Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Comparative Supervised Machine Learning for Predicting Cardiovascular Disease among Indian Adults 45 Years or Older: Evidence from the LASI Wave 1

**Vikas Kamble**

vikas.iips007@gmail.com

International Institute for Population Sciences

**Wahengbam Bigyananda Meitei**

International Institute for Population Sciences

**Murali Dhar**

International Institute for Population Sciences

**Research Article**

**Keywords:** Cardiovascular disease, machine learning, Older Adults, India

**Posted Date:** November 17th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11390178/v1

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

**Additional Declarations:** No competing interests reported.

# Comparative Supervised Machine Learning for Predicting Cardiovascular Disease among Indian Adults 45 Years or Older: Evidence from the LASI Wave 1

Vikas Kamble [1], Wahengbam Bigyananda Meitei [2], Murali Dhar [3]

1. PhD Scholar
   Department of Biostatistics and Demography,
   International Institute for Population Sciences,
   City: Mumbai, Country: India
   Email: vikas.iips007@gmail.com

2. Junior Research Scientist, GENDER-Project
   International Institute for Population Sciences
   City: Mumbai, Country: India
   Email: bigyanandawahengbam@gmail.com

3. Professor
   Department of Bio-Statistics & Epidemiology
   International Institute for Population Sciences
   City: Mumbai, Country: India
   Email: m.dhar@iipsindia.ac.in

**Corresponding Author:**
Vikas Kamble
Department of Biostatistics and Demography,
International Institute for Population Sciences, Mumbai, India
City: Mumbai, Country: India
Email: vikas.iips007@gmail.com

**Abstract**

**Background:** Cardiovascular disease (CVD) remains a leading cause of death in world, where rapid demographic aging and epidemiological transition are intensifying the burden of older adults. Machine learning (ML) offers potential to improve early risk stratification beyond traditional clinical risk scores, yet population-level evidence from nationally representative Indian data remains limited. Our study aimed to develop and validate ML models for predicting CVD among older adults using the Longitudinal Ageing Study in India (LASI) wave 1, comparing seven supervised algorithms and identifying key predictive features.

**Methods:** The analytic sample consists of 64,447 older adults. Using stratified sampling, the data were divided into a training set (56%, n = 36,089), a calibration set (14%, n = 9,023), and a held-out test set (30%, n = 19,335). The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set only, to correct class imbalance. Seven algorithms were evaluated: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, LightGBM, and Multilayer Perceptron (MLP). Test-set performance was assessed using $AUC_{ROC}$, $AUC_{PRC}$, Brier score, and decision curve analysis, with bootstrap 95% confidence intervals. SHapley Additive exPlanations (SHAP) were used to interpret feature contributions.

**Results:** Random Forest showed the best discrimination $AUC_{ROC}$ 72.8%, 95% CI 72.1–73.6; $AUC_{PRC}$ 53.86%, 95% CI 52.5–55.2), with LightGBM and XGBoost following closely ($AUC_{ROC}$ 72.6% each). These three ensemble ML algorithms formed a distinct upper tier, about three percentage points above the individual classifiers, Logistic Regression, SVM, Decision Tree, and MLP. Random Forest had the lowest Brier score (0.177), achieving a sensitivity of 61.6% and a specificity of 71.9% at its selected threshold (0.309). Age, having diabetes, higher body mass index, and waist-hip ratio, and having sleep difficulty were identified as the most influential feature variables. Predicted risk rises sharply between ages 58 and 60, then plateaus.

**Conclusion:** Our findings suggest that an ensemble ML algorithm, the Random Forest classifier, could help identify adults aged 45 years or older at higher risk of CVD using routinely collected survey data, thereby aiding policy-makers in making evidence-based decisions while implementing accessible chronic disease management services.

**Keywords:** Cardiovascular disease; machine learning; Older Adults; India

## Introduction

Cardiovascular disease (CVD) consists of a range of diseases affecting the heart and blood vessels, including coronary heart disease, hypertension, and stroke (1). These conditions represent a major public health problem because of their high prevalence, severity, and interrelated nature, as well as the substantial burden they impose on individuals and health systems (2). A family history of CVD-related conditions has been associated with approximately a two-fold higher risk, indicating a genetic predisposition (3). Globally, CVD accounts for 17.9 million deaths annually, representing nearly one-third of all deaths (4). The burden of cardio-metabolic non-communicable diseases (NCD), particularly CVD and diabetes, is increasing rapidly in low- and middle-income countries (LMICs), with deaths occurring at younger ages as opposed to the declining burden in high-income countries (5). Despite a near doubling of the global CVD burden between 1990 and 2017, mortality continues to rise in lower-income countries while declining in high-income countries, underscoring persistent inequalities in disease prevention and treatment (6, 7).

In India, CVD has become a significant public health challenge and is currently the leading cause of overall and premature mortality (8). It accounts for approximately a quarter of all deaths in the country, with ischemic heart disease and stroke contributing to more than 80% of the total CVD burden (9). India's burden is further compounded by the relatively earlier onset of these conditions compared with Western populations. Rapid urbanization, changing dietary patterns, and the increasing prevalence of hypertension and diabetes have intensified the cardiovascular health burden and placed additional strain on an already overstretched health system. These challenges underscore the urgent need for more comprehensive prevention strategies and strengthened efforts toward early detection and timely intervention (10).

Machine learning (ML) algorithms have been increasingly employed to predict CVD risk, enabling more accurate risk estimates and earlier disease detection than traditional approaches of risk assessment based on limited clinical parameters (11). A key advantage of LM is its ability to model complex, non-linear relationships and higher-order interactions among multiple predictors, relationships that may not be adequately captured by conventional regression-based risk assessment studies (12). This has resulted in the increasing use of ML in a variety of healthcare settings, such as the interpretation of diagnostic images, risk stratification based on electronic health records, and population health screening based on

survey data, often with performance comparable to or better than traditional statistical methods when trained on adequate data (13).

Several studies have explored the application of ML to predict CVD risk using clinical and electronic medical records. Comparative assessments of algorithms, including Naïve Bayes, Decision tree, k-nearest neighbours, and Artificial neural Networks, have reported varying predictive performance, with Decision tree and Naïve Bayes models demonstrating comparatively better performance than the other models (14). In fact, Konstantina and Antonis (13) evaluated six machine learning algorithms using data from 2,943 participants in the Ludwigshafen Risk and Cardiovascular Health (LURIC) cohort and identified Logistic Regression as the most reliable model, achieving 72.2% accuracy. Moreover, the potential clinical utility of ML-based prediction has been demonstrated, with Miranda and Irwansyah (15) reporting a Naïve Bayes classifier that achieved accuracy, sensitivity, and specificity above 80%, whose predictive performance was subsequently validated by clinicians. Evidence from hospital-based studies in LMIC settings has also highlighted the importance of incorporating context-specific predictors; a study conducted in Somaliland identified culturally specific behaviours, including chewing and smoking, as important predictors of CVD risk (16). Similarly, the Prospective Urban Rural Epidemiology (PURE) study, conducted in 21 LMICs, between 2013-2014, identified several socio-demographic and behavioural determinants, including lower educational attainment, tobacco use, harmful alcohol consumption, physical inactivity, and unhealthy diet, and further reported that these factors exert a greater influence in LMICs (17).

Building on these foundation, using the Longitudinal Ageing Study in India, Wave-I (LASI-1), a nationally representative survey of Indian adults aged 45 years or older, our study systematically compares multiple supervised ML algorithms, employs advanced resampling techniques to address class imbalance, and applies variable importance analysis to identify key CVD predictors, with the overall aim of developing an accurate early prediction model whose outputs are interpreted through explainability techniques and translated into a useful risk prediction tool for deployment in Indian healthcare settings based on the routinely collected survey-data, with the potential relevance for population-level screening and risk assessment.

## Materials and Methods

Our study employed supervised machine learning classification approaches to predict CVD risk among adults aged 45 years or older in India. The analytical pipeline comprised of six sequential stages: identification of data source, exploratory data analysis, data preprocessing and feature engineering, class imbalance correction, hyperparameter optimization, and model training and evaluation. A detailed description of each stage is provided in the subsequent sections.

### *Data*

We used data from the LASI-1, a nationally representative survey of the middle-aged and older adult population (45+ years, including their spouses, irrespective of age) in India, covering all states and Union Territories. LASI-1 is a comprehensive nationwide survey that collects information on functional health, socioeconomic well-being, access to health care, social activities, and disease burden among older adults. LASI-1 interviewed 73,396 individuals, of which 64,447 were included in our analysis. **Figure 1** provides detailed information on how we arrived at the final sample size for the study.

[**Figure 1 Here**]

### *Outcome Variable*

The outcome variable in our study is CVD status, defined on the respondents' self-reported diagnosis of hypertension, chronic heart disease, or stroke by a qualified health professional. Respondents who reported having at least one of these conditions were classified as having CVD and coded as 1, whereas those who did not report any of these conditions were classified as not having CVD and coded as 0. This binary classification of the outcome variable provides a simple and focused framework for assessing the relationship between the predictor variables and CVD status.

