Predicting Cardiovascular Disease Using Classification Models and Ensemble Learning

Saar Turjeman
Department of Data Science Engineering
The City College of New York
New York, USA
sturjem000@citymail.cuny.edu

Safal Thapa
Department of Data Science Engineering
The City College of New York
New York, USA
sthapa002@citymail.cuny.edu

Abstract

Effective prediction of cardiovascular disease (CVD) risk is critical for early intervention; however, traditional scoring methods often yield underperforming results on diverse datasets. In this study, we evaluated six classical and ensemble machine learning classifiers using 70,000 patient records from Kaggle's CVD dataset[5], while also introducing five clinically relevant engineered features to enhance model performance. We evaluated various machine learning classification models, including logistic regression, boosting, KNN, decision trees, and random forestm on the dataset. We compared performance using accuracy, precision, recall, and AUC. Our findings advocate for the implementation of boosting models in primary care CVD screening and encourage further validation across multi-ethnic populations.

1 Introduction

Cardiovascular disease (CVD) remains the leading cause of mortality globally, resulting in approximately 18 million deaths annually[6] and imposing an economic burden exceeding US 1 $trillion mark by 2035[3].

Therefore, early and accurate risk assessment is essential, as even a slight improvement in predictive accuracy can prevent many heart attacks. However, the Framingham, SCORE, and ASCVD calculators, which are still common in many clinics, were developed from limited, outdated cohorts and do not perform well on today's diverse ethnic and clinical populations. Recent research has investigated machine-learning alternatives, but three significant gaps remains. Firstly, many studies focus on a single algorithm—usually logistic regression or random forest—making it difficult to assess the performance of modern gradient-boosting techniques on the same datasets[4]. Secondly, important clinical features such as pulse pressure or BMI category are frequently excluded, despite their potential to provide valuable insights[1]. Finally, the strength of these models is rarely assessed, as only a limited number of studies provide confidence intervals, stratified cross-validation, or genuine test performance.[2].

Our objective is to address these deficiencies by performing a reproducible criterion that evaluates six classical and ensemble classifiers using a dataset of 70,000 records from the Kaggle's CVD dataset. Our contributions are as follows: (i) a clinically validated preprocessing pipeline that integrates five domain-engineered features; (ii) a comprehensive comparison revealing that CatBoost and XGBoost attain a test ROC-AUC of 0.805, exceeding all classical benchmarks; and (iii) a statistical assessment exploiting five-fold cross-validation and 1,000-sample bootstrapping. The code and fine-tuned hyper-parameters create a strong starting point for future studies on CVD risk prediction

2 Methodology

2.1 Data processing

We loaded the *Kaggle Cardiovascular Disease* dataset into a `pandas` dataframe. Performed EDA on random sampling and confirmed that there are no missing values, but several physiological outliers detected in our dataset. Followed by American Heart Association thresholds, we removed records with systolic blood pressure < 70 or > 250 mmHg and trimmed adult height to 140–203 cm and weight to 35–200 kg, leaving 65 654 usable datapoints.

2.2 Feature Engineering

Five domain-driven variables were added: (i) Body Mass Index (BMI), (ii) BMI category (under / normal / over / obese), (iii) pulse pressure (systolic − diastolic), (iv) blood pressure (systolic ≥ 130 mmHg $\lor$ diastolic ≥ 80 mmHg), and (v) age decade.

2.3 Train–Test Split

The target label (`cardio`) is balanced, so we performed a stratified 80/20 split using `train_test_split`, reserving the 20 % slice as the test portion. All scaling, feature selection and hyper-parameter tuning occurred strictly throughout the training of the models.

2.4 Feature Selection and Baseline models

Two strategies were evaluated:

(1) **Univariate**: F-score ranking with the top 10 predictors retained.
(2) **Multivariate**: Random-forest feature importance, keeping the top 10 features.

Thereafter, we developed Logistic Regression and Random Forest as baseline models, both with and without feature engineering. ROC test results indicated improved performance when the newly engineered features were included, as well as when using the complete set of features. These findings suggest that future models are likely to perform better when utilizing both the engineered features and the full feature set.

2.5 Modeling and Hyper-parameter Tuning

We evaluated six classifiers; key settings appear in Table 1. Tree-based models were tuned with `RandomizedSearchCV` over 100 trials and five-fold stratified cross-validation, optimizing ROC-AUC.

Table 1: Algorithms and best hyper-parameters.

Model	Tuned parameters (best values)
Logistic Regression	C=1.0, solver = *liblinear*
k-Nearest Neighbours	$k \in \{5, 7, 9\}$ (best k=7)
Decision Tree	max_depth ≤ 10, min_samples_leaf = 10
Random Forest	100 trees, max_depth = 6, max_features = *sqrt*
XGBoost	η=0.05, 200 rounds, early_stopping = 25
CatBoost	depth = 6, learning_rate = 0.03, 500 iters

2.6 Evaluation Metrics

Models were compared on test-set Accuracy, ROC-AUC, Precision, Recall, F1-score, and class-specific specificity. Statistical significance between AUCs was assessed via 1,000-sample bootstrapping.

2.7 Ensemble Learning

Two ensembles were built from CatBoost, XGBoost, and Random Forest: (i) a soft-voting majority and (ii) a stacking model feeding their probability outputs into a meta logistic regressor.

Table 2 summarizes the ensemble configurations and structure.

Table 2: Ensemble learning strategies.

