Title

Interpretable Ensemble ML Model for Heart Disease Prediction

1. Team Details

1.1 Team Members

Ghazal Rezaee 1— Email: rezaeeghazal3@gmail.com

External Collaborator: Sara Ghazizadeh1— Email: saraghazizadeh77@gmail.com

1.2 Mentor

External Collaborator: Fatemehzahraseyedkolbadi1 —

Email:Fatemehzahraseyedkolbadi@gmail.com

1.3 Affiliation:

Faculty of Medicine, Hormozgan University of Medical Sciences, Bandar Abbas, Iran.

2. Introduction

Cardiovascular diseases continue to be the leading cause of morbidity and mortality globally, causing approximately 17.9 million deaths annually according to the World Health Organization and imposing significant economic burdens on healthcare systems worldwide. The multifactorial etiology of the development of cardiovascular disease with intricate interactions among genetic predisposition, lifestyle factors, environmental exposures, and comorbidities complicates risk prediction and timely intervention within the clinical setting.

Early diagnosis and accurate prediction of risk for heart disease are essential components of ideal cardiovascular care, since early identification of at-risk patients enables application of evidence-based prevention strategies that can markedly restrict cardiovascular events and promote improved long-term patient outcomes. Traditional methods of cardiovascular risk stratification, while clinically helpful, rely on reduced-function risk calculators or global clinical judgment that fail to consistently fully represent the complex, non-linear interplay among many risk factors and the occurrence of cardiovascular disease.

The advancement of machine learning technology in medicine presents unparalleled opportunities for maximizing cardiovascular risk prediction with the help of sophisticated analytical approaches that are able to identify subtle patterns and intricate relationships in large clinical databases. These sophisticated computational methods are likely to surpass the performance of existing risk assessment tools by taking into account multiple clinical variables in combination, revealing non-linear associations between risk factors, and making patient-specific risk estimates that take into account individual patient profiles and clinical presentation.

This project aims to develop a comprehensive, end-to-end machine learning system that directly predicts the likelihood of heart disease onset in patients based on readily available clinical, demographic, and physiological features that are readily collected as part of standard clinical practice. The system aims to bridge the gap between sophisticated machine learning capability and meaningful clinical utility so that predictive analytics can easily be translated into valuable clinical decision-making aids.

The overall objective of this study is to construct accurate predictive models through strict application of a blend of complementary machine learning algorithms like Logistic Regression for simple linear modeling with explainable coefficients, Random Forest for uncovering complex feature interactions and non-linear relationships, Gradient Boosting for sequential learning and error correction capabilities, Support Vector Machines for maximum margin classification with kernel adaptability, and Extreme Gradient Boosting (XGBoost) for high-boosting capability with built-in regularization. This multi-algorithm framework guarantees extensive evaluation of different modeling paradigms as well as identification of the best approaches to cardiovascular risk prediction.

The research emphasizes data-driven results through meticulous Exploratory Data Analysis procedures designed to value intrinsic feature distributions, identify patterns in missing data potentially affecting model performance, investigate correlations between single risk factors and cardiovascular outcomes, and identify possible data quality defects potentially devastating to predictive accuracy. This thorough groundwork of analysis ensures subsequent model building is founded on a solid understanding of the dataset characteristics and clinical correlations.

Robust preprocessing techniques are a central component of the research design that incorporate robust feature selection techniques based on clinical relevance and statistical relevance, wide categorical encoding techniques maintaining clinical meaning with the ability to be compatible with machine learning models, and pertinent feature scaling techniques that enhance technical proficiency along various algorithmic routes. All the preprocessing techniques strive to preserve clinical interpretability while technically maximizing performance.

Sophisticated evaluation frameworks involve critical testing of model performance via stratified k-fold cross-validation maintaining class balance between validation folds, various converging evaluation measures like accuracy, precision, recall, F1-score, and ROC-AUC that provide complementary perspectives on classification performance, and advanced statistical validation procedures like confidence interval estimation and significance testing to enable credible statistical inference regarding model performance differences.

Along with its simple predictive capability, this research combines clinical guideline systems that provide evidence-based advice to patients as well as to clinicians, combining single-risk-factor evaluation with actionable guidance regarding how to enhance cardiovascular health. The overall approach is an acknowledgment of the fact that the best clinical decision support will extend beyond the prediction of risk to include actionable recommendations on risk factor change, lifestyle modification, and management.

The integrated system combines sophisticated predictive analytics with clinically actionable recommendations to provide a comprehensive decision support tool that is capable of assisting healthcare professionals in the early detection of high-risk patients during early stages of disease progression so that timely preventive measures can be established, which can significantly improve patient outcomes. The system is designed to harmoniously integrate into existing clinical practices while providing health care practitioners with easy-to-use interfaces for risk evaluation as well as clinical decision-making.