### *Feature Variables*

Based on the existing literature, we included a range of features to examine their relationships with CVD. These variables encompassed socio-demographic, behavioral, physiological, and

health-related characteristics. Socio-demographic variables includes, age, sex (male, female), place of residence (rural, urban), marital status (Married, Widowed, Others), education (no education, less than secondary education, secondary education), monthly per capita expenditure (MPCE) quintile (poorest, poorer, middle, richer, richest), caste (scheduled tribe, scheduled caste, other backward classes, others), and religion (Hindu, Muslim, Christian, other). Behavioural risk factors include, smoking (never smoked, ever smoked) and alcohol consumption (never, ever). Physiological measures included anthropometric measures such as body mass index (BMI) (categorized as underweight, healthy overweight, and obese), and waist-to-hip ratio (WHR). Functional status was assessed using activities of daily living (ADL) and instrumental activities of daily living (IADL), both dichotomized as "yes" or "no". Additionally, the presence of comorbid conditions, including cancer, diabetes, chronic lung disease, chronic bone disease, psychiatric depression, high cholesterol, chest pain, and other diseases, was also recorded as a binary variable (yes or no). Collectively, these covariates were selected to capture the multifactorial determinants of CVD risk within the study population.

### *Analytical Approach*

We employed supervised ML classification approaches, and the ML pipeline is discussed in detail in the following sections. The process begins with data acquisition from the LASI-1, followed by exploratory data analysis and data preprocessing steps, including missing-value treatment, one-hot encoding of categorical variables, and variable scaling for distance-sensitive algorithms. In the next stage, the dataset is partitioned into training, calibration, and held-out test sets using a stratified three-way split, after which the Synthetic Minority Over-sampling Technique (SMOTE) is applied exclusively to the training partition to correct class imbalance. Finally, seven supervised classification algorithms undergo randomized hyperparameter optimization via cross-validation, and the best-performing model is selected and evaluated on the held-out test set using a comprehensive suite of performance metrics, including Area Under the Receiving Operating Curves ($\mathrm{AUC}_{ROC}$) and Area Under the Precision-Recall Curve ($\mathrm{AUC}_{PRC}$), calibration (Brier score), and also SHapley Additive Explanations (SHAP).

### *Data Preprocessing*

These preprocessing steps constitute a critical component of the ML pipeline, as they substantially influence both the quality of the input data and the predictive performance of the

models (14). Prior to model development, these procedures were applied sequentially to the LASI-1 data. The prevalence of CVD was 29.2%, indicating a moderate degree of class imbalance that required appropriate correction during model training.

All feature variables were initially examined to ensure consistency in data types. Since all variables were of integer type, except for age and WHR, no further data type conversion was needed (18). The LASI-1 employs standardized non-response codes, including ".d" to denote "don't know" responses and ".r" to denote refusal. Observations with these codes were either excluded using complete-case analysis or recoded into a separate missing category. Thus, the final analytic sample consisted of 64,447 older adults.

All categorical features, including socio-demographic, behavioral, physiological, and health-related variables, were transformed using one-hot encoding, with the first category omitted to avoid multi-collinearity. For distance-sensitive models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP), variable scaling was applied within the pipeline using Z-score standardization (mean = 0, standard deviation = 1). On the other hand, tree-based models were trained on unscaled variables. The final design matrix consists of 64,447 observations and 36 variables, including two continuous variables (age and WHR), and 34 binary variables, which were subsequently used in all analyses.

***Data Partitioning***

The dataset was partitioned using a three-way stratified hold-out approach comprising training, calibration, and test sets to facilitate independent model training, probability calibration, and unbiased assessment of predictive performance (19). Stratification on the binary CVD outcome was applied at each split to preserve the observed prevalence of 29.2% across all three partitions. Relative to the total sample ($N$ = 64,447), 30% of the data ($n$ = 19,335) were allocated to the test set, 14% ($n$ = 9,023) to the calibration set, and 56% ($n$ = 36,090) training dataset. Reporting all partitions as proportions of the overall sample ensured clarity and eliminated the potential ambiguity of nested percentages. A fixed random seed of 42 was used throughout the analysis to ensure reproducibility.

### *Class Imbalance*

To address the moderate class imbalance in the dataset, SMOTE was applied exclusively to the training dataset, resulting in a balanced training dataset of 51,126 observations. SMOTE was not applied to the calibration and test datasets to preserve their natural class distributions and ensure the model evaluation under realistic conditions. Consistent with the recommendations of Chawla (20), restricting SMOTE to the training dataset while retaining the natural distribution in the test dataset enhances the model's ability to generalize to unseen real-world data and provides a more reliable assessment of predictive performance.

### *Calibration Curve*

A calibration curve is a useful technique in machine learning and forecasting that helps us to understand and improve the accuracy of predicted probabilities from classification models. But to use these probabilities, we need a well-calibrated model. A perfectly calibrated model is characterized by a calibration curve that follows the 45-degree diagonal line extending from (0, 0) to (1, 1), indicating complete concordance between predicted probabilities and the observed proportion of positive cases (21).

### *Hyperparameter*

Hyperparameter optimization was performed using a randomized search strategy, in which 50 randomly selected hyperparameter combinations were evaluated for each model (22). Model performance for each combination was assessed using the $\mathrm{AUC}_{ROC}$, a threshold-independent metric well-suited for imbalanced binary classification. Randomized search was preferred over exhaustive grid search due to its computational tractability. In fact, Bergstra and Bengio (22) demonstrated that random sampling of hyperparameter configurations typically matches the performance of grid search with substantially fewer evaluations, particularly when the effective dimensionality of the search space is low. The optimal hyperparameter configuration for each model was defined as the hyperparameter set yielding the highest mean $\mathrm{AUC}_{ROC}$ across the three inner cross-validation folds (nested within the five-fold outer cross-validation) and was subsequently retained for final model training on the full training dataset.

The optimal classification threshold for each model was determined using Youden's J statistic (J = sensitivity + specificity − 1), selecting the probability cutoff that maximized J on the ROC curve.

### *Machine Learning Algorithms*

Seven supervised machine learning algorithms were employed to predict CVD: logistic regression, decision tree, random forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Multilayer Perceptron (MLP). These algorithms were chosen based on their demonstrated effectiveness in clinical prediction studies and their ability to capture diverse linear and non-linear relationships between feature variables and the outcome variable. A brief overview of each algorithm is provided below to contextualize its application in the present study.

Logistic regression is a supervised learning algorithm commonly used for binary and multiclass classification problems. It estimates the probability of class membership by applying the logistic (sigmoid) function to a linear combination of the feature variables (23). For a binary outcome variable $Y$ and a set of feature variables $X_1$, $X_2$, …, $X_n$, the relationship between the feature variables and the outcome is expressed through the logit function as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Where, $p$ denotes the predicted probability of CVD, $\beta_0$ is the intercept, and $\beta_1$ to $\beta_n$ are the regression coefficients estimated using the maximum likelihood method. To improve model generalizability and reduce over fitting, $L_2$ regularization was used by penalizing large coefficient estimates. The regularization parameter (C), which controls the strength of the penalty, was optimized using a randomized hyperparameter search.

A decision tree is a supervised learning algorithm that constructs a tree-like model by recursively partitioning the variable space into a series of hierarchical binary splits (24). In this structure, each internal node represents a decision rule based on a predictor variable, each branch corresponds to the outcome of that decision, and each terminal (leaf) node assigns a class label. At every split, the algorithm selects the variable and threshold that minimize

information gain or, equivalently, minimize Gini impurity, thereby improving the homogeneity of the resulting subsets. Although decision trees are highly interpretable and computationally efficient, they are susceptible to overfitting, particularly when allowed to grow without constraints. To improve model generalizability, hyperparameters controlling tree complexity, including the maximum tree depth, the minimum number of samples required to split an internal node, and the minimum number of samples required at a leaf node, were optimized using a randomized hyperparameter search.

Random forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrap samples of the training data and predicts the final class through majority voting across individual trees (25). At each node split, a random subset of variables is considered, a method known as variable bagging, which reduces correlation among trees and enhances generalization. Owing to its strong predictive performance, random forest has been widely used in biomedical research, including CVD diagnosis and risk prediction (26). The number of trees, maximum tree depth, minimum samples per leaf, and the number of variables considered at every split were optimized through randomized hyperparameter search.

SVM is a supervised learning algorithm that identifies the optimal hyperplane separating two classes by maximizing the margin between the nearest data points of each class, known as support vectors (27). In our study, a linear kernel was employed using Linear SVC, with the regularization parameter $C$ controlling the trade-off between margin maximization and training error minimization. Since SVM does not naturally produce probability-calibrated estimates, sigmoid calibration was performed after training to convert decision function scores into interpretable CVD risk probabilities, using a hold-out calibration set.

