JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

# Recursive Feature Elimination with Cross-Validation (RFECV) for Feature Subset Selection in Cardiovascular Disease Prediction: A Leakage-Free Evaluation of Ten Classifiers

**Raphael Ozighor Enihe**
Department of Computer Science, Baze University, Abuja, Nigeria

## ABSTRACT

Clinical datasets often contain redundant or weakly informative variables that add noise, increase computational cost and obscure the predictors that matter most for cardiovascular disease (CVD). This study applies Recursive Feature Elimination with Cross-Validation (RFECV) to 1,000 patient records (874 CVD-positive, 126 CVD-negative) from the University of Abuja Teaching Hospital (UATH), Nigeria. To prevent information leakage, the records were first divided into a stratified training set (80%) and a hold-out test set (20%). SMOTE-Tomek balancing was then applied to the training set only (800 to 1,372 records), and RFECV, with an Extra Trees Classifier, stratified five-fold cross-validation and ROC-AUC scoring, reduced the fourteen predictors to five: Age, Smoking, Weight, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP). Weight (importance 0.452) and SBP (0.186) ranked highest. Ten classifiers were trained with and without feature selection and evaluated on the untouched test set using accuracy, precision, recall, F1-score, specificity, balanced accuracy, ROC-AUC and PR-AUC, and the findings were checked over 20 repeated stratified splits. On the primary split, feature selection raised ROC-AUC and specificity for all ten models (for example, ROC-AUC for the Support Vector Machine and Extra Trees rose from 0.70 to 0.80), whereas accuracy, recall and F1-score were unchanged or slightly lower for most models. The repeated splits showed the same pattern, with ROC-AUC gains ranging from negligible (LightGBM) to +0.11 (KNN). The impact of RFECV, however, depended on the model and the metric used: in this instance it resulted in a smaller predictor set and improved identification of patients without CVD at a relatively small price in terms of sensitivity. Any decrease in clinical testing, cost/time is a potential implication and needs to be clinically proven in the future.

**KEYWORDS:** Cardiovascular Disease, Feature Selection, Recursive Feature Elimination (RFECV), SMOTE-Tomek, Extra Trees, Machine Learning, Nigeria.

## I. INTRODUCTION

Cardiovascular disease (CVD) is the most common cause of death globally (World Health Organization, 2021) and machine learning (ML) is currently being applied widely to predict the occurrence of the disease. In clinical data sets, however, there are many more variables to be considered than are required to make accurate predictions. Irrelevant, redundant or noisy features can degrade model performance because the model fits noise rather than signal, and they increase computational cost and reduce interpretability (Wan, 2019). Feature selection addresses this problem by retaining only the variables that contribute most to the target outcome (Alfebi & Anasanti, 2023; Ghosh et al., 2021). Feature selection methods fall into three broad categories: filter, wrapper and embedded methods. Filter methods rank features using statistical measures independent of any learning algorithm. Wrapper methods evaluate candidate subsets using the performance of a specific model. Embedded methods perform selection as part of model training, as in LASSO regression and tree-based importance.



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

Wrapper methods are computationally more intensive, but they can yield stronger subsets because they account for interactions between features. Recursive Feature Elimination (RFE) is a wrapper method that repeatedly fits a model, ranks the features by importance and removes the weakest one (Guyon et al., 2002). Recursive Feature Elimination with Cross-Validation (RFECV) extends RFE by using cross-validation to determine the number of features that gives the best score, which removes the need to specify that number in advance.

## II. RELATED WORK

Feature selection continues to be an active area of CVD-prediction research. Akhter and Miller (2026) compared four feature-selection strategies, including a hypergraph-based method, for CVD risk prediction on NHANES data, and found that the best-performing feature subset, deployed in a public web application, improved both accuracy and AUC over a correlation-filter baseline. Uddin et al. (2026) used XGBoost-based feature importance to reduce thirteen Cleveland-dataset predictors to nine before ensembling, and reported negligible performance loss. Khawla et al. (2026) deliberately retained all available predictors because their datasets contained few redundant features, which illustrates that the benefit of feature selection is dataset-dependent. Tabbassum et al. (2026) used Random Forest-based importance ranking on a combined Kaggle/UCI dataset and identified ST-segment slope and chest-pain type as the dominant predictors, which shows that the dominant predictors vary with the feature set and the population sampled. A SMOTE balanced, feature selected Random Forest model as a web tool presented a recall of 96% on a large, real world data set (Chakraborty et al., 2026), and the feature selection based on univariate and LASSO models with ensemble voting resulted in a high coronary heart disease risk discrimination (AUC = 0.906) (Hassan et al., 2026). Riswanda et al. (2026) and Yuliana et al. (2026) examined SMOTE-Tomek balancing with several classifiers, and Li et al. (2026) and Pidishetty and Bukaita (2026) addressed CVD risk modelling with interpretable approaches. Together, these studies indicate that principled feature reduction remains central to CVD prediction pipelines, but that its benefit depends on the model, the dataset and the evaluation metric. This motivates the controlled comparison reported here. This study is part of a broader investigation of CVD prediction using the UATH clinical dataset (Enihe et al., 2025). Its research question is: how does RFECV, applied to a SMOTE-Tomek-balanced training set, affect the performance of ML models for CVD prediction, and which predictors does it retain? The main contributions are as follows. First, the study reports a leakage-free pipeline in which the train/test split precedes balancing, feature selection and model training. Second, it justifies the choice of the RFECV estimator and the cross-validation scheme with a direct empirical comparison. Third, it compares ten classifiers with and without feature selection on an untouched test set using clinically relevant metrics (sensitivity, specificity, F1-score, ROC-AUC and PR-AUC), and checks the findings over 20 repeated stratified splits. Fourth, it reports the complete experimental configuration so that the experiment can be reproduced.

## III. MATERIALS AND METHODS

### A. Dataset and Class Distribution

The study used 1,000 anonymised patient records retrospectively abstracted from the medical files of patients at UATH, Abuja, Nigeria. Ethical clearance and permission to use the records for research were obtained from the UATH Ethics Committee. Each record contains 14 predictors and a binary outcome (Cvd: 1 = CVD present, 0 = CVD absent). The predictors are Age, Gender, Smoke, Alcohol, Def_in_VitD (vitamin D deficiency), Diabetes, High_Cholesterol, Weight, Chestpain, FamilyHistoryCVD, SBP, DBP, HBP (high blood pressure indicator) and Proteinuria. Weight is body weight in kilograms (range 59–164 kg); height was not recorded, so body mass index could not be derived. The dataset contains no



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

missing values or duplicate records. Of the 1,000 records, 874 (87.4%) are CVD-positive and 126 (12.6%) are CVD-negative, so the dataset is imbalanced and the minority class is the CVD-negative group. Table 1 shows the class distribution at each stage of the pipeline. The records carry no admission date or other timestamp, so no temporal ordering of patients can be established.

**Table 1:** Class distribution at each stage of the pipeline.

| Data subset | n | CVD-positive, n (%) | CVD-negative, n (%) | Role |
|---|---|---|---|---|
| **Complete dataset** | 1,000 | 874 (87.4) | 126 (12.6) | Source data |
| **Training set (80%)** | 800 | 699 (87.4) | 101 (12.6) | Before balancing |
| **Training set after SMOTE-Tomek** | 1,372 | 686 (50.0) | 686 (50.0) | RFECV and model training |
| **Hold-out test set (20%)** | 200 | 175 (87.5) | 25 (12.5) | Final evaluation only |

## B. Experimental Pipeline and Prevention of Data Leakage

The exact sequence of operations, shown in Figure 1, was as follows: (1) the 1,000 records were split into training (80%) and test (20%) sets using stratified sampling (random_state = 42); (2) SMOTE-Tomek balancing was applied to the training set only; (3) RFECV was fitted on the balanced training set; (4) the ten classifiers were trained on the balanced training set, once with all 14 features and once with the RFECV subset; and (5) all models were evaluated once on the untouched test set, which retained the original class ratio. Balancing, feature selection, hyperparameter tuning and feature scaling were fitted on training data only. Balancing before the split would allow synthetic records derived from test-set patients to enter training and would inflate performance estimates (Kapoor & Narayanan, 2023). The drawback of this sequence is that the internal cross-validation folds of RFECV are generated from the balanced training set, which also includes synthetic records; this means that the internal scores are optimistic (Santos et al., 2018). Only the internal scores were used for the comparison of the size of the subsets. All performance assertions in this paper use the original set of tests.

