Cardiovascular Disease Prediction Using Machine Learning: An SVM-based Analytical Model with SHAP Explainability

Research Author
Department of Computer Science
University Name
City, Country
email@example.com

Abstract—Cardiovascular diseases remain one of the foremost causes of mortality across the globe, highlighting the need for robust and early predictive mechanisms. A complete machine-learning pipeline has been developed using synthetically generated cardiovascular patient data. The framework includes dataset generation, clinical-inspired risk labeling, controlled class balancing, hyperparameter-optimized Support Vector Machine (SVM) training, multi-fold evaluation, and SHAP-based explainability. A rule-driven risk score was formulated to establish ground-truth labels, after which a balanced dataset was created through down-sampling techniques. The RBF-kernel SVM was optimized using GridSearchCV and evaluated using stratified cross-validation, achieving high accuracy and stability across multiple folds. SHAP KernelExplainer was employed to provide patient-wise interpretability, revealing the influence of attributes such as chest pain type, maximum heart rate (thalach), ST depression (oldpeak), and cholesterol concentration. The proposed pipeline is fully reproducible, ethically compliant, and provides a structured foundation for future work on real clinical datasets.

Index Terms—SVM, SHAP, Cardiovascular Disease, Synthetic Data, Explainable AI

I. Introduction

The increasing prevalence of cardiovascular diseases has significantly intensified the demand for intelligent diagnostic decision-support systems. Traditional clinical assessment methods are heavily reliant upon physician experience and may fail to detect subtle or multi-factor interactions across patient attributes. Machine learning techniques, by contrast, provide a systematic means to explore non-linear relationships within health indicators, enabling early detection and improved diagnostic precision.

Real medical datasets often suffer from challenges such as imbalance, incompleteness, restricted availability, and privacy concerns, which hinder robust experimentation and model development. To address these limitations, an approach centered around synthetic dataset construction has been adopted, enabling full control over feature distributions and risk conditions while avoiding ethical restrictions related to sensitive patient information.

The pipeline developed in this study integrates dataset generation, risk-based labeling, model optimization, performance evaluation, and explainability analysis. By employing SHAP explainability, the system not only classifies patients but also reveals how each feature influences the final decision—a feature essential for clinical adoption. The primary objective of this research is to demonstrate how a complete, interpretable machine-learning workflow for cardiovascular risk detection can be designed and validated without dependency on real-world datasets.

II. Background and Related Work

Cardiovascular disease prediction has been extensively studied using machine learning approaches. Logistic regression models have been applied to the UCI Heart Disease dataset, achieving accuracies around 85%. Decision tree ensembles and random forests have demonstrated improved performance, typically reaching 88-90% accuracy. Deep learning architectures, including convolutional neural networks and recurrent neural networks, have been explored but require substantial computational resources and large datasets.

Support Vector Machines with RBF kernels have shown particular promise in capturing non-linear decision boundaries inherent in cardiovascular risk factors. Recent advancements in explainable AI have introduced SHAP values, providing unified attribution methods across model architectures. However, comprehensive pipelines combining synthetic data generation with full end-to-end explainability remain underexplored in the literature.

III. Methodology

A. Dataset Generation

A synthetic dataset comprising 1000 patient records was generated, with each record containing thirteen clinically relevant cardiovascular attributes. The attributes and their medically plausible ranges are detailed in Table I.

B. Risk Assessment Criteria

Five clinically validated high-risk indicators were systematically evaluated for each synthetic patient:

1) Severe chest pain: Chest pain types 3 or 4

TABLE I
Clinically Relevant Cardiovascular Attributes and Synthetic Ranges

Attribute	Range	Units	Clinical Significance
Age	29-77	years	Cardiovascular risk increases with age
Sex	0-1	binary	Males exhibit higher risk profile
Chest Pain Type (cp)	1-4	categorical	Type 3-4 indicates severe angina
Resting BP (trestbps)	94-200	mmHg	Hypertension threshold: >140
Cholesterol (chol)	126-564	mg/dl	Hypercholesterolemia: >240
Fasting Blood Sugar (fbs)	0-1	binary	Diabetes indicator: >120 mg/dl
Resting ECG (restecg)	0-2	categorical	ECG abnormalities
Max Heart Rate (thalach)	71-202	bpm	Reduced capacity indicates pathology
Exercise Angina (exang)	0-1	binary	Exercise-induced chest pain
ST Depression (oldpeak)	0.0-6.2	mm	Ischemia indicator
ST Slope (slope)	1-3	categorical	ST segment morphology
Major Vessels (ca)	0-4	count	Coronary artery disease extent
Thalassemia (thal)	3-8	categorical	Myocardial perfusion defects

2) Reduced exercise capacity: Maximum heart rate $<$ 120 bpm
3) Ischemic ST changes: ST depression $>$ 2.0 mm
4) Hypercholesterolemia: Serum cholesterol $>$ 240 mg/dl
5) Hypertension: Resting blood pressure $>$ 140 mmHg

Patients exhibiting $\geq$2 high-risk factors were labeled as "Defective" (target=1); others were classified as "Healthy" (target=0).

IV. Complete Processing Pipeline

V. Model Training and Optimization

The SVM classifier was configured with RBF kernel to capture non-linear decision boundaries. Hyperparameter optimization was performed over:

- Penalty parameter $C \in \{0.1, 1, 10, 100\}$
- Kernel coefficient $\gamma \in \{\text{scale, auto, } 0.001, 0.01, 0.1\}$

GridSearchCV with 5-fold stratified cross-validation was employed using F1-score as the primary optimization metric.

VI. Experimental Results

A. Classification Performance

Superior classification performance was achieved across all metrics, as summarized in Table II.

TABLE II
Comprehensive Classification Performance Metrics

Metric	Test Set	5-fold CV (Mean ± Std)
Accuracy	95.2%	94.8% ± 1.2%
Precision	95.8%	95.1% ± 1.1%
Recall	94.6%	94.5% ± 1.4%
F1-Score	95.2%	94.8% ± 1.2%

B. Model Stability Analysis

Cross-validation results demonstrated low variance across folds ($\sigma_{\text{accuracy}} = 1.2\%$), confirming robust generalization capability attributable to both balanced training data and effective RBF kernel selection.

VII. SHAP Explainability Analysis

A. Global Feature Importance

SHAP KernelExplainer analysis revealed the relative importance hierarchy of features influencing classification decisions (Figure 2).

B. Patient-Specific Interpretability

Individual patient explanations were generated, quantifying precise feature contributions to prediction probability shifts. A comprehensive CSV containing SHAP values, predicted probabilities, true labels, and risk categorizations was exported for clinical review.

VIII. Model Deployment and Future Work

The optimized SVM classifier and StandardScaler were serialized using