XGBoost is a scalable gradient boosting algorithm that sequentially ensembles decision trees, with each tree correcting the residual errors of its predecessors (28). By incorporating L1 and L2 regularization into its objective function, XGBoost effectively reduces overfitting and has demonstrated excellent performance in structured clinical prediction tasks involving high-dimensional datasets. The optimized hyperparameters included the number of estimators, learning rate, maximum tree depth, subsample ratio, variable subsampling ratio, and regularization coefficients.

LightGBM is an efficient gradient boosting framework that employs tree-based learning algorithms and is optimized for large-scale datasets. It improves computational efficiency of conventional GBDT implementations through Gradient-based One-Side Sampling (GOSS)**,** which retains data instances with larger gradients while randomly discarding those with smaller gradients, thereby reducing the number of data records evaluated without significant loss of predictive information, and Exclusive Feature Bundling (EFB)**,** which combines sparse, mutually exclusive variables to a single variable, effectively reducing the dimensionality of the variable space and accelerating tree construction (29).

MLP is a feed-forward artificial neural network comprising an input layer, one or more hidden layers, and an output layer. Each node in the input layer represents one variable in the input dataset. Hidden-layer neurons apply a nonlinear activation function to the weighted sum of their inputs, enabling the model to learn complex relationships, while the output layer produces the predicted class probability (30). The hyperparameters optimized included the number and size of hidden layers, the activation function, the L2 regularization strength (alpha), and the initial learning rate.

*SHapley Additive Explanations (SHAP)*

In interpreting SHAP-based variable importance plots, the width of the effect range **is** a reasonable approximation to permutation importance, as its magnitude may be disproportionately influenced by a small number of extreme observations. A broad dispersion of all points across the plot may indicate greater permutation importance, reflecting a significant variation in the variable's contribution to model predictions. SHAP value represents the estimated contribution of a given variable to the prediction. Accordingly, a pattern extending from upper left to lower right indicates that lower values of the variable of interest are associated with higher predicted outcomes. On the other hand, a dispersed or mixed pattern suggests that the direction of the variable importance is not consistent across observations. In such cases, both high and low values of the variable can have positive or negative effects on the prediction outcome, indicating potential heterogeneity in the variable-outcome relationship (31).

*Performance Evaluation with Bootstrap Confidence Intervals*

We evaluated models on the held-out test set (N = 19,335) using a comprehensive panel of metrics. To assess discrimination, we calculated the $\mathrm{AUC}_{ROC}$ and the $\mathrm{AUC}_{PRC}$, the latter being especially informative in the presence of class imbalance (32). The Classification performance measures at the chosen threshold include sensitivity (recall for the positive class), specificity, positive predictive value (precision), negative predictive value, and the F1 score. To assess calibration quality, we employed the Brier score, which combines discrimination and calibration into a single proper scoring rule.

All data preprocessing, model building, and statistical analysis were performed in Google Colab. ML-based models were built using the scikit-learn library (33) . Class imbalance was addressed by the SMOTE implementation of the imbalanced-learn library, while model interpretability was measured using the SHAP (SHapley Additive explanations) package.

**Result**

**Table 1** presents the prevalence of CVD by background characteristics among 64,447 older adults, of whom 29.17% (95% CI: 28.82–29.52) had CVD. The prevalence of CVD was observed to be highest among those with high cholesterol, followed by those with diabetes, representing a nearly 2.6-fold higher prevalence among diabetics: 71.38% (95% CI: 69.43–73.25) among those with high cholesterol vs. 27.71% (95% CI: 27.36–28.06) among those with no cholesterol, and 63.09% (95% CI: 62.00–64.16) among those with diabetes vs. 24.59% (95% CI: 24.23–24.94) with no diabetes. The CVD prevalence also increases significantly with increasing BMI. The prevalence of CVD was much higher in those who were obese (41.13%; 95% CI: 40.45–41.82) than in those who were underweight (17.42%; 95% CI: 16.67–18.21). Having psychiatric problems (46.93%; 95% CI: 44.23–49.64), ADL (41.76%; 95% CI: 40.70–42.83), and IADL limitations (35.14%; 95% CI: 34.49–35.81), and chest pain (40.29%; 95% CI: 39.48–41.11) also had a high prevalence of CVD. Socio-demographic gradients were evident across socioeconomic status: 22.31% (95% CI: 21.60–23.04) among those belonging to poorest MPCE quintile, and 36.05% (95% CI: 35.22–36.89) among those belonging to richest MPCE quintile, 36.87% (95% CI: 36.24–37.50) among urban residents, and 25.10% (95% CI: 24.69–25.52) among rural residents. Interestingly, older adults using smokeless tobacco (25.26%; 95% CI: 24.69–25.83) have a lower prevalence of CVD than those who do

not use (31.24%; 95% CI: 30.80–31.68). A similar pattern was also observed with alcohol consumption: 24.24% (95% CI: 23.44–25.07) among those who consume alcohol, and 30.15% (95% CI: 29.76–30.53) among those who do not consume. The mean age of participants was 57.70 years (95% CI: 57.61–57.79), and the mean WHR was 0.93 (95% CI: 0.93–0.93).

[**Table 1 Here**]

**Table 2** presents the cross-validation AUC score for hyperparameter optimization. We optimized the hyperparameters using Randomized Search strategy with 50 iterations and 5-fold stratified cross-validation, and SMOTE was applied only on each training fold to prevent data leakage. All seven models demonstrated moderate cross-validation AUC performance (range: 67%-72%), comparable to population-based CVD prediction studies using survey data. Random Forest had the best performance with a hyper-parameter tuning of (71.66%; 95% CI: 70.52-72.80) followed by XGBoost (71.46%; 95% CI: 70.64-72.28) and LightGBM (71.37%; 95% CI: 70.58-72.16). All three ensemble methods outperformed the individual classifiers, such as MLP (69.16%; 95% CI: 68.49-69.83), Logistic Regression (68.96%; 95% CI: 67.89-70.03), SVM (69.02%; 95% CI: 67.95-70.09), and Decision Tree (68.14%; 95% CI: 66.90-69.38).

[**Table 2 Here**]

**Figure 2** shows the calibration curves of different models using the Brier score. It shows that the best fit to the perfect calibration diagonal was achieved by Random Forest (Brier score = 0.177), followed by XGBoost and LightGBM (both Brier score = 0.178), with a slight overestimation at predicted probabilities above 0.50. Logistic Regression (Brier score = 0.184), Decision Tree (Brier score = 0.184), and MLP (Brier score = 0.186) exhibited moderate calibration with higher deviation in the mid-range. SVM was the worst-calibrated model (Brier score = 0.213), systematically underestimating CVD risk across all probability bins and exhibiting a compressed prediction range (0.20–0.80), likely due to distorted probability scaling caused by the Linear SVC's sigmoid calibration.

[**Figure 2 Here**]

**Figure 3** presents the AUC of the Receiver Operating Curve and Precision-Recall Curve of the seven machine learning models evaluated on the held-out test set. The $\text{AUC}_{ROC}$ for all seven models were much higher than the random baseline. This suggests that each method has a high degree of discriminative power for predicting CVD. Random Forest showed the best discrimination with the highest $\text{AUC}_{ROC}$ (72.8%; 95% CI: 72.1-73.6). XGBoost (72.6%; 95% CI: 71.8-73.3) and LightGBM (72.6%; 95% CI: 71.8-73.3) were close, forming a distinct ensemble performance tier about three percentage points above single classifiers.

[**Figure 3 Here**]

With respect to the Precision-Recall Curve, all seven models significantly outperformed the no-skill baseline with $\text{AUC}_{PRC}$ scores ranging from 48.41% to 53.9%. Among the models, Random Forest achieves the highest $\text{AUC}_{PRC}$ (53.9%; 95% CI: 52.5–55.2), followed closely by XGBoost (53.0%; 95% CI: 51.7–54.4) and LightGBM (52.9%; 95% CI: 51.5–54.2). These three tree-based ensemble models constituted a higher-performing group, with overlapping confidence intervals indicating no statistically clear difference in their predictive performance. The remaining models showed comparatively lower $\text{AUC}_{PRC}$ values, including Logistic Regression (50.1%; 95% CI: 48.9–51.5), SVM (50.0%; 95% CI: 48.7–51.3), MLP (48.7%; 95% CI: 47.4–50.1), and Decision Tree (48.4%; 95% CI: 47.2–49.8). The non-overlapping confidence intervals between Random Forest and SVM, as well as between Random Forest and Decision Tree, suggest that the observed differences in performance are unlikely to be attributable solely to sampling variability.