## C. Class Balancing with SMOTE-Tomek

The class imbalance issue was addressed using the Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) and Tomek-link cleaning (Tomek, 1976; Batista et al., 2004). SMOTE makes synthetic minority-class instances by interpolating between a minority instance and its closest minority neighbours, while Tomek-link cleaning eliminates borderline pairs of minority and majority instances. The implementation used the imbalanced-learn library (Lemaître et al., 2017) with default settings (five nearest neighbours) and random_state = 42. On the training set, SMOTE increased the CVD-negative class from 101 to 699 records, and Tomek-link cleaning then removed 13 records from each class, giving 686 records per class (1,372 in total; Table 1). The test set was not resampled.



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

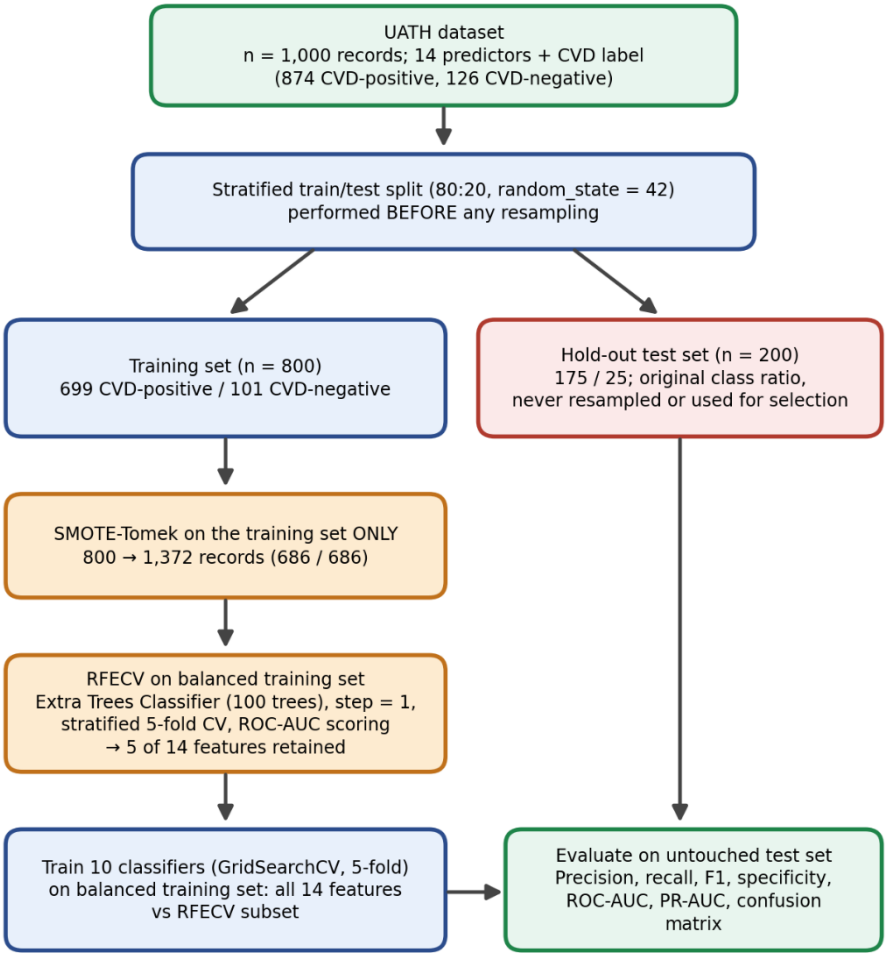



**Figure 1:** Experimental pipeline

The hold-out test set is separated before any resampling and is used only for the final evaluation.

## D. Recursive Feature Elimination with Cross-Validation

RFECV starts with all 14 features, fits the base estimator, ranks the features by importance, and removes the least important one (step size = 1). The process is repeated until one feature remains, and at each subset size the cross-validated score is recorded. The subset size with the highest mean cross-validated score is selected as optimal. The importance of feature $j$ in a tree ensemble is the mean decrease in impurity (Louppe, 2014; Darji, 2024; Tsiligaridis, 2023):

$$\text{Importance}(f_j) = (1/T) \sum_{t=1..T} \sum_{n \in t:\, v(n)=j} p(n) \cdot \Delta i(n) \qquad (1)$$

where $T$ is the number of trees, $p(n)$ is the proportion of samples reaching node $n$, $v(n)$ is the feature used to split at node $n$, and $\Delta i(n)$ is the impurity decrease produced by that split. Algorithm 1 gives the complete procedure, and Table 2 lists every setting required to reproduce it.

**Algorithm 1.** Leakage-free RFECV feature selection and model evaluation.

Input: dataset $D$ (1,000 records; 14 predictors $X$; binary label $y$)
1. Split $D$ into ($X_{train}$, $y_{train}$) and ($X_{test}$, $y_{test}$) with test size 0.20, stratified on $y$, random_state = 42; set the test set aside.



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

2. Apply SMOTE-Tomek (random_state = 42) to ($X_{train}$, $y_{train}$) to obtain ($X_{bal}$, $y_{bal}$).
3. Initialise the estimator: ExtraTreesClassifier (n_estimators = 100, random_state = 42).
4. Define the cross-validation scheme: StratifiedKFold (n_splits = 5, shuffle = True, random_state = 42).
5. Initialise RFECV (estimator, step = 1, min_features_to_select = 1, cv, scoring = ROC-AUC).
6. Fit RFECV on ($X_{bal}$, $y_{bal}$); obtain the support mask and the optimal number of features $n^*$.
7. Define the selected subset $S$ = features retained in the support mask.
8. For each classifier, tune hyperparameters by grid search on ($X_{bal}$[all 14], $y_{bal}$) and on ($X_{bal}$[$S$], $y_{bal}$).
9. Predict on $X_{test}$ (all 14) and $X_{test}$[$S$]; compute all evaluation metrics.
Output: subset $S$, importance ranking, and test-set metrics with and without feature selection.

***(a). Choice of the RFECV estimator***

The base estimator was an Extra Trees (Extremely Randomised Trees) ensemble (Geurts et al., 2006), which is fast, handles mixed binary and continuous predictors without scaling, and yields impurity-based importance scores. Because the outcome is binary, the Extra Trees Classifier is the natural choice and was used for the final selection. An Extra Trees Regressor fitted to the 0/1 label is nevertheless a defensible alternative: for a binary label with positive-class proportion $p$, the variance criterion used by the regressor, $p(1 - p)$, is proportional to the Gini impurity used by the classifier, $2p(1 - p)$ (Breiman et al., 1984), so the two estimators evaluate split candidates in the same way. They differ mainly in default settings (in scikit-learn, the regressor considers all features at each split, whereas the classifier considers the square root of the number of features) and in the scoring function that RFECV optimises (Pedregosa et al., 2011). To confirm that the classifier is appropriate for the feature-selection objective, the two estimators were compared directly (Section 3.1).

***(b). Choice of the cross-validation scheme***

TimeSeriesSplit assumes that observations are ordered in time and always validates on later observations than those used for training (Bergmeir & Benítez, 2012; Pedregosa et al., 2011). The UATH records are cross-sectional chart abstractions without admission dates or any other timestamp, so the row order carries no temporal information and a time-ordered splitter has no basis. In addition, SMOTE-Tomek appends its synthetic CVD-negative records after the original records, so a time-ordered split of the balanced data produces folds that reflect the resampling procedure rather than any property of the patients (Section 3.1). Stratified k-fold cross-validation with shuffling preserves the class proportions in every fold (Kohavi, 1995) and was therefore adopted for RFECV and for hyperparameter tuning.