By integrating advanced machine learning predictions with human-interpretable clinical recommendations, the developed platform not only provides accurate cardiovascular risk prediction but also enables actionable, evidence-based recommendations for the prevention and management of individual cardiovascular risk factors. The two-pronged mechanism ensures that high-end predictive attributes are translated into clinically significant interventions that can potentially improve patient care and reduce cardiovascular disease burden.

The research closes significant gaps in current cardiovascular risk assessment approaches by applying the power of machine learning to identify subtle patterns and associations not necessarily observable via traditional analytical methods, without compromising clinical interpretability and real-world implementability within contemporary healthcare settings. The systematic methodology ensures that powerful analytics assumes a significant role in improved clinical decision-making as well as improved patient outcomes in cardiovascular medicine.

3. Methodology

3.1 Study Design and Data Source

This research employed a retrospective cross-sectional study design to develop and validate an interpretable machine learning heart disease prediction model. The research utilized the combined Heart Failure Prediction Dataset on Kaggle, which combines five well-known cardiovascular datasets: Cleveland, Hungarian, Statlog, Long Beach VA, and Switzerland datasets. It is an integrated dataset with 920 anonymous patient records, providing a good source of model development for diverse international medical centers with different patterns of cardiovascular disease and clinical presentation.

3.2 Dataset Description and Feature Selection

The original dataset had eleven clinical features commonly employed in cardiovascular risk prediction: age, gender, chest pain type (typical angina, atypical angina, non-anginal pain, asymptomatic), resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, ST depression (Old peak), and ST slope features. The target variable was the binary presence or absence of heart disease.

3.2.1 Statistical Feature Selection Process

A strict feature selection procedure was performed using Stata 16 to choose the most predictive features in a clinically interpretable manner. All features were first examined with univariate logistic regression, which was then followed by forward stepwise multivariate logistic regression with an entry criterion of $p < 0.05$. Among the categorical predictors, ST slope was the best predictor (Pseudo $R^2 = 0.3018$, $p < 0.001$), followed by exercise-induced angina (OR = 10.62, 95% CI: 7.59-14.86, $p < 0.001$) and type of chest pain (Pseudo $R^2 = 0.2269$, $p < 0.001$). For continuous variables, maximum heart rate possessed high predictive ability (OR = 0.964, 95% CI: 0.957-0.970, $p < 0.001$), and age

(OR = 1.066, 95% CI: 1.050-1.083, $p < 0.001$) and cholesterol (OR = 0.995, 95% CI: 0.994-0.997, $p < 0.001$) possessed important correlations.

The last multivariate model worked very well statistically (LR $\chi^2 = 655.46$, Pseudo $R^2 = 0.5193$, $p < 0.001$). Following statistical significance requirements, clinical availability requirements, and interpretability demands, the five most significant features were kept: age, cholesterol, chest pain type, exercise-induced angina, and ST slope features. Gender was excluded to highlight modifiable risk factors, and fasting blood sugar and resting ECG were excluded according to the least clinical availability and the weakest statistical association, respectively.

3.3 Data Preprocessing Pipeline

3.3.1 Missing Value Handling and Data Quality

The initial data quality check did not reveal many missing values among the selected features. Listwise deletion was used for missing target variables, mean imputation for numeric variables (age, cholesterol), and mode imputation for categorical variables, if necessary, in a conservative approach.

3.3.2 Feature Scaling and Categorical Encoding

The categorical features were systematically encoded through scikit-learn's Label Encoder: chest pain type (four-level category), exercise-caused angina (binary: 1/0), and ST slope (three-level category). Encoding was performed by a custom Data Preprocessor class that maintains distinct encoder objects to provide uniform representations for the training and testing data.

Continuous variables were standardized using StandardScaler z-score normalization to optimize performance across algorithms. The fit-transform approach prevented data leakage by training scalers on training data only and applying the same transformation to test data.

3.4 Model Development Strategy

3.4.1 Multi-Algorithm Comparative Architecture

Five complementary machine learning algorithms were utilized to model various aspects of cardiovascular risk prediction:

Logistic Regression: Linear probabilistic model providing interpretable coefficients and a good statistical baseline (max_iter=1000, random_state=42).

Random Forest: Tree-based ensemble capable of detecting non-linear relationships and feature interactions with importance measures built in, robust to outliers (random_state=42).

Gradient Boosting: Adaptive boosting method for sequential learning, error correction as major emphasis, with sound performance on structured data (random_state=42).

Support Vector Machine: Maximum margin classifier with kernel support, efficient on high-dimensional space (probability=True for ROC-AUC calculation, random_state=42).

XGBoost: Cutting-edge boosting with regularization built in, missing value support, and detailed feature importance (eval_metric='logloss', random_state=42).