**Figure 4** presents the decision curve analysis of the seven machine learning models. The Y-axis is net clinical benefit. The dotted line at zero represents the "treat none" strategy, whereas the dashed line represents the "treat all" strategy. The net benefit in the clinically relevant threshold range of 25%–50% was best for Random Forest. SVM resulted in negative net benefit above approximately 38% implying net clinical harm at higher decision thresholds. All other models showed positive net benefit above both reference strategies over the entire threshold

range, confirming clinical utility for population-based CVD screening in the LASI-1 sample. Taken together, Random Forest outperformed the other models in discrimination, calibration, and net clinical benefit, justifying its selection as the primary model for predicting CVD risk in the LASI-1.

[**Figure 4 Here**]

**Figure 5(a)** presents the feature importance plot of model interpretability using SHAP values for the Random Forest model on 2000 test observations and 36 feature variables. Age was the strongest predictor (mean |SHAP| = 0.040), followed by severe sleep difficulty (mean |SHAP| = 0.034), diabetes (mean |SHAP| = 0.031), obesity (mean |SHAP| = 0.027), and BMI category overweight (mean |SHAP| = 0.025). The next top three predictors were smokeless tobacco use (mean |SHAP| = 0.020), WHR (mean |SHAP| = 0.019), and being a member of a household belonging to the scheduled caste and tribe (mean |SHAP| = 0.017). The importance curve declined steadily through the top ten predictors and flattened out around rank twelve, indicating diminishing marginal contributions beyond that point and suggesting that a model with the top ten to twelve predictors would capture most of the predictive information.

[**Figure 5 Here**]

The SHAP dependence plot (**Figure 5(b)**) for age indicated a nonlinear pattern in the model. SHAP contributions were negative (range: −0.08 to −0.01) for patients aged 28–55 years, corresponding to lower predicted CVD probability in younger individuals, and transitioned sharply to positive values (+0.03 to +0.05) between ages 58 and 60, after which they plateaued beyond age 60.

## Discussion

CVD remains a leading cause of mortality globally (8). Delayed detection of CVD can adversely affect population prognosis and increase the risk of complications and mortality. Therefore, timely and accurate identification of individuals at high risk of CVD is essential for strengthening preventive strategies and reducing CVD-related mortality (34). In this context,

ML offers a promising approach for capturing complex and potentially non-linear patterns with large-scale health datasets and developing predictive models to support risk stratification and informed clinical decision-making.

Our study adds to the growing evidence on ML-based CVD prediction by evaluating seven supervised ML algorithms on 64,447 Indian adults aged 45 years or older, using nationally representative data from the LASI-1. Ensemble-based approaches, particularly Random Forest, demonstrated better performance than other individual classifiers, with an $\mathrm{AUC}_{ROC}$ of 72.8%, Brier score of 0.177, and favourable results in decision curve analysis. SHAP-based analysis further identified key predictors and revealed a prominent non-linear association with age, with predicted CVD risk increasing sharply from age approximately 58 to 60 years. These findings suggest that ML models may facilitate clinically meaningful CVD risk stratification using routinely collected population-level survey data, potentially reducing reliance on expensive clinical biomarkers or diagnostic testing.

The seven models demonstrate moderate and comparable discriminative performance, with $\mathrm{AUC}_{ROC}$ ranging from 69.5% to72.8%. These findings are consistent with previous population-based CVD prediction studies that relied on self-reported survey data without clinical biomarkers, supporting the potential utility of questionnaire-based information for population-level CVD risk screening (35). Among the evaluated models, Random Forest, XGBoost, and LightGBM formed a distinct ensemble performance group, achieving AUC values approximately three percentage points higher than those of single classifiers, including SVM, Logistic Regression, Decision Tree, and MLP. This observation is broadly consistent with the meta-analysis findings of Krittanawong et al. (36), which examined 103 cohorts comprising 3,377,318 participants and reported that boosting algorithms achieved the highest discriminatory performance for coronary artery disease (CAD) prediction, with a pooled $\mathrm{AUC}_{ROC}$ of 88%, (95% CI: 84–91) and custom-built algorithms had a pooled AUC of 93% (95% CI: 85–97). In contrast, Forrest et al. (31) reported much higher Random Forest $\mathrm{AUC}_{ROC}$ values ranging from 91% to 95% for CAD using structured electronic health record (EHR) data from the USA and UK Biobank. Such differences may reflect the greater availability of detailed clinical variables and other information in electronic health records and biobank datasets. Accordingly, the moderate AUC values observed in our study may primarily reflect the

information content and measurement limitations of the survey-based features rather than the limitations of the evaluated algorithms.

Decision curve analysis also indicated that the Random Forest model yielded the highest net clinical benefit across the clinically relevant threshold range of 25–50%, outperforming both the "treat all" and "treat none" reference strategies. These findings suggest that Random Forest may have utility for CVD risk stratification and could guide tiered screening approaches: intensive monitoring for high-risk individuals, standard care for moderate-risk individuals, and routine screening for low-risk individuals. Given its reliance on routinely collected survey variables, such an approach may also offer a scalable and relatively low-cost framework for population-level risk stratification across diverse geographical and socioeconomic settings in India.

Krittanawong et al. (37) have highlighted substantial methodological heterogeneity across published ML-based CVD prediction studies. These studies differed in their choice of hyperparameters, the approach to partitioning the data, the method of feature selection and had no standardized reporting or head-to-head comparisons with traditional risk scores. Such heterogeneity makes direct comparison across studies difficult and limits the feasibility of meaningful meta-analytic synthesis. Our study addresses some of these methodological concerns through a transparent and systematic modelling framework, including 50 iterations of randomized hyperparameter search, three-way stratified data splitting into training, calibration, and held-out test set, application of SMOTE exclusively within the training data to minimize data leakage, selection of the optimal classification threshold using Youden's optimal cutoff, and estimation of 95% confidence intervals using bootstrap resampling. For comparison, Hossain et al. (38) reported near-perfect performance of a Random Forest model ($AUC_{ROC}$ = 99%, accuracy = 98.04%) using a clinical CVD dataset from Bangladesh. However, the relatively small sample size (n = 651) and lack of external validation may limit the generalizability of these findings. In contrast, our study uses a substantially larger, nationally representative sample of 64,447 adults, thereby providing broader population-level evidence for ML-based CVD risk prediction in India.

The five most influential SHAP features, age, sleep difficulty, diabetes, obesity, and WHR

align with established cardiovascular epidemiology, providing biological plausibility for the feature importance identified by the ML-models (39,40). Similarly, Hossain et al. (38), using a clinical CVD dataset from Bangladesh (n = 651), identified having high cholesterol, hypertension, abnormal heart rhythms, and sleep-related problems as key predictors of CVD. The overlap between the findings of Hossain et al. (38) and those of our study, particularly for age and metabolic comorbidities, further supports the clinical and biological relevance of the model-derived feature rankings.

The identification of socioeconomic characteristics, including caste, and household expenditure quintile among the important predictors also highlights the potential contributions of sociocultural and socioeconomic determinants of CVD risk among older Indian adults. Such factors are less commonly captured in Western or hospital-based clinical Multi-Ethnic Study of Atherosclerosis (MESA) data from the National Sleep Research Resource data. Furthermore, the importance of sleep-related characteristics in our study is consistent with findings from imaging-based and clinical ML studies (41). In addition, Forrest et al. (31) demonstrated the potential translational value of interpretable ML models for each CVD detection, reporting that 46% of high-risk individuals identified by their model had previously undiagnosed CAD. Collectively, these findings support the relevance of both conventional clinical risk factors and context-specific socioeconomic and behavioural characteristics in ML-based CVD risk prediction.

The higher prevalence of CVD observed among urban residents, individuals belonging to higher-income households, and those with higher educational attainment challenges the conventional view that CVD burden is predominantly concentrated among socioeconomically disadvantaged populations. Instead, these findings may reflect India's ongoing epidemiological transition, in which CVD is increasingly affecting relatively advantaged groups undergoing rapid changes in lifestyle and behavioural patterns (42,43), a pattern also observed in the PURE study (17). The pronounced co-occurrence of CVD with psychiatric conditions, functional limitations in activities of daily living (ADL/IADL), and sleep problems further highlights the multifactorial nature of CVD risk among older adults. These patterns underscore the potential value of integrated and comprehensive risk assessment approaches that consider multiple interrelated health and social factors alongside conventional disease-specific screening.