## E. Classifiers and Hyperparameter Tuning

Ten classifiers were evaluated: Light Gradient Boosting Machine (LightGBM; Ke et al., 2017), Support Vector Machine (SVM; Cortes & Vapnik, 1995), Extra Trees (Geurts et al., 2006), CatBoost (Prokhorenkova et al., 2018), Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2001), XGBoost (Chen & Guestrin, 2016), K-Nearest Neighbours (KNN; Cover & Hart, 1967), Decision Tree (Breiman et al., 1984) and AdaBoost (Freund & Schapire, 1997). Each classifier was tuned by grid search with stratified five-fold cross-validation on the balanced training set, scoring by ROC-AUC, separately for the all-feature and RFECV-subset conditions and using the same search grid for both. SVM and KNN were wrapped in a pipeline



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

with standardisation fitted within each training fold; tree-based models were not scaled. The search grids and the selected hyperparameters are given in Appendix A (Table A1), and all other settings are in Table 2.

**Table 2.** Complete Experimental Configuration

| Component | Setting |
|---|---|
| **Data split** | Stratified 80:20 hold-out (train_test_split, test_size = 0.20, random_state = 42); 800 training and 200 test records |
| **Balancing** | SMOTE-Tomek (imbalanced-learn), training set only; random_state = 42; SMOTE k = 5; default Tomek-link cleaning |
| **RFECV estimator** | ExtraTreesClassifier; n_estimators = 100; random_state = 42; criterion = Gini; max_features = sqrt; remaining settings at library defaults |
| **RFECV settings** | step = 1; min_features_to_select = 1; scoring = ROC-AUC; cv = StratifiedKFold(5, shuffle = True, random_state = 42) |
| **Classifier tuning** | GridSearchCV; cv = StratifiedKFold(5, shuffle = True, random_state = 42); scoring = ROC-AUC; random_state = 42 for all stochastic models; grids in Table A1 |
| **Scaling** | StandardScaler inside the pipeline for SVM (RBF kernel, probability = True) and KNN only |
| **Decision threshold** | 0.5 on the predicted probability of CVD |
| **Repeated analysis** | 20 stratified 80:20 splits (random_state = 100–119); the full pipeline, including balancing and RFECV, repeated for each split; classifier hyperparameters fixed at the values in Table A1 |
| **Software** | Python 3.12.3; scikit-learn 1.8.0; imbalanced-learn 0.14.2; LightGBM 4.7.0; XGBoost 3.4.1; CatBoost 1.2.10 |

### F. Performance Evaluation

The CVD-positive class was treated as the positive class. Accuracy, precision, recall (sensitivity), specificity, F1-score (for the CVD-positive class), ROC-AUC and PR-AUC (average precision, CVD-positive class) were computed on the test set, and confusion-matrix counts were recorded (TP, FN, FP and TN). Specificity is TN/(TN + FP), the proportion of CVD-negative patients correctly identified, and balanced accuracy is the mean of sensitivity and specificity (Brodersen et al., 2010). The metric definitions follow Hicks et al. (2022) and Orozco-Arias et al. (2020). Precision-recall curves are more informative than ROC curves when classes are imbalanced (Saito & Rehmsmeier, 2015). Because 87.5% of the test records are CVD-positive, a classifier that always predicts CVD would reach an accuracy of 0.875, a balanced accuracy of 0.50 and a specificity of 0; accuracy alone is therefore not a sufficient indicator, and ROC-AUC, balanced accuracy, sensitivity and specificity are given equal prominence (Monaghan et al., 2021). The test set contains only 25 CVD-negative records, so single-split estimates are uncertain. To assess robustness, the whole pipeline was repeated on 20 further stratified splits (Table 2), and the paired differences between the RFECV-subset and all-feature conditions were summarised descriptively as means, standard deviations and win/loss counts. These repeated splits overlap in their training data and are not independent, so no significance tests were applied.

## IV. RESULTS AND DISCUSSION

### A. Choice of RFECV Estimator and Cross-Validation Scheme

Table 3 compares four RFECV configurations on the primary split. With TimeSeriesSplit and the Extra Trees Classifier, the ROC-AUC could not be computed for the internal folds, because



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

the last two validation folds contained no CVD-positive records (Table 4); RFECV consequently defaulted to a single feature. With the regressor, both cross-validation schemes returned larger subsets that included Gender, Diabetes, Vitamin D deficiency and Proteinuria. The stratified classifier configuration returned the most compact subset.

**Table 3.** RFECV Configurations Compared on the Primary Split (balanced training set, 1,372 records).

| Estimator | Cross-validation | n | Features selected | Test ROC-AUC* |
|---|---|---|---|---|
| **Classifier** | Stratified 5-fold | 5 | Age, Smoke, Weight, SBP, DBP | 0.799 |
| **Classifier** | TimeSeriesSplit(5) | 1 | Weight (internal CV scores undefined) | 0.736 |
| **Regressor** | Stratified 5-fold | 9 | Age, Gender, Smoke, Def_in_VitD, Diabetes, Weight, SBP, DBP, Proteinuria | 0.729 |
| **Regressor** | TimeSeriesSplit(5) | 8 | Age, Gender, Smoke, Def_in_VitD, Diabetes, Weight, SBP, DBP | 0.731 |
| **All 14 features** | — | 14 | — | 0.704 |

*Test ROC-AUC of a default Extra Trees Classifier trained on the selected subset (a quick comparison, not the tuned models of Section 3.4).

**Table 4.** Class composition of the validation folds in the RFECV cross-validation of the balanced training set (CVD-positive / CVD-negative).

| Fold | TimeSeriesSplit(5) | Stratified 5-fold |
|---|---|---|
| **1** | 195 / 33 | 138 / 137 |
| **2** | 199 / 29 | 137 / 138 |
| **3** | 86 / 142 | 137 / 137 |
| **4** | 0 / 228 (single class) | 137 / 137 |
| **5** | 0 / 228 (single class) | 137 / 137 |

The comparison was repeated on ten further splits. The classifier selected 4–5 features on every split (mean 4.4), whereas the regressor selected between 4 and 13 features (mean 8.7). When four reference models (LightGBM, Random Forest, XGBoost and CatBoost) were trained on each subset, the mean test ROC-AUC was 0.800 for the classifier-selected subsets, 0.782 for the regressor-selected subsets and 0.785 with all 14 features. The Extra Trees Classifier with stratified cross-validation and ROC-AUC scoring was therefore retained: it gave a stable subset size and aligned the selection criterion with the evaluation metric, whereas the regressor gave unstable subset sizes without a performance gain.

## B. Selected Feature Subset

Figure 2 shows the mean cross-validated ROC-AUC against the number of retained features. The score peaked at five features (0.972), so RFECV reduced the predictor set from 14 to 5 on the primary split. These internal scores are higher than the test-set scores reported below because the internal folds contain synthetic records (Section 2.2). Across the 20 repeated splits, RFECV retained four features (Age, Weight, SBP and DBP) in 16 splits and five in four splits;



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

Age, Weight, SBP and DBP were retained in all 20 splits, whereas Smoking was retained in only one of them (Table 5). The four-feature core is therefore the stable result, and the inclusion of Smoking on the primary split should be regarded as marginal.