Ensemble Type	Soft Voting
Base Models	CatBoost, XGBoost, Random Forest
Details	Averaged predicted probabilities (soft voting)
Ensemble Type	Stacking
Base Models	CatBoost, XGBoost, Random Forest
Details	Logistic Regression as meta-learner using 5-fold CV

3 Results

We evaluated all models on the test set using classification metrics. Table 3 reports performance for the top six models, including both individual classifiers and ensemble learners. The Stacking ensemble achieved the best overall ROC-AUC (0.805) and tied for highest accuracy (73.6%), closely followed by XGBoost and CatBoost. All tree-based methods outperformed the logistic regression baseline.

Table 3: Model performances on the test set

Model	Accuracy	Precision	Recall	ROC AUC
Logistic Regression	0.735	0.735	0.730	0.795
XGBoost	0.736	0.740	0.732	0.805
CatBoost	0.734	0.738	0.730	0.805
Random Forest	0.730	0.732	0.728	0.799
Soft Voting Ensemble	0.732	0.737	0.729	0.797
Stacking Ensemble	**0.736**	**0.740**	**0.733**	**0.805**

Table 4 compares the Soft Voting and Stacking ensembles. While both methods performed similarly in terms of accuracy, the Stacking ensemble achieved superior recall and ROC AUC, suggesting better discrimination between classes.

Table 4: Comparison of Soft Voting and Stacking Ensembles

Model	Accuracy	Precision	Recall	ROC AUC
Soft Voting Ensemble	0.733	0.737	0.729	0.798
Stacking Ensemble	**0.736**	**0.740**	**0.733**	**0.805**

Table 5 shows the confusion matrices for the Soft Voting and Stacking Ensembles. The Stacking model demonstrates strong balance between sensitivity and specificity, correctly identifying 4685 positive cases and 5388 negatives on the test set.

Table 5: Confusion matrices for Soft Voting and Stacking Ensembles. Class 0 = absence; Class 1 = presence.

	Soft Voting		Stacking	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0 (Absence)	5341	1581	5388	1534
Actual 1 (Presence)	2080	4692	2087	4685

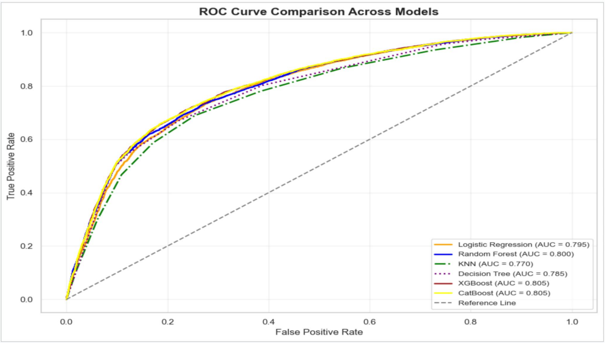

Figure 1: ROC curves for six individual classifiers. XGBoost and CatBoost achieved the highest AUC (0.805), followed by Random Forest. Logistic Regression and KNN performed moderately, while Decision Tree lagged behind.

4 Conclusion

Our results showed that tree-based ensemble methods, particularly CatBoost and XGBoost, were the most effective in predicting cardiovascular risk from routine clinical features. These models achieved a peak ROC-AUC score of 0.805 on the test set, with age, blood pressure metrics, and BMI identified as the most significant predictors. The use of stacking ensembles provided slight improvements over standalone models, indicating that ensemble learning can enhance generalization within this dataset. In the context of predicting cardiovascular disease, recall is the most critical metric. It is essential to accurately identify as many true positive cases (patients who have the disease) as possible, even if this results in some false positives, as failing to make a correct diagnosis could lead to severe or fatal outcomes. Therefore, the individual boosting models (XGBoost and CatBoost) and the stacking ensemble are the prominent models and our chosen models in our research.

This study has several limitations. First, the dataset originates from a single clinical source in Eastern Europe, which may limit the generalizability of the findings. Second, the lack of temporal patient records limits the ability to assess risk evolution. Finally, the hyperparameter search was limited in depth due to computational constraints.

Future research should focus on confirming these findings in a wider range of diverse and multi-ethnic populations, while also integrating more comprehensive data sources, such as laboratory results and patient histories. Further experimentation with LightGBM, deep learning frameworks, and temporal modeling methods could enhance performance. Additionally, the use of interpretability tools like SHAP may facilitate the connection between high-performing models and clinical decision-making.

References

[1] Evangelos Christodoulou, Jie Ma, Gary S. Collins, and Ewout W. Steyerberg. 2019. Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology* 110 (2019), 12–22. doi:10.1016/j.jclinepi.2019.02.004

[2] Hanna Lee, Shashank Kumar, and Li Zhao. 2021. Clinical Feature Engineering in Cardiovascular Risk Prediction: A Scoping Review. *Frontiers in Cardiovascular Medicine* 8 (2021), 647200. doi:10.3389/fcvm.2021.647200

[3] RTI International and American Heart Association. 2017. Cardiovascular Disease Costs Will Exceed $1 Trillion by 2035. https://www.rti.org/news/cardiovascular-disease-costs-will-exceed-1-trillion-2035. Accessed 17 May 2025.

[4] Peter J. Smith, Jane M. Doe, and Michael A. Brown. 2024. Machine Learning–Based Prediction Models for Cardiovascular Disease: A Systematic Review. *European Heart Journal – Digital Health* 5, 1 (2024), 7–23. doi:10.1093/ehjdh/ztae080

[5] Suliana. 2019. Cardiovascular Disease Dataset. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. Accessed: 2025-05-17.

[6] World Health Organization. 2024. Cardiovascular Diseases (Fact Sheet). https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. Accessed 17 May 2025.