These findings have important implications for CVD prevention policy in India. The successful application of ML to large-scale survey data demonstrates the feasibility of predicting population CVD risk without costly clinical investigation with only routinely collected socio-demographic, anthropometric, and behavioral features and the good prediction performance (34). The non-linear age threshold identified through SHAP dependence analysis offers a hypothesis-generating cut point with direct relevance for age-based eligibility criteria in national screening programmes. This inflection should be considered a feature of how the ensemble model partitions age, not definitive evidence of a biological threshold, and would need prospective verification before informing formal screening criteria. These findings are directly relevant to India's existing NCD screening infrastructure. The National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases and Stroke (NPCDCS) currently relies on universal opportunistic screening for adults aged 30 and above, using a simple point-based Community-Based Assessment Checklist (CBAC) administered by frontline health workers; a validated ML-based risk model built entirely from routinely available variables could complement this checklist by enabling more granular, data-driven risk stratification within the eligible population without requiring additional infrastructure or cost (44,45). Because Ayushman Bharat Health and Wellness Centres already record sociodemographic, anthropometric, and behavioural information as part of routine comprehensive primary care, the predictors used in this model align closely with data already collected at the point of care, suggesting a feasible pathway for integration into existing digital health platforms. Furthermore, the socioeconomic and demographic gradients observed in CVD prevalence, particularly the elevated burden among urban, wealthier, and more educated subgroups, support supplementing NPCDCS's uniform population-wide approach with targeted risk communication for these emerging high-risk groups, consistent with India's National Multisectoral Action Plan for NCDs, which emphasizes strengthened surveillance and early detection. Embedding such tiered, ML-informed risk stratification within existing primary care infrastructure could thus improve the efficiency of India's NCD screening efforts without necessitating parallel systems or additional clinical resources (44).

Our study has a few limitations that should be considered when interpreting the findings. First, we used data from only the first wave of LASI, which makes it impossible to infer causality or determine the temporal association between the identified features and CVD outcomes.

However, with the release of subsequent waves of LASI in future, a longitudinal analysis would be performed to capture the sequence of risk factor exposure and disease onset, thereby reinforcing predictive validity. Second, both CVD status and behavioral characteristics were self-reported, making the findings susceptible to recall bias, social desirability bias and outcome misclassification, which are common concerns in survey-based epidemiological studies. Third, the model's discriminative performance may have been constrained by the absence of detailed clinical assessments, biomarker measurements, and comprehensive dietary information. Fourth, despite the use of SMOTE to address class imbalance, synthetic oversampling generates artificial observations that may not fully reflect the underlying distribution of the minority class, and could potentially introduce distributional bias during model training.

**Conclusion**

Our study demonstrates the potential of ensemble ML methods, particularly Random Forest, to accurately and meaningfully predict CVD risk among older adults in India using nationally representative survey data, without requiring expensive clinical tests or biomarker profiling. SHAP analysis identified age, having diabetes, sleep problems, being obese and having a higher WHR as the key predictors of CVD, indicating that the model captured biologically realistic CVD risk factors. Importantly, the nonlinear association between age and predicted CVD risk observed in the ensemble models, but not by the linear classifiers, highlights the potential importance of accounting for age-specific CVD screening strategies in India. These findings suggest that ML-based risk stratification could provide a cost-effective and scalable approach for early CVD risk detection in resource-limited situations. Future research that incorporates longitudinal study designs, clinical and biomarker information, and larger and more diverse populations could further strengthen the robustness, predictive validity, and generalizability of CVD risk prediction models.

**Ethics, Consent to Participate, and Consent to Publish**: Not applicable. Our study uses de-identified secondary data from a publicly available dataset. No individual person data in any form (including individual details, images, or videos) are presented in this manuscript.

**Availability of data and materials**: The datasets analysed during the current study are available from the International Institute for Population Sciences (IIPS), Mumbai, through the LASI data request process https://www.iipsindia.ac.in/content/LASI-data

**Competing interests**: The authors declare that they have no competing interests.

**Funding**: This research received no specific grant from any funding agency in the public, commercial, Or not-for-profit sectors.

**Author's contributions**: VK and WBM conceptualized the study design and methodology. VK conducted the statistical analysis, including developing the machine learning models, and preparing the figures and tables. VK wrote the first draft of the manuscript. VK, WBM, and MD critically reviewed and revised the manuscript, reviewed the revised version, and approved it for submission.

**Clinical trial number**: Not applicable

**References:**


1. *Kundu J, Kundu S. Cardiovascular disease (CVD) and its associated risk factors among older adults in India: Evidence from LASI Wave 1. Clinical Epidemiology and Global Health. 2022 Jan 1;13:100937. doi:10.1016/j.cegh.2021.100937*

2. *Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. Journal of the American College of Cardiology. 2020 Dec 22;76(25):2982–3021. doi:10.1016/j.jacc.2020.11.010*

3. *Bachmann JM, Willis BL, Ayers CR, Khera A, Berry JD. Association Between Family History and Coronary Heart Disease Death Across Long-Term Follow-Up in Men. Circulation. 2012 Jun 26;125(25):3092–8. doi:10.1161/CIRCULATIONAHA.111.065490 PubMed PMID: 22623718; PubMed Central PMCID: PMC3631594.*

4. *WHO. Cardiovascular diseases (CVDs) [Internet]. 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed January 21, 2024.*

5. *Miranda JJ, Barrientos-Gutiérrez T, Corvalan C, Hyder AA, Lazo-Porras M, Oni T, et al. Understanding the rise of cardiometabolic diseases in low- and middle-income countries. Nature Medicine. 2019 Nov;25(11):1667–79. doi:10.1038/s41591-019-0644-7*

6. *Mensah GA, Roth GA, Fuster V. The Global Burden of Cardiovascular Diseases and Risk Factors: 2020 and Beyond. Journal of the American College of Cardiology. 2019 Nov 19;74(20):2529–32. doi:10.1016/j.jacc.2019.10.009*

7. *Wurie HR, Cappuccio FP. Cardiovascular disease in low- and middle-income countries: an urgent priority. Ethn Health. 2012 Jan 1;17(6):543–50. doi:10.1080/13557858.2012.778642 PubMed PMID: 23534502; PubMed Central PMCID: PMC7613448.*

8. *Gull M, Saini TC. Global and Indian cardiovascular disease mortality dynamics across multiregional epidemiological transitions differential risk factor burdens and integrated policy frameworks for advancing sustainable development goal 3.4. Discover Social Science and Health. 2026 Mar 21;6(1):71. doi:10.1007/s44155-026-00405-7*

9. *Prabhakaran D, Jeemon P, Roy A. Cardiovascular Diseases in India. Circulation. 2016 Apr 19;133(16):1605–20. doi:10.1161/CIRCULATIONAHA.114.008729*

10. *Reddy KS, Yusuf S. Emerging Epidemic of Cardiovascular Disease in Developing Countries. Circulation. 1998 Feb 17;97(6):596–601. doi:10.1161/01.CIR.97.6.596*

11. *Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal - Digital Health. 2025 Jan 22;6(1):7–22. doi:10.1093/ehjdh/ztae080*

12. *Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers [Internet]. 2022 Jan 1 [cited 2026 Aug 11]. Available from: https://www.degruyterbrill.com/document/doi/10.1515/med-2022-0508/html?lang=en*

13. *Tsarapatsani K, Sakellarios AI, Pezoulas VC, Tsakanikas VD, Kleber ME, März W, et al. Machine Learning Models for Cardiovascular Disease Events Prediction. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) [Internet]. 2022 [cited 2026 Jun 22]. p. 1066–9. Available from: https://ieeexplore.ieee.org/abstract/document/9871121 doi:10.1109/EMBC48229.2022.9871121*

14. *Soni J, Ansari U, Sharma D, Soni S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. IJCA. 2011 Mar 31;17(8):43–8. doi:10.5120/2237-2860*

15. *Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Healthc Inform Res. 2016;22(3):196. doi:10.4258/hir.2016.22.3.196*

16. *Yh M, Ma H, H A, N B, K H, I D, et al. Relationship between lifestyle factors and cardiovascular disease prevalence in Somaliland: A supervised machine learning approach using data from Hargeisa Group Hospital, 2024. PubMed [Internet]. [cited 2026 Jun 22]. Available from: https://pubmed.ncbi.nlm.nih.gov/39828104/*

17. *Yusuf S, Joseph P, Rangarajan S, Islam S, Mente A, Hystad P, et al. Modifiable risk factors, cardiovascular disease and mortality in 155,722 individuals from 21 high-, middle-, and low-income countries. Lancet. 2020 Mar 7;395(10226):795–808. doi:10.1016/S0140-6736(19)32008-2 PubMed PMID: 31492503; PubMed Central PMCID: PMC8006904.*

18. *Abdelouahed SM, Abla R, Asmae E, Abdellah A. Harnessing feature engineering to improve machine learning: A review of different data processing techniques. In: 2024 International Conference on Intelligent Systems and Computer Vision (ISCV) [Internet]. 2024 [cited 2026 Jun 22]. p. 1–6. Available from: https://ieeexplore.ieee.org/abstract/document/10620105 doi:10.1109/ISCV60512.2024.10620105*

19. *Bennett M, Nekouei M, Mehta APR, Kleczyk E, Hayes K. Methodology to Create Analysis-Naive Holdout Records as well as Train and Test Records for Machine Learning Analyses in Healthcare [Internet]. arXiv; 2022 [cited 2026 Jun 22]. Available from: http://arxiv.org/abs/2205.03987 doi:10.48550/arXiv.2205.03987*

20. *Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 Jan 1;16:321–57. doi:10.1613/jair.953*