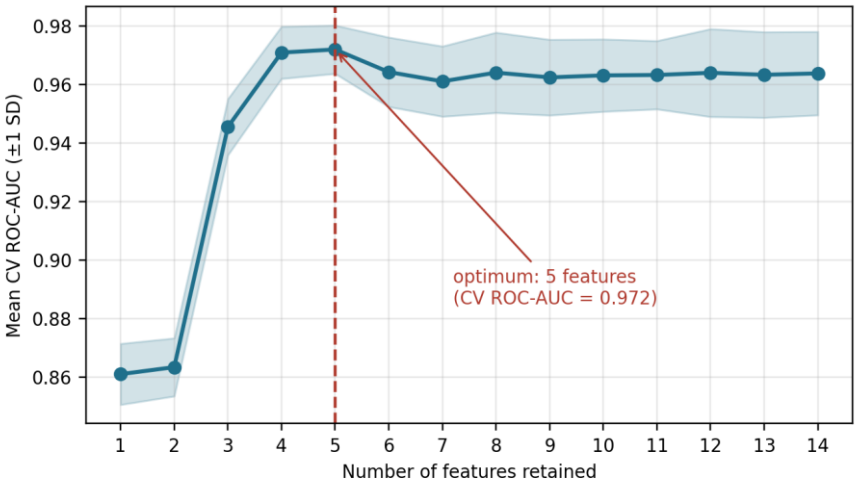



**Figure 2.** Mean Cross-validated ROC-AUC (±1 SD) against the Number of Features Retained by RFECV on the Primary Split

**Table 5.** Predictors, RFECV selection on the primary split and selection frequency over 20 repeated splits.

| Predictor | Retained (primary split) | Importance | Retained in Repeated Splits |
|---|---|---|---|
| **Weight (kg)** | Yes | 0.452 | 20/20 |
| **Systolic BP** | Yes | 0.186 | 20/20 |
| **Age** | Yes | 0.161 | 20/20 |
| **Diastolic BP** | Yes | 0.144 | 20/20 |
| **Smoking** | Yes | 0.057 | 1/20 |
| **Proteinuria** | No | — | 2/20 |
| **Vitamin D deficiency** | No | — | 1/20 |
| **Gender** | No | — | 0/20 |
| **Alcohol** | No | — | 0/20 |
| **Diabetes** | No | — | 0/20 |
| **High cholesterol** | No | — | 0/20 |
| **Chest pain** | No | — | 0/20 |
| **Family history of CVD** | No | — | 0/20 |
| **High blood pressure (indicator)** | No | — | 0/20 |

## C. Ranking of Selected Predictors

Figure 3 ranks the five features selected on the primary split. Weight was the most important predictor (0.452), followed by SBP (0.186), Age (0.161), DBP (0.144) and Smoking (0.057). Mean body weight was 96.9 kg (SD 15.4) in CVD-positive records and 78.7 kg (SD 12.3) in CVD-negative records, and mean SBP was 165.7 mmHg (SD 33.4) and 158.2 mmHg (SD 36.1), respectively. The prominence of weight is consistent with the established link between excess adiposity and cardiovascular risk (Powell-Wiley et al., 2021; Guni et al., 2023; Hristova et al., 2023), that of SBP and DBP with the established association between blood pressure and


https://doi.org/10.5281/zenodo.23410758

cardiovascular risk (Cheng et al., 2023; Lee et al., 2022; Razo et al., 2022; Chen et al., 2024), and that of age with vascular ageing (Zhao et al., 2024; Shi, 2023; Hussain et al., 2024). However, body weight is only a proxy for adiposity, because height was not recorded and BMI could not be computed. The size of the weight effect may also partly reflect how the CVD-negative records were drawn from a hospital population, so it should not be read as a population-level effect size.

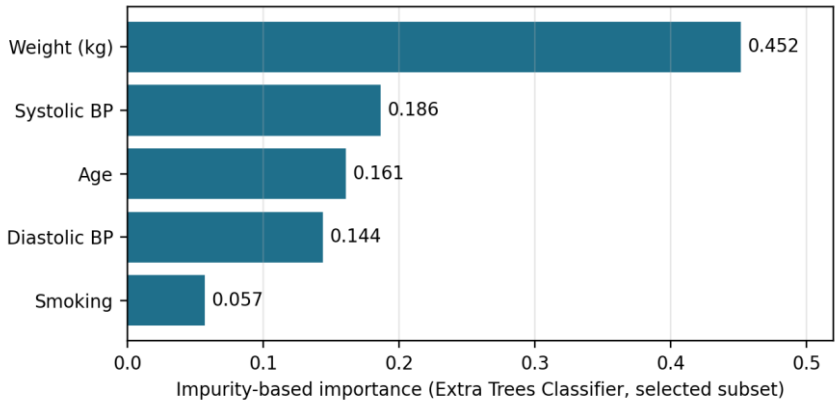



**Figure 3.** Impurity-based Importance of the Five Features Selected by RFECV on the Primary Split

Several variables that are established CVD risk factors were not retained, namely diabetes, high cholesterol, family history, proteinuria and alcohol use (Herrington et al., 2016; Antoine et al., 2020; Katakami et al., 2018; Florescu et al., 2020; Stephen et al., 2014; Kelly et al., 2023; Parmar et al., 2023). Their exclusion means that they added little predictive information beyond the retained variables in this dataset, not that they lack clinical relevance. For example, the HBP indicator was positive in 999 of the 1,000 records and therefore carried almost no information, and it is in any case implied by SBP and DBP.

## D. Effect of Feature Selection on Model Performance

Tables 6 and 7 and Figure 4 give the test-set results on the primary split, with all 14 features (BFS, before feature selection) and with the RFECV subset (AFS, after feature selection). Precision, recall and F1-score refer to the CVD-positive class.

ROC-AUC increased after feature selection for all ten models, with the largest gains for KNN (0.607 to 0.785), SVM (0.704 to 0.798) and Extra Trees (0.701 to 0.795); the smallest gains were for AdaBoost (+0.019), LightGBM (+0.024) and Gradient Boosting and XGBoost (+0.025 each). Specificity also rose for all ten models, and balanced accuracy and PR-AUC rose for nearly all. The other metrics did not improve in the same way. Accuracy was unchanged for Extra Trees, higher for KNN (0.680 to 0.785), and lower for the remaining eight models (by 0.005 for XGBoost up to 0.050 for AdaBoost). Recall and F1-score fell for every model except KNN. The best-performing model after selection by ROC-AUC was CatBoost (0.817), followed by AdaBoost and Random Forest (0.811 each); before selection it was AdaBoost (0.792).

**Table 6.** Accuracy, precision, recall (sensitivity) and F1-score on the test set (n = 200). BFS = all 14 features; AFS = RFECV subset.

| Model | Acc. BFS | Acc. AFS | Prec. BFS | Prec. AFS | Recall BFS | Recall AFS | F1 BFS | F1 AFS |
|---|---|---|---|---|---|---|---|---|



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **LightGBM** | 0.860 | 0.850 | 0.911 | 0.914 | 0.931 | 0.914 | 0.921 | 0.914 |
| **SVM** | 0.835 | 0.820 | 0.913 | 0.954 | 0.897 | 0.834 | 0.905 | 0.890 |
| **Extra Trees** | 0.825 | 0.825 | 0.898 | 0.917 | 0.903 | 0.880 | 0.900 | 0.898 |
| **CatBoost** | 0.860 | 0.850 | 0.911 | 0.934 | 0.931 | 0.891 | 0.921 | 0.912 |
| **Random Forest** | 0.845 | 0.825 | 0.919 | 0.932 | 0.903 | 0.863 | 0.911 | 0.896 |
| **Gradient Boosting** | 0.850 | 0.835 | 0.914 | 0.933 | 0.914 | 0.874 | 0.914 | 0.903 |
| **XGBoost** | 0.845 | 0.840 | 0.914 | 0.939 | 0.909 | 0.874 | 0.911 | 0.905 |
| **KNN** | 0.680 | 0.785 | 0.894 | 0.946 | 0.720 | 0.800 | 0.797 | 0.867 |
| **Decision Tree** | 0.770 | 0.755 | 0.896 | 0.909 | 0.834 | 0.800 | 0.864 | 0.851 |
| **AdaBoost** | 0.820 | 0.770 | 0.954 | 0.957 | 0.834 | 0.771 | 0.890 | 0.854 |

**Table 7.** Specificity, balanced accuracy, ROC-AUC and PR-AUC on the test set (n = 200).

| **Model** | **Spec. BFS** | **Spec. AFS** | **Bal. acc. BFS** | **Bal. acc. AFS** | **ROC-AUC BFS** | **ROC-AUC AFS** | **PR-AUC BFS** | **PR-AUC AFS** |
|---|---|---|---|---|---|---|---|---|
| **LightGBM** | 0.360 | 0.400 | 0.646 | 0.657 | 0.741 | 0.765 | 0.949 | 0.955 |
| **SVM** | 0.400 | 0.720 | 0.649 | 0.777 | 0.704 | 0.798 | 0.925 | 0.961 |
| **Extra Trees** | 0.280 | 0.440 | 0.591 | 0.660 | 0.701 | 0.795 | 0.928 | 0.959 |
| **CatBoost** | 0.360 | 0.560 | 0.646 | 0.726 | 0.780 | 0.817 | 0.942 | 0.966 |
| **Random Forest** | 0.440 | 0.560 | 0.671 | 0.711 | 0.775 | 0.811 | 0.950 | 0.962 |
| **Gradient Boosting** | 0.400 | 0.560 | 0.657 | 0.717 | 0.760 | 0.785 | 0.949 | 0.957 |
| **XGBoost** | 0.400 | 0.600 | 0.654 | 0.737 | 0.778 | 0.803 | 0.954 | 0.960 |
| **KNN** | 0.400 | 0.680 | 0.560 | 0.740 | 0.607 | 0.785 | 0.915 | 0.948 |
| **Decision Tree** | 0.320 | 0.440 | 0.577 | 0.620 | 0.574 | 0.628 | 0.892 | 0.905 |
| **AdaBoost** | 0.720 | 0.760 | 0.777 | 0.766 | 0.792 | 0.811 | 0.956 | 0.959 |