3.4.2 Training and Validation Framework

The dataset was split for training and testing in a stratified (80-20) fashion, maintaining class balance in subsets. All the algorithms were trained independently on the same preprocessed data with a probabilistic output setup for end-to-end assessment, including ROC-AUC calculation.

3.5 Hyperparameter Optimization and Performance Evaluation

3.5.1 Hyperparameter Optimization

Hyperparameter optimization was performed exhaustively using RandomizedSearchCV with 5-fold stratified cross-validation. Significant parameter ranges were:

- Random Forest: n_estimators (100-500), max_depth (3,5,7,10,None), min_samples_split (2,5,10)

- XGBoost: n_estimators (100-300), learning_rate (0.01-0.2), max_depth (3-6), subsample (0.8-1.0)

- Gradient Boosting: n_estimators (100-300), learning_rate (0.01-0.2), max_depth (3-5)

- Logistic Regression: C (0.001-100), solver (liblinear, lbfgs)

- SVM: C (0.001-100), kernel (rbf, linear, poly), gamma (scale, auto, 0.001-0.1)

ROC-AUC was employed as the primary optimization metric due to robustness against class imbalance, with parallel processing to maximize computational efficiency.

3.5.2 Overall Performance Evaluation

Model performance was evaluated with standard metrics: accuracy (overall classification performance), precision (positive predictive value), recall/sensitivity (true positive rate), F1-score (harmonic mean of precision and recall), and ROC-AUC (threshold-independent discriminative power). Each metric was calculated with scikit-learn methods under the same evaluation protocols for every algorithm

3.5.3 Statistical Significance Testing

It was accomplished by the ModelEvaluator class with a full statistical framework

Stratified K-Fold Cross-Validation (k=5) maintained class distribution throughout folds with a non-static random state for reproducibility.

95% Confidence Intervals approximated via normal distribution:

$CI = \text{mean} \pm (1.96 \times SE)$, where $SE = \sigma/\sqrt{n}$.

McNemar's Test evaluated classification precision between model pairs through contingency tables and chi-square distribution with continuity correction ($\alpha = 0.05$).

Wilcoxon Signed-Rank Test provided a non-parametric comparison of cross-validation performance distributions against all metrics with paired data from CV folds.

3.6 Model Interpretability and Clinical Decision Support

3.6.1 Interpretability Framework

A number of techniques made clinical interpretability certain: Logistic Regression coefficients provided point odds ratio interpretations, Random Forest employed impurity-based feature importance, XGBoost employed gain-based importance measures, and SHAP (Shapley Additive exPlanations) analysis provided model-agnostic descriptions of both global feature importance ranking and local individual prediction explanations.

3.6.2 Clinical Decision Support System

An integrated prediction interface was developed, providing: validation of clinical feature input, estimation of probability of risk with confidence intervals, clinical risk stratification (low, moderate, high), and personalized advice according to patient-specific risk profiles. An evidence-based advice engine provides patient-specific lifestyle modification, clinical follow-up advice, risk factor management strategies, and emergency care criteria for high-risk patients.

3.7 Bias Assessment and Ethical Issues

3.7.1 Fairness Evaluation

Model performance was stringently tested across demographic subgroups like age groups (≤50, 51-65, >65 years) and gender group comparisons despite sex exclusion from characteristics. Sensitivity, specificity, positive and negative predictive values, and ROC-

AUC were compared between subgroups to identify fair performance as well as identify possible inequities.

3.7.2 Data Privacy and Ethical Implementation

All patient data consisted entirely of anonymized records with no personally identifiable information. Robust privacy measures entailed encrypted storage, secure computing environments, and access controls to authorized personnel only. Ethical rationale is that machine learning predictions enhance clinical judgment but do not replace it, through open model design and clear communication of limitations to clinicians.

3.8 Software Implementation and Reproducibility

3.8.1 Technical Infrastructure

Analysis pipeline was carried out using Python 3.8+ and key libraries: scikit-learn (machine learning), XGBoost (gradient boosting), pandas (data manipulation), numpy (numerical computation), matplotlib/seaborn (visualization), SHAP (interpretability), and streamlit (user interface). Object-oriented programming was implemented using specialized classes: DataLoader, DataPreprocessor, ModelTrainer, ModelEvaluator, Predictor, Visualizer, and HeartRecommendationSystem.

3.8.2 Reproducibility Measures

Fixed random seeds (random_state=42) were established for all the stochastic operations like data splitting, cross-validation, tuning of hyperparameters, and model training. Comprehensive documentation encompasses everything that is done in preprocessing, version control using Git, rigorous requirements specification, and automated test cases for quality assurance.