21. *Austin PC, Putter H, Giardiello D, Van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagn Progn Res. 2022 Dec;6(1):2. doi:10.1186/s41512-021-00114-6*

22. *Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012 Feb 1;13(null):281–305.*

23. *Nusinovici S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of Clinical Epidemiology. 2020 Jun 1;122:56–69. doi:10.1016/j.jclinepi.2020.03.002*

24. *Quinlan JR. Induction of decision trees. Machine Learning. 1986 Mar;1(1):81–106. doi:10.1007/BF00116251*

25. *Breiman L. Random Forests. Machine Learning. 2001 Oct;45(1):5–32. doi:10.1023/A:1010933404324*

26. *Maniruzzaman Md, Kumar N, Menhazul Abedin Md, Shaykhul Islam Md, Suri HS, El-Baz AS, et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer Methods and Programs in Biomedicine. 2017 Dec 1;152:23–34. doi:10.1016/j.cmpb.2017.09.004*

27. *Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995 Sep;20(3):273–97. doi:10.1007/BF00994018*

28. *Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2026 Jun 22]. p. 785–94. (KDD '16). Available from: https://dl.acm.org/doi/10.1145/2939672.2939785 doi:10.1145/2939672.2939785*

29. *Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2026 Jun 22]. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html*

30. *El-Hassani FZ, Amri M, Joudar NE, Haddouch K. A New Optimization Model for MLP Hyperparameter Tuning: Modeling and Resolution by Real-Coded Genetic Algorithm. Neural Processing Letters. 2024 Mar 14;56(2):105. doi:10.1007/s11063-024-11578-0*

31. *Forrest IS, Petrazzini BO, Duffy Á, Park JK, Marquez-Luna C, Jordan DM, et al. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. The Lancet. 2023 Jan;401(10372):215–25. doi:10.1016/S0140-6736(22)02079-7*

32. *T S, M R. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PubMed [Internet]. 2015 [cited 2026 Aug 13]. Available from: https://pubmed.ncbi.nlm.nih.gov/25738806/?utm_source=chatgpt.com*

33. *Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python [Internet]. [cited 2026 Sep 23]. Available from: https://www.jmlr.org/papers/v12/pedregosa11a.html?utm_source=chatgpt.com*

34. *Karwa V, Wanjari A, Kumar S, Dhondge RH, Patil R, Kothari M. Optimizing Cardiovascular Health: A Comprehensive Review of Risk Assessment Strategies for Primary Prevention. Cureus. 2024 Aug 6. doi:10.7759/cureus.66341*

35. *Wang H, Tucker WJ, Jonnagaddala J, Schutte AE, Jalaludin B, Liaw ST, et al. Using machine learning to predict cardiovascular risk using self-reported questionnaires: Findings from the 45 and Up Study. International Journal of Cardiology. 2023 Sep 1;386:149–56. doi:10.1016/j.ijcard.2023.05.030*

36. *Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific Reports. 2020 Sep 29;10(1):16057. doi:10.1038/s41598-020-72685-1*

37. *Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology. 2017 May 30;69(21):2657–64. doi:10.1016/j.jacc.2017.03.571*

38. *Hossain S, Hasan MK, Faruk MO, Aktar N, Hossain R, Hossain K. Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023. BMC Cardiovascular Disorders. 2024 Apr 18;24(1):214. doi:10.1186/s12872-024-03883-2*

39. *Pradeepa R, Mohan V. Prevalence of type 2 diabetes and its complications in India and economic costs to the nation. European Journal of Clinical Nutrition. 2017 Jul;71(7):816–24. doi:10.1038/ejcn.2017.40*

40. *Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, editor. PLoS ONE. 2017 Apr 4;12(4):e0174944. doi:10.1371/journal.pone.0174944*

41. *Gourishetti SC, Taylor R, Isaiah A. Stratifying the Risk of Cardiovascular Disease in Obstructive Sleep Apnea Using Machine Learning. The Laryngoscope. 2022;132(1):234–41. doi:10.1002/lary.29852*

42. *Banerjee K, Dwivedi LK. Burden assessment of infectious and cardio-vascular diseases in India for the decade 2004-2014. Epidemiol Health. 2016 Dec 14;e2016057. doi:10.4178/epih.e2016057*

43. *Geldsetzer P, Manne-Goehler J, Theilmann M, Davies JI, Awasthi A, Danaei G, et al. Geographic and sociodemographic variation of cardiovascular disease risk in India: A cross-sectional study of 797,540 adults. Peiris D, editor. PLoS Med. 2018 Jun 19;15(6):e1002581. doi:10.1371/journal.pmed.1002581*

44. *qps.nhsrcindia.org/sites/default/files/2021-03/Operational Guideline Comprehensive Primary Health Care.pdf [Internet]. [cited 2026 Sep 24]. Available from: https://qps.nhsrcindia.org/sites/default/files/2021 03/Operational%20Guideline%20Comprehensive%20Primary%20Health%20Care.pdf*

45. *NCD Programme [Internet]. [cited 2026 Jun 22]. Available from: https://ncd.mohfw.gov.in/ncdlandingassets/aboutus.html*