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

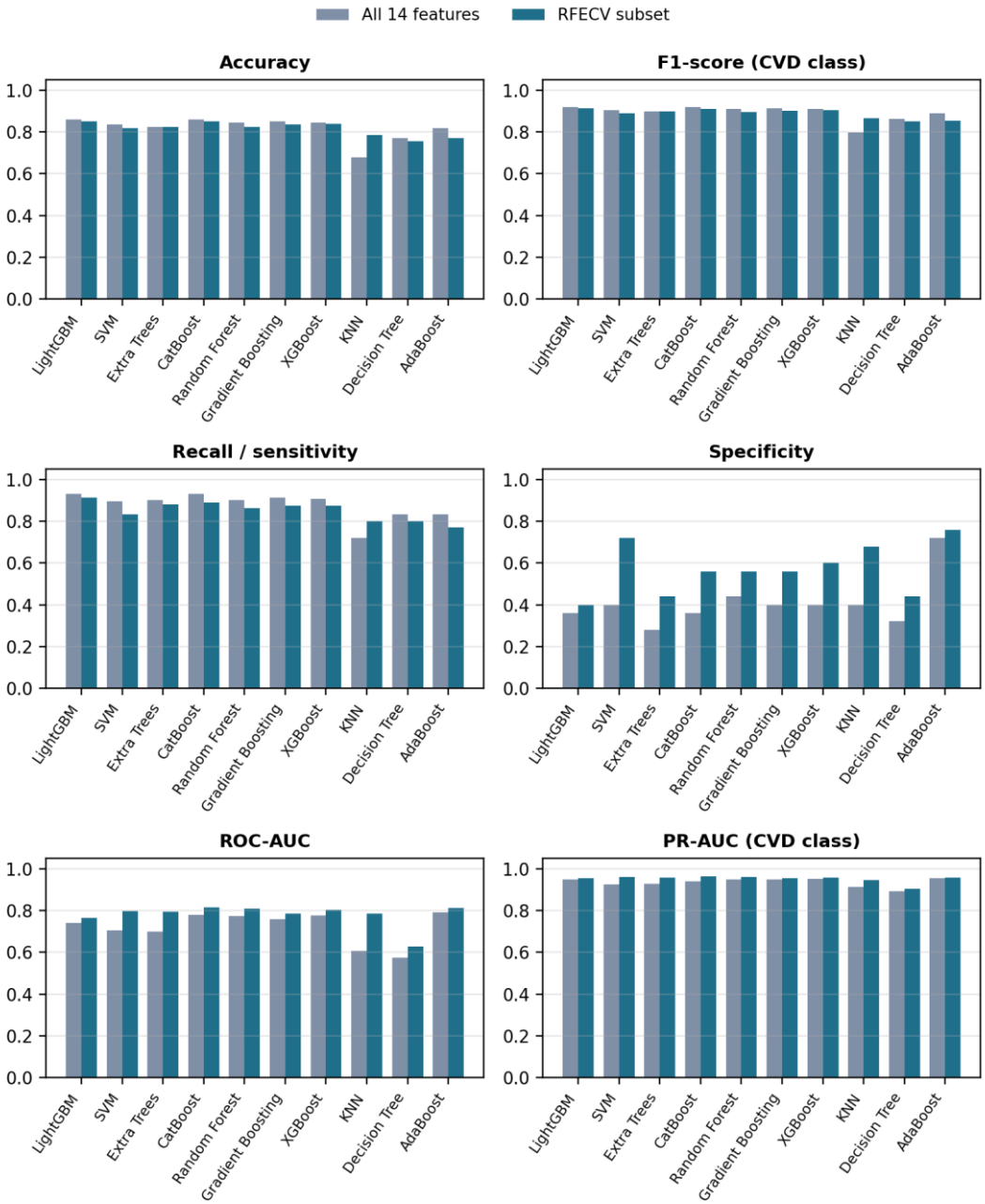



**Figure 4.** Test-set Performance of the Ten Classifiers with all 14 Features and with the RFECV Subset.

These results do not support the view that RFECV universally improves model performance. The effect depended on the model and on the metric: discrimination (ROC-AUC) and the identification of CVD-negative patients (specificity) improved, whereas sensitivity, accuracy and F1-score generally declined slightly. All accuracies (0.68–0.86) were below the 0.875 obtained by always predicting CVD, so the models mainly add value through discrimination rather than raw accuracy, and ROC-AUC values of 0.63–0.82 after selection indicate moderate discrimination at best.

The repeated-split analysis (Table 8, Figure 5) showed the same pattern. Averaged over the ten models, feature selection changed ROC-AUC by +0.037, specificity by +0.117, balanced accuracy by +0.044, recall by −0.030, F1-score by −0.010 and accuracy by −0.011. The ROC-AUC change was negligible for LightGBM (+0.003; higher in only 9 of 20 splits), and the



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

largest changes were for KNN (+0.109), Extra Trees (+0.071) and SVM (+0.047). The boosted-tree models (LightGBM, CatBoost, XGBoost and Gradient Boosting) benefited least, whereas KNN, SVM and Extra Trees benefited most; a plausible explanation is that distance-based and kernel methods are more sensitive to uninformative inputs, but this was not tested directly.

**Table 8.** Paired Differences over 20 Repeated Stratified Splits

| Model | Δ ROC-AUC (wins) | Δ Specificity | Δ Recall | Δ F1-score | Δ Accuracy |
|---|---|---|---|---|---|
| **LightGBM** | +0.003 ± 0.036 (9/20) | +0.056 ± 0.121 | −0.029 ± 0.026 | −0.013 ± 0.017 | −0.019 ± 0.028 |
| **SVM** | +0.047 ± 0.043 (18/20) | +0.202 ± 0.115 | −0.053 ± 0.032 | −0.018 ± 0.021 | −0.021 ± 0.032 |
| **Extra Trees** | +0.071 ± 0.054 (17/20) | +0.168 ± 0.105 | −0.021 ± 0.030 | −0.001 ± 0.015 | +0.003 ± 0.023 |
| **CatBoost** | +0.011 ± 0.035 (11/20) | +0.078 ± 0.103 | −0.031 ± 0.024 | −0.013 ± 0.012 | −0.018 ± 0.018 |
| **Random Forest** | +0.012 ± 0.039 (15/20) | +0.082 ± 0.092 | −0.023 ± 0.030 | −0.008 ± 0.018 | −0.010 ± 0.027 |
| **Gradient Boosting** | +0.017 ± 0.042 (12/20) | +0.082 ± 0.112 | −0.031 ± 0.024 | −0.012 ± 0.013 | −0.017 ± 0.020 |
| **XGBoost** | +0.020 ± 0.039 (13/20) | +0.082 ± 0.102 | −0.037 ± 0.029 | −0.016 ± 0.016 | −0.022 ± 0.025 |
| **KNN** | +0.109 ± 0.062 (18/20) | +0.154 ± 0.133 | +0.003 ± 0.049 | +0.012 ± 0.029 | +0.022 ± 0.040 |
| **Decision Tree** | +0.042 ± 0.062 (16/20) | +0.122 ± 0.149 | −0.020 ± 0.046 | −0.004 ± 0.025 | −0.002 ± 0.036 |
| **AdaBoost** | +0.039 ± 0.029 (18/20) | +0.146 ± 0.069 | −0.054 ± 0.023 | −0.024 ± 0.016 | −0.029 ± 0.023 |