3.9 Study Limitations

3.9.1 Dataset and Design Limitations

The retrospective design is vulnerable to selection bias, and restricted demographic representation can restrict generalizability. External validation was not performed since it was impossible to find suitable external datasets with corresponding feature sets and sufficient quality. The step of reducing features could have left out predictive factors for specific patient subgroups.

3.9.2 Model Limitations

The binary classification approach has the risk of oversimplification of the range of cardiovascular risk, and static models cannot handle the temporal progression of disease or the dynamic change of risk factors. The planned ensemble approach was ultimately discounted since individual algorithms performed at high enough levels to make additional complexity unnecessary for clinical use.

3.9.3 Implementation Challenges

Clinical deployment involves extensive testing in diverse populations, compatibility with electronic health records, comprehensive clinician training, and ongoing performance monitoring. These limitations support the necessity of systematic clinical testing with patient safety and effectiveness as the initial priority before wide deployment.

This strategy provides a rigorous scientific foundation for heart disease prediction with ongoing clinical interpretability, statistical validity, and feasible implementation in current healthcare settings.

4. Results

4.1 Dataset Characteristics

The final dataset consisted of 918 patient records, featuring five clinical features and binary outcomes for heart disease. No missing values were present, ensuring robust model training.

Continuous Variables:

- **Age**: Mean 53.5 years (SD=9.4, range: 28-77, median: 54.0)
- **Cholesterol**: Mean 198.8 mg/dL (SD=109.4, range: 0-603, median: 223.0)

Categorical Variables:

- **Chest Pain Type**: Asymptomatic (54.03%), Non-anginal pain (22.11%), Atypical angina (18.85%), Typical angina (5.01%)
- **Exercise-Induced Angina**: Absent (59.59%), Present (40.41%)
- **ST Slope**: Flat (50.11%), Upsloping (43.03%), Down sloping (6.86%)

Target Distribution: 55.34% positive cases (n=508), 44.66% negative cases (n=410)

Correlations: Age showed a moderate positive correlation with heart disease (r=0.282), while cholesterol showed a weak negative correlation (r=-0.233).

4.2 Model Performance

4.2.1 Cross-Validation Results (5-fold stratified)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	**83.65%** (80.53-86.78%)	**85.07%** (81.14-89.00%)	85.72% (84.62-86.81%)	**85.35%** (82.86-87.85%)	**90.79%** (87.24-94.34%)
SVM	83.79% (80.46-87.12%)	81.77% (78.00-85.54%)	**91.39%** (88.70-94.08%)	86.23% (83.71-88.76%)	90.33% (87.48-93.19%)
Gradient Boosting	83.38% (79.93-86.83%)	84.66% (80.26-89.06%)	85.97% (82.24-89.71%)	85.16% (82.37-87.96%)	90.27% (86.65-93.88%)
Random Forest	82.15% (78.57-85.74%)	81.98% (78.08-85.88%)	87.21% (83.57-90.85%)	84.42% (81.45-87.39%)	89.08% (85.80-92.36%)
XGBoost	81.47% (77.51-85.42%)	82.58% (78.00-87.16%)	84.75% (81.30-88.19%)	83.55% (80.26-86.84%)	88.43% (84.73-92.14%)

4.2.2Statistical Significance Testing

McNamar's Test revealed significant performance differences between Logistic Regression and ensemble methods (all p<0.001), except SVM (p=0.065). The **Wilcoxon Signed-Rank Test** showed consistent performance across cross-validation folds for most comparisons.

4.3 Clinical Decision Support Example

Sample Patient: Age 58, Cholesterol 280 mg/dL, Atypical Angina, Exercise-induced Angina Present, Flat ST Slope

Model Predictions (High Risk Consensus):

- Logistic Regression: 82.51%
- Random Forest: 96.00%
- Gradient Boosting: 90.97%
- SVM: 84.97%
- XGBoost: 98.25%

Clinical Recommendations Generated:

- Age 58: Enhanced monitoring due to elevated cardiovascular risk
- Cholesterol 280: High level requiring medication consideration
- Exercise-induced angina: Urgent cardiology evaluation needed
- Flat ST slope: Possible ischemia requiring further assessment

4.4 Key Findings

1. **Best Performance**: Logistic Regression achieved the highest accuracy (83.65%) and ROC-AUC (90.79%)
2. **Clinical Utility**: All models showed >80% accuracy with excellent ROC-AUC (>88%)

3. **Feature Importance**: Age and exercise-induced angina emerged as the strongest predictors
4. **Decision Support**: System successfully provided personalized risk assessment and evidence-based recommendations
5. **Statistical Validation**: Robust cross-validation with significant performance differences confirmed between linear and ensemble approaches

The developed system demonstrates strong predictive performance suitable for clinical decision support, with interpretable results and actionable recommendations for cardiovascular risk management.