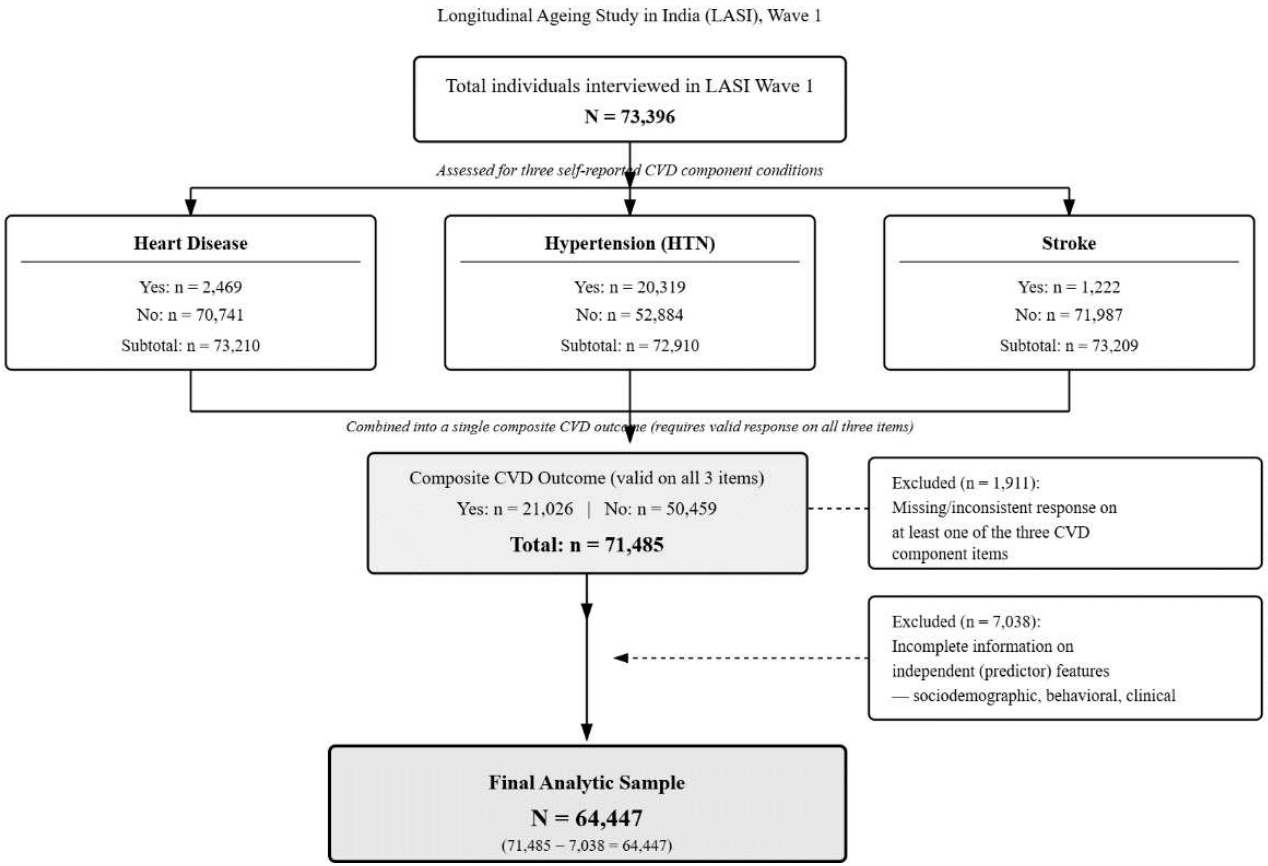



**Figure 1**: sample selection flowchart

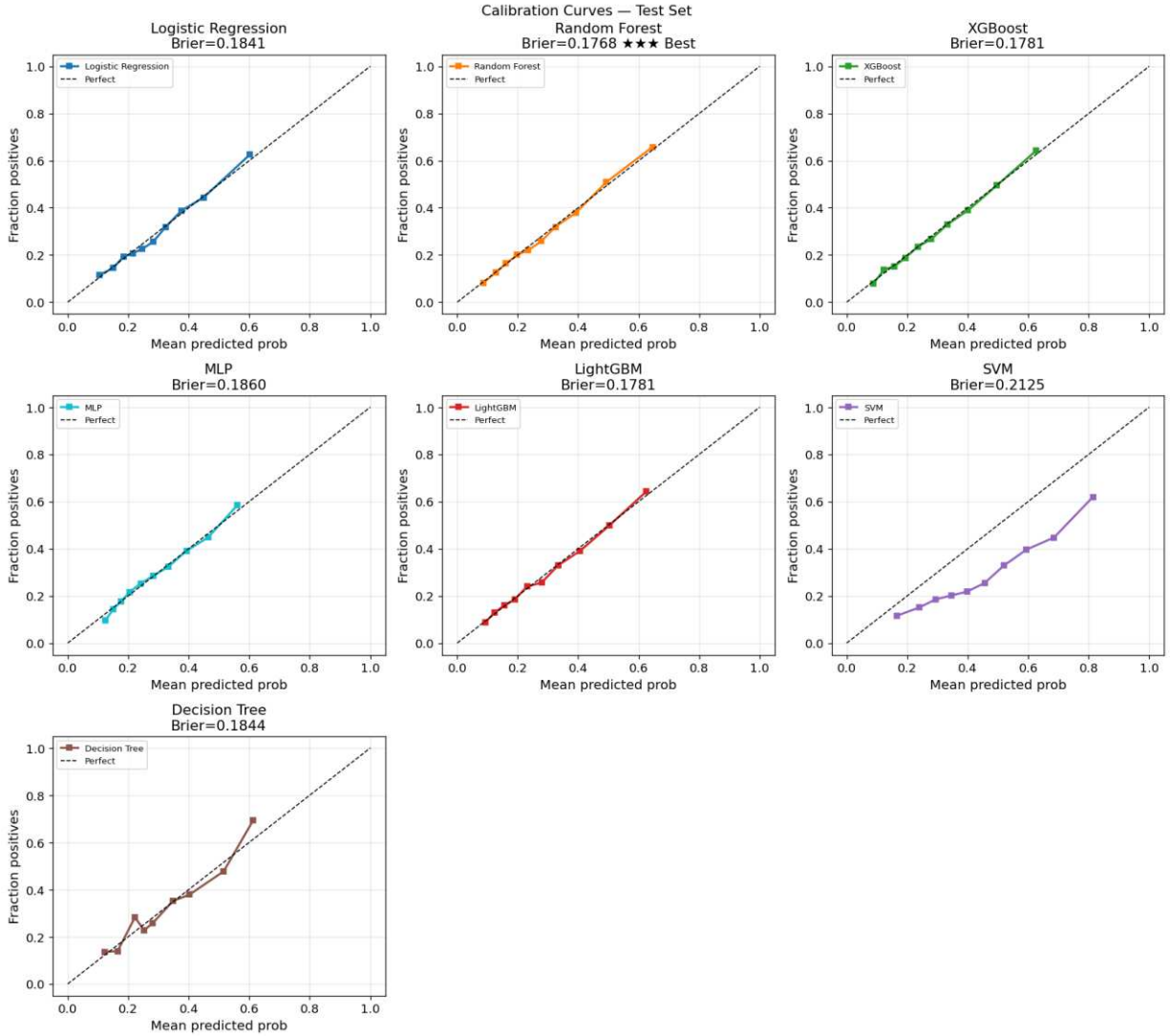



**Figure 2**: calibration curves

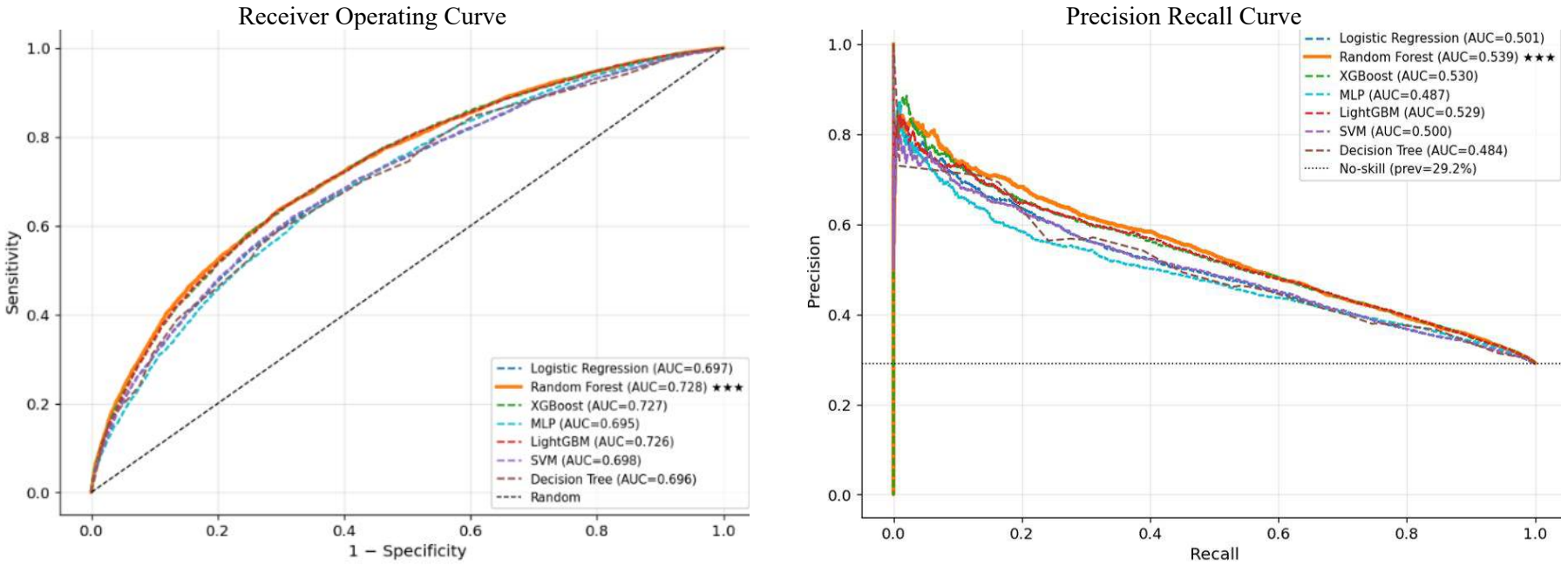



***Figure 3**: Area under the curve of the Receiver Operating Curve and Precision Recall Curve*

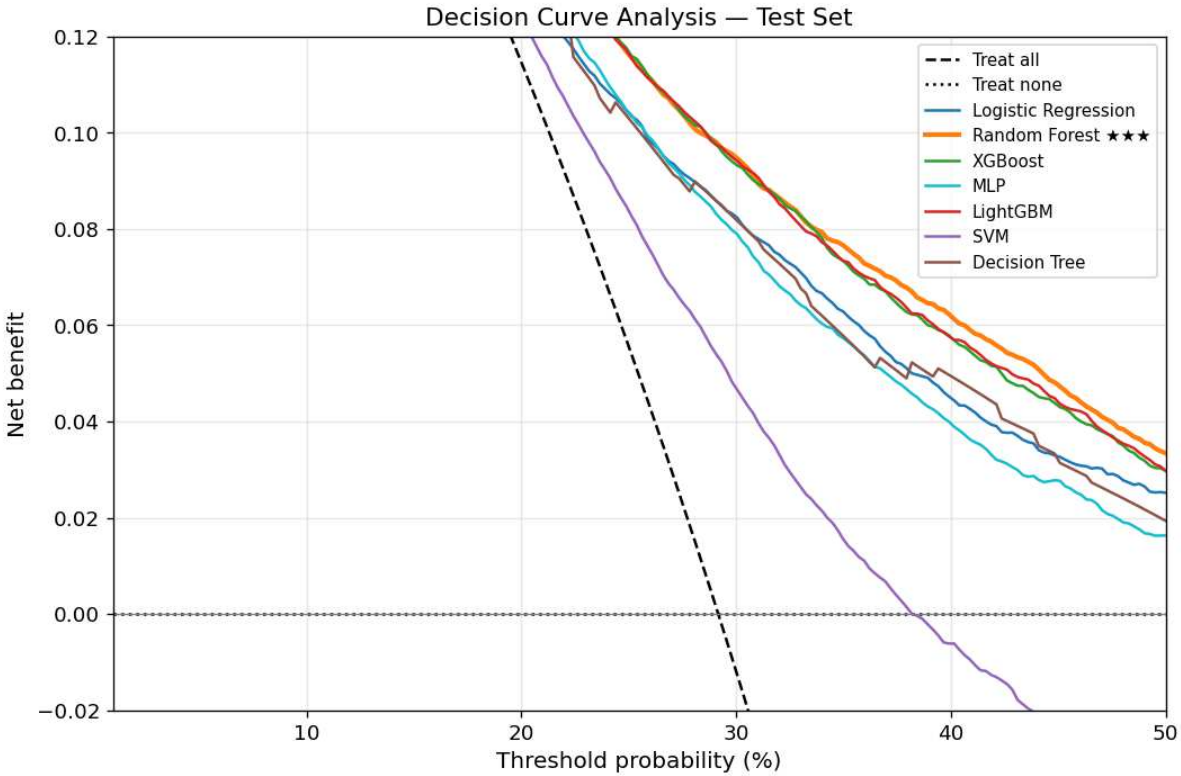



**Figure 4**: Decision curve

**(a) SHAP Feature Importance**

**(b) SHAP dependence plot**

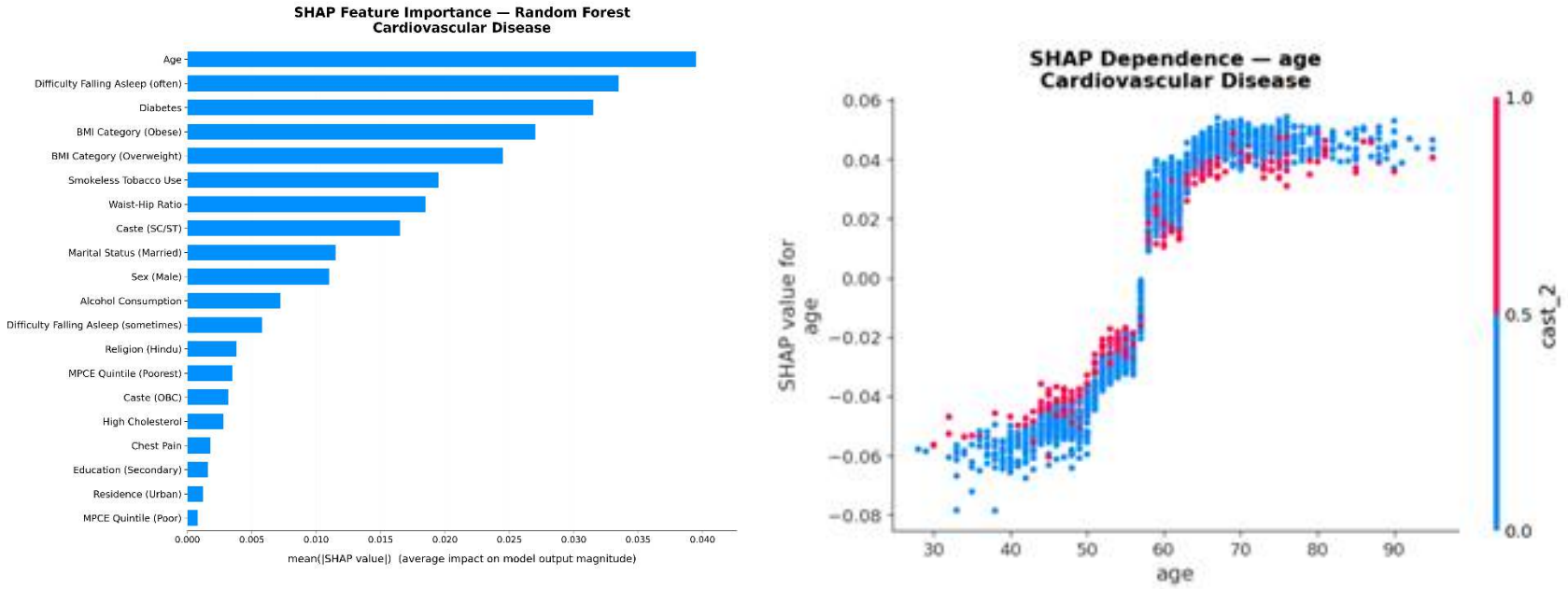



**Figure 5**: SHAP Feature Importance and Age Dependence Plots for Cardiovascular Disease Prediction

**Table 1**: **Prevalence of Cardiovascular disease of older adult's population in India by its background characteristics**

| Variables (Features) | Prevalence (95% CI) |
|---|---|
| **Age#** | |
| Mean | 57.70%(57.61 - 57.79) |
| **WHR#** | |
| Mean | 0.93% (0.93 - 0.93) |
| **Sex** | |
| Female | 30.80% (30.33–31.27) |
| Male | 26.92% (26.40–27.45) |
| **Residence** | |
| Rural | 25.10% (24.69–25.52) |
| Urban | 36.87% (36.24–37.50) |
| | |
| **Religion** | |
| Muslim | 35.95% (34.87–37.03) |
| Christian | 25.97% (24.92–27.06) |
| Hindu | 27.89% (27.48–28.29) |
| Others | 38.25% (36.59–39.94) |
| **Social Category** | |
| OBC | 29.12% (28.56–29.70) |
| Others | 35.73% (35.03–36.44) |
| SC | 27.11% (26.28–27.96) |
| ST | 21.05% (20.31–21.80) |
| **Marital Status** | |
| Married | 27.01% (26.63–27.41) |
| Widowed | 37.87% (37.03–38.71) |
| Others | 27.14% (25.21–29.16) |
| **Education** | |
| None | 26.57% (26.07–27.08) |

| | |
|---|---|
| < secondary Education | 30.50% (29.79–31.22) |
| > secondary Education | 32.15% (31.48–32.82) |
| **MPCE Quintile** | |
| Poorest | 22.31% (21.60–23.04) |
| Poorer | 25.90% (25.16–26.66) |
| Middle | 29.31% (28.54–30.10) |
| Richer | 32.33% (31.53–33.13) |
| Richest | 36.05% (35.22–36.89) |
| **Smokeless Tobacco** | |
| Yes | 25.26% (24.69–25.83) |
| No | 31.24% (30.80–31.68) |
| **Alcoholic Beverages** | |
| Yes | 24.24% (23.44–25.07) |
| No | 30.15% (29.76–30.53) |