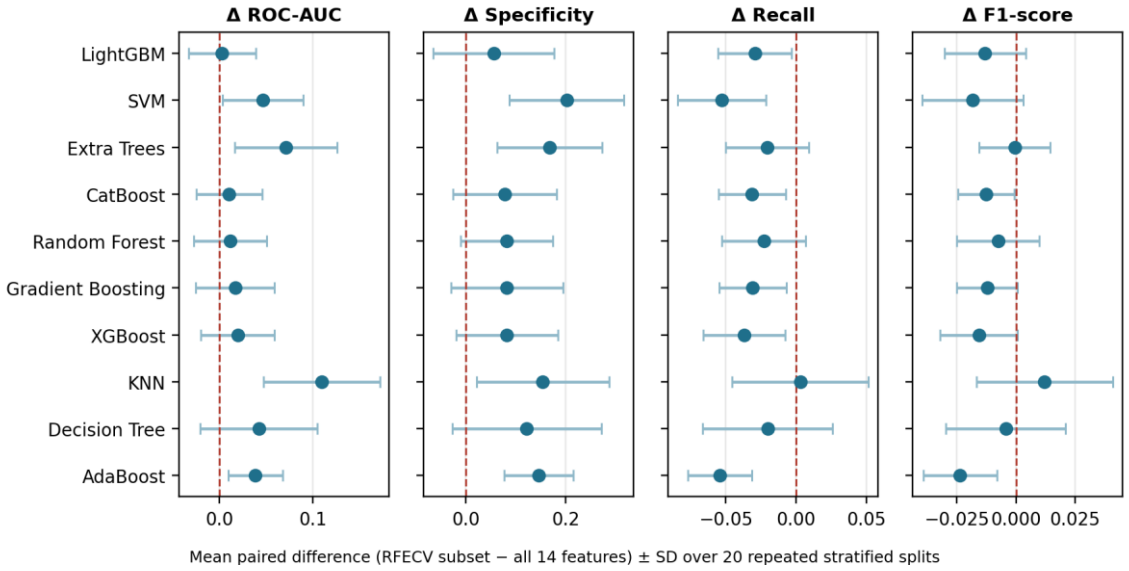



**Figure 5.** Mean paired Differences (RFECV subset minus all 14 features) ± SD over 20 Repeated Stratified Splits



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

Paired differences (RFECV subset minus all 14 features), mean ± SD over 20 repeated stratified splits; the number of splits in which ROC-AUC was higher after selection is shown in parentheses.

## E. Confusion-Matrix Analysis

Figures 6 and 7 and Table 9 show the confusion matrices. A false negative is a CVD patient classified as CVD-negative, and a false positive is a CVD-negative patient classified as CVD-positive.

**Table 9.** Confusion-matrix Counts on the Test Set (175 CVD-positive and 25 CVD-negative records).

| Model | TP BFS | TP AFS | FN BFS | FN AFS | FP BFS | FP AFS | TN BFS | TN AFS |
|---|---|---|---|---|---|---|---|---|
| **LightGBM** | 163 | 160 | 12 | 15 | 16 | 15 | 9 | 10 |
| **SVM** | 157 | 146 | 18 | 29 | 15 | 7 | 10 | 18 |
| **Extra Trees** | 158 | 154 | 17 | 21 | 18 | 14 | 7 | 11 |
| **CatBoost** | 163 | 156 | 12 | 19 | 16 | 11 | 9 | 14 |
| **Random Forest** | 158 | 151 | 17 | 24 | 14 | 11 | 11 | 14 |
| **Gradient Boosting** | 160 | 153 | 15 | 22 | 15 | 11 | 10 | 14 |
| **XGBoost** | 159 | 153 | 16 | 22 | 15 | 10 | 10 | 15 |
| **KNN** | 126 | 140 | 49 | 35 | 15 | 8 | 10 | 17 |
| **Decision Tree** | 146 | 140 | 29 | 35 | 17 | 14 | 8 | 11 |
| **AdaBoost** | 146 | 135 | 29 | 40 | 7 | 6 | 18 | 19 |

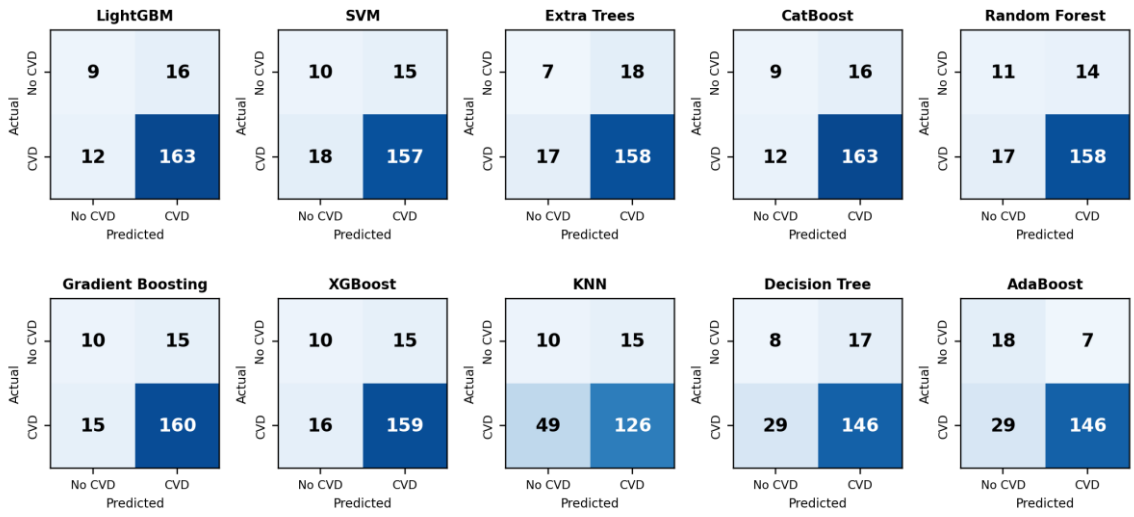



**Figure 6.** Confusion Matrices before Feature Selection (all 14 features)



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758

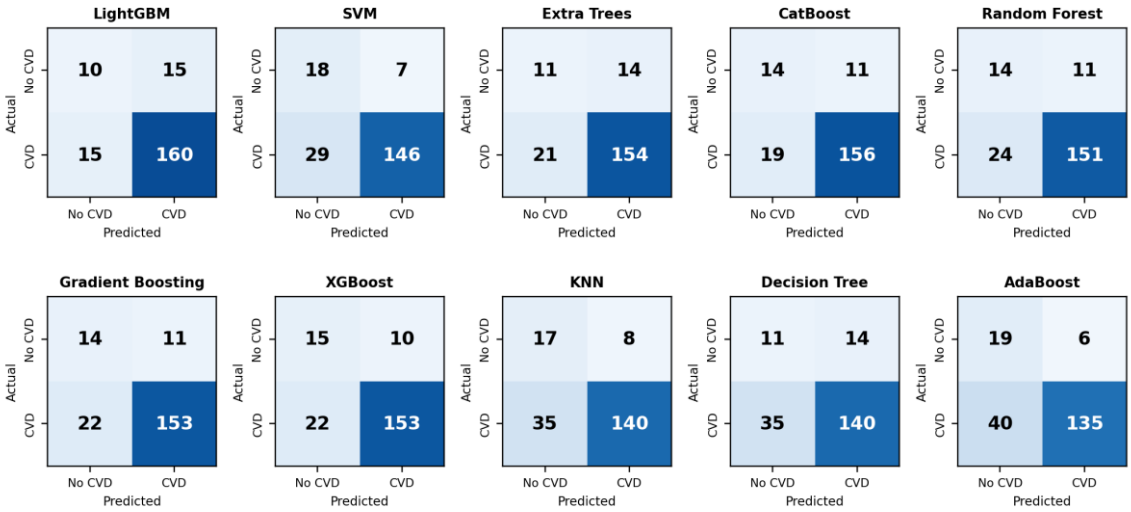



**Figure 7.** Confusion Matrices after Feature Selection (RFECV subset).

Feature selection reduced false positives for all ten models (for example, SVM from 15 to 7 and CatBoost from 16 to 11), which corresponds to the higher specificity. It increased false negatives for nine models (for example, LightGBM from 12 to 15 and AdaBoost from 29 to 40); only KNN reduced them (from 49 to 35). In a screening context, missed CVD cases are generally more costly than false alarms, so the subset should not be described as clinically superior on this evidence. The balance between sensitivity and specificity depends on the decision threshold, which was fixed at 0.5 here; threshold calibration or cost-sensitive learning would be needed before any deployment.

### F. Potential Clinical Implications and Limitations

A model that needs four or five inputs is simpler to interpret and to implement, and Age, Weight, SBP and DBP are routine bedside measurements. A smaller predictor set could therefore, in principle, reduce data collection and reporting effort. This is a potential implication only. The removal of a predictor variable from a model does not imply that the corresponding clinical measurement is not needed in practice since the clinical measurement is also used for making other diagnoses, risk stratification and treatment decisions that are not covered by the model. Any reduction in tests, cost or time would have to be established through external and prospective validation, following the TRIPOD reporting guidance (Collins et al., 2015), and with clinical review of which measurements can safely be omitted. The study has several limitations. The data come from a single hospital and contain only 1,000 records, of which 126 are CVD-negative and only 25 fall in the test set, so the estimates are imprecise. The records were abstracted retrospectively, and the CVD-negative group may not be representative of the general population, which may exaggerate the importance of variables such as weight. Height was not recorded, so BMI could not be examined. The internal RFECV scores are optimistic because the balanced training folds contain synthetic records, and the classifier hyperparameters in the repeated analysis were fixed at the values tuned on the primary split. Finally, no external validation, threshold optimisation or probability calibration was performed.

## V. CONCLUSION

This study applied RFECV with an Extra Trees Classifier and stratified cross-validation to a UATH clinical dataset, using a leakage-free pipeline in which the train/test split precedes


https://doi.org/10.5281/zenodo.23410758

balancing, feature selection and model training. Of the 14 predictors, RFECV narrowed it down to five, with 4 remaining stable across 20 repeated splits (Age, Weight, SBP and DBP). In terms of feature selection, the balanced accuracy and specificity scores were generally increased, while sensitivity scores were somewhat decreased, with the exception of several classifiers such as Artificial Neural Networks, which were not affected; the F1-score increased for most of the ten classifiers, and the ROC-AUC score increased for most of the classifiers. The scores of specificity and balanced accuracy increased in most or all of the classifiers, whereas the scores of sensitivity and accuracy decreased in most of the classifiers, and the score of false positive decreased while the score of false positive increased for most of the classifiers. The effect of RFECV is thus model and metric dependent and should not be claimed as a general improvement. Though the smaller predictor set is appealing due to interpretability and simplicity, any decrease in clinical testing, cost or time would be an implication that would need clinical validation. Future studies should validate the selected subset on independent multi-centre cohorts with height measured to compare the BMI and weight; tune the decision thresholds and calibrate the probabilities with sensitivity in mind; resample within each fold of the CV; and compare RFECV with other (embedded and stability-based) selection methods.

**Declarations**

**Ethical Approval:** Ethical clearance and permission to use anonymised patient records for research was obtained from the UATH Ethics Committee.
**Funding:** This research was self-funded.

## REFERENCES

Akhter, S., & Miller, J. H. (2026). Web-based cardiovascular disease risk prediction using machine learning. Frontiers in Artificial Intelligence, 9, 1690664. https://doi.org/10.3389/frai.2026.1690664

Alfebi, F. H., & Anasanti, M. D. (2023). Improving cardiovascular disease prediction by integrating imputation, imbalance resampling, and feature selection techniques into machine learning model. Indonesian Journal of Computing and Cybernetics Systems, 17(1), 55-66. https://doi.org/10.22146/ijccs.80214

Antoine, M. L. K., Xu, Y., & Djakpo, D. K. (2020). Atherosclerotic disease and diabetes mellitus. International Journal of Medicine, 8(1), 4-7. https://doi.org/10.14419/ijm.v8i1.30213

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In Proceedings of the 20th International Conference on Pattern Recognition (pp. 3121–3124). IEEE. https://doi.org/10.1109/ICPR.2010.764

Chakraborty, A., Das, U. K., Sazzad, S., Das, P., & Khan, M. M. H. (2026). Explainable machine learning for early heart disease risk prediction: Insights from a clinical dataset in Bangladesh. Intelligence-Based Medicine, 13, 100352. https://doi.org/10.1016/j.ibmed.2026.100352

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, C., Huang, Y., Wu, S., & Feng, Y. (2024). Association of blood pressure trajectories since young adulthood with cardiovascular diseases in later life. Journal of Hypertension, 42(Suppl 1).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cheng, Y.-J., Luo, D.-L., Bi, W.-T., Mei, W.-Y., Wu, S.-H., Li, Z.-Y., & Xie, J.-D. (2023). Systolic blood pressure and cardiovascular risk in normotensive adults. Mayo Clinic Proceedings, 98(3), 386-397. https://doi.org/10.1016/j.mayocp.2022.10.011

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ, 350, g7594. https://doi.org/10.1136/bmj.g7594

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Darji, P. A. (2024). Utilizing an ensemble of extra tree model for classifying mesothelioma cancer. African Journal of Biological Sciences, 6(12), 535-545. https://doi.org/10.48047/AFJBS.6.12.2024.535-545

Enihe, R. O., Prasad, R., Ogwueleka, F. N., & Abdullahi, F. B. (2025). The effect of imbalance data mitigation techniques on cardiovascular disease prediction. Journal of the Nigerian Society of Physical Sciences, 7(2), 2385. https://doi.org/10.46481/jnsps.2025.2385

Florescu, M. C., Youssef, I., Shenoy, A., & Hawkins, J. L. (2020). Proteinuria and albuminuria as CVD markers in diabetes and chronic kidney disease: Evaluation and management (pp. 127-148). https://doi.org/10.2174/9789811422003120010012

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatius, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. IEEE Access, 9, 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759

Guni, M., Karanxha, J., & Dede, M. (2023). Impact of BMI in early outcomes in patients undergoing on CABG. International Journal of Science and Research, 12(11), 1099-1102. https://doi.org/10.21275/sr231115162034

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Hassan, F. H., Wang, S., & Miron, A. (2026). Machine learning for predicting coronary heart disease risk in patients with hypertension: An ensemble modeling approach. Healthcare Informatics Research, 32(1), 28-37. https://doi.org/10.4258/hir.2026.32.1.28

Herrington, W., Lacey, B., Sherliker, P., Armitage, J., & Lewington, S. (2016). Epidemiology of atherosclerosis and the potential to reduce the global burden of atherothrombotic

disease. Circulation Research, 118(4), 535-546. https://doi.org/10.1161/CIRCRESAHA.115.307611

Hicks, S. A., Strumke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979. https://doi.org/10.1038/s41598-022-09954-8

Hristova, P., & Platikanova, M. (2023). Body Mass Index (BMI) - Predictor of diseases onset and regulator of prevention? Journal of IMAB, 29(2), 4947-4950. https://doi.org/10.5272/jimab.2023292.4947

Hussain, M. M., Rafi, U., Imran, A., Rehman, M. U., & Abbas, S. K. (2024). Risk factors associated with cardiovascular disorders. Pakistan BioMedical Journal, 03-10. https://doi.org/10.54393/pbmj.v7i02.1034

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Katakami, N. (2018). Mechanism of development of atherosclerosis and cardiovascular disease in diabetes mellitus. Journal of Atherosclerosis and Thrombosis, 25(1), 27-39. https://doi.org/10.5551/jat.RV17014

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (Vol. 30, pp. 3146–3154).

Kelly, J. T., Gonzalez-Ortiz, A., St-Jules, D. E., & Carrero, J. J. (2023). Animal protein intake and possible cardiovascular risk in people with chronic kidney disease: Mechanisms and evidence. Advances in Kidney Disease and Health, 30(6), 480-486. https://doi.org/10.1053/j.akdh.2023.06.003

Khawla, R., Talukder, M. A., Layek, M. A., Das, U. K., Alazab, A., & Kazi, M. (2026). Enhanced cardiovascular disease risk prediction using explainable machine learning and data balancing. Discover Computing, 29(1), 100. https://doi.org/10.1007/s10791-026-09973-3

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (Vol. 2, pp. 1137–1143).