| **Fall Asleep** | |
| Occasionally/frequently | 39.39% (38.47–40.31) |
| 1–2 night | 33.54% (32.77–34.32) |
| Never | 24.79% (24.36–25.21) |
| **BMI** | |
| Underweight (<18) | 17.42% (16.67–18.21) |
| Healthy (18–22.9) | 23.12% (22.61–23.64) |
| Overweight (23–24.9) | 32.06% (31.13–33.01) |
| Obese (≥25) | 41.13% (40.45–41.82) |
| **Diabetes** | |
| Yes | 63.09% (62.00–64.16) |
| No | 24.59% (24.23–24.94) |
| **Cancer** | |
| Yes | 41.21% (36.48–46.10) |
| No | 29.09% (28.74–29.45) |
| **Chronic Lung Disease** | |
| Yes | 41.75% (40.12–43.41) |

| | |
|---|---|
| No | 28.46% (28.10–28.82) |
| **Chronic Bone** | |
| Yes | 42.81% (41.79–43.83) |
| No | 26.95% (26.58–27.32) |
| **Psychiatric problems** | |
| Yes | 46.93% (44.23–49.64) |
| No | 28.80% (28.45–29.16) |
| **Cholesterol** | |
| Yes | 71.38% (69.43–73.25) |
| No | 27.71% (27.36–28.06) |
| **Chest Pain** | |
| Yes | 40.29% (39.48–41.11) |
| No | 26.13% (25.74–26.51) |
| **Other Chronic Disease** | |
| No other Chronic | 26.66% (26.27–27.05) |
| Other chronic | 36.88% (36.13–37.64) |
| **Feel Depressed** | |
| Depressed | 31.84% (31.27–32.42) |
| Never | 27.46% (27.02–27.91) |
| **ADL** | |
| No | 27.34% (26.97–27.71) |
| Yes | 41.76% (40.70–42.83) |
| **IADL** | |
| No | 26.46% (26.06–26.88) |
| Yes | 35.14% (34.49–35.81) |

*Note: CVD = cardiovascular disease; CI = confidence interval; MPCE = monthly per capita consumption expenditure; ADL = activities of daily living; IADL = instrumental activities of daily living; OBC = Other Backward Classes; SC = Scheduled Castes; ST = Scheduled Tribes, BMI=Body Mass Index, WHR=Waist Hip Ratio, # = Mean*

**Table 2**: Cross-validation AUC score for hyperparameter optimization

| Models | AUC (CV) |
|---|---|
| **Logistic Regression** | 68.85 [67.89-70.03] |
| **Random Forest** | 71.66 [70.52-72.80] |
| **Decision Tree** | 68.14 [66.90-69.38] |
| **MLP** | 69.16 [68.49-69.83] |
| **XGBoost** | 71.46 [70.64-72.28] |
| **LightGBM** | 71.37 [70.58-72.16] |
| **SVM** | 69.02 [67.95-70.09] |