Lee, J. Y., Hong, J. H., Lee, S., An, S., Shin, A., & Park, S. K. (2022). Binary cutpoint and the combined effect of systolic and diastolic blood pressure on cardiovascular disease mortality: A community-based cohort study. PLoS ONE, 17(6). https://doi.org/10.1371/journal.pone.0270510

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1–5.

Li, J., Xiang, W., Shang, D., Li, S., & Li, Q. (2026). Federated learning for cardiovascular disease prediction: A systematic review of clinical applications, validation, and translation readiness. Frontiers in Cardiovascular Medicine, 13, 1831342. https://doi.org/10.3389/fcvm.2026.1831342

Louppe, G. (2014). Understanding random forests: From theory to practice [Doctoral dissertation, University of Liège]. arXiv:1407.7502.

Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., & Dmochowski, R. R. (2021). Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (Kaunas, Lithuania), 57, 503. https://doi.org/10.3390/medicina57050503

Orozco-Arias, S., Pina, J. S., Tabares-Soto, R., Castillo-Ossa, L. F., Guyot, R., & Isaza, G. (2020). Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. Processes, 8, 638. https://doi.org/10.3390/pr8060638

Parmar, M. P., Kaur, M., Bhavanam, S., Mulaka, G. S. R., Ishfaq, L., Vempati, R., C, M. F., Kandepi, H. V., ER, R., Sahu, S., & Davalgi, S. (2023). A systematic review of the effects of smoking on the cardiovascular system and general health. Cureus, 15(4), e38073. https://doi.org/10.7759/cureus.38073

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Pidishetty, H., & Bukaita, W. (2026). Human-AI clinical decision support for heart disease risk prediction using interpretable and reliable machine learning. Medical Research Archives, 14(5). https://doi.org/10.18103/mra.v14i5.7474

Powell-Wiley, T. M., Poirier, P., Burke, L. E., Després, J.-P., Gordon-Larsen, P., Lavie, C. J., et al. (2021). Obesity and cardiovascular disease: A scientific statement from the American Heart Association. Circulation, 143(21), e984–e1010. https://doi.org/10.1161/CIR.0000000000000973

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems (Vol. 31, pp. 6638–6648).

Razo, C., Welgan, C. A., Johnson, C. O., McLaughlin, S. A., Iannucci, V., Rodgers, A., Wang, N., LeGrand, K. E., Sorensen, R. J. D., He, J., Zheng, P., Aravkin, A. Y., Hay, S. I., Murray, C. J. L., & Roth, G. A. (2022). Effects of elevated systolic blood pressure on ischemic heart disease: a Burden of Proof study. Nature Medicine, 28(10), 2056-2065. https://doi.org/10.1038/s41591-022-01974-1

Riswanda, M. A., Abadi, F., Mazdadi, M. I., Faisal, M. R., & Herteno, R. (2026). The effect of Smote-Tomek on the classification of chronic diseases based on health and lifestyle data. Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, 8(1), 130-140. https://doi.org/10.35882/ijeeemi.v8i1.324

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. IEEE Computational Intelligence Magazine, 13(4), 59–76. https://doi.org/10.1109/MCI.2018.2866730

Shi, P. (2023). Treatments and prevention of CVDs by taking an insight into the mechanism of cardiovascular aging. Highlights in Science, Engineering and Technology. https://doi.org/10.54097/h2m33r10

Stephen, R., Jolly, S. E., Nally, J. V., & Navaneethan, S. D. (2014). Albuminuria: When urine predicts kidney and cardiovascular disease. Cleveland Clinic Journal of Medicine, 81(1), 41-50. https://doi.org/10.3949/ccjm.81a.13040

Tabbassum, S., Manoj, V. V. R., Praveen, S. P., Lakshmi, A. N., Reddy, B. R., & Muthukumar, P. (2026). Heart disease prediction for enhanced cardiovascular health management by using machine learning algorithms: A cross-sectional study. Premier Journal of Science, 15, 100209. https://doi.org/10.70389/PJS.100209

Tomek, I. (1976). Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452

Tsiligaridis, J. (2023). Tree-based ensemble models, algorithms and performance measures for classification. Advances in Science, Technology and Engineering Systems Journal, 8(6), 19-25. https://doi.org/10.25046/aj080603

Uddin, N., Mahmud, P., Ahammed, M., Uddin, M. A., Kazi, M., & Talukder, M. A. (2026). Explainable ensemble learning model for cardiovascular disease prediction with feature optimization and data balancing. Discover Computing, 29(1), 44. https://doi.org/10.1007/s10791-026-09930-0

Wan, C. (2019). Feature selection paradigms. In Hierarchical Feature Selection for Knowledge Discovery. Springer. https://doi.org/10.1007/978-3-319-97919-9_3

World Health Organization. (2021). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yuliana, Y., Robet, R., & Hoki, L. (2026). Comparative analysis of XGBoost, KNN, and SVM algorithms for heart disease prediction using SMOTE-Tomek balancing. Sinkron: Jurnal dan Penelitian Teknik Informatika, 10(1), 305-314. https://doi.org/10.33395/sinkron.v10i1.15469

Zhao, D., Wang, Y., Wong, N. D., & Wang, J. (2024). Impact of aging on cardiovascular diseases: From chronological observation to biological insights. JACC: Asia, 4(5), 345-358. https://doi.org/10.1016/j.jacasi.2024.02.002

## Appendix A. Hyperparameter Search Grids and Selected Values

**Table A1.** Search grids and selected hyperparameters (all other parameters at library defaults; random_state = 42).

| Model | Search grid | Selected: all 14 features | Selected: RFECV subset |
|---|---|---|---|
| **LightGBM** | num_leaves {31, 41}; learning_rate {0.05, 0.1}; n_estimators {100, 150} | learning_rate = 0.1; n_estimators = 150; num_leaves = 41 | learning_rate = 0.1; n_estimators = 150; num_leaves = 41 |
| **SVM** | C {0.1, 1, 10}; gamma {scale, auto} | C = 10; gamma = scale | C = 10; gamma = scale |
| **Extra Trees** | n_estimators {100, 150}; max_depth {None, 10, 20}; min_samples_split {2, 5} | max_depth = None; min_samples_split = 2; n_estimators = 150 | max_depth = 20; min_samples_split = 2; n_estimators = 150 |
| **CatBoost** | iterations {100, 200}; learning_rate {0.01, 0.1}; depth {4, 6, 8} | depth = 8; iterations = 200; learning_rate = 0.1 | depth = 8; iterations = 200; learning_rate = 0.1 |
| **Random Forest** | n_estimators {100, 150}; max_depth {None, 10, 20}; min_samples_split {2, 5} | max_depth = 20; min_samples_split = 2; n_estimators = 150 | max_depth = 20; min_samples_split = 2; n_estimators = 150 |
| **Gradient Boosting** | learning_rate {0.05, 0.1}; n_estimators {100, 150}; max_depth {3, 4} | learning_rate = 0.1; max_depth = 4; n_estimators = 150 | learning_rate = 0.1; max_depth = 4; n_estimators = 150 |
| **XGBoost** | learning_rate {0.05, 0.1}; n_estimators {100, 150}; max_depth {3, 4} | learning_rate = 0.1; max_depth = 4; n_estimators = 150 | learning_rate = 0.1; max_depth = 4; n_estimators = 150 |
| **KNN** | n_neighbors {3, 5, 7}; weights {uniform, distance} | n_neighbors = 7; weights = distance | n_neighbors = 7; weights = distance |
| **Decision Tree** | max_depth {None, 10, 20}; min_samples_split {2, 5} | max_depth = None; min_samples_split = 5 | max_depth = 10; min_samples_split = 5 |
| **AdaBoost** | n_estimators {50, 100}; learning_rate {0.01, 0.1, 1} | learning_rate = 1; n_estimators = 100 | learning_rate = 1; n_estimators = 100 |



Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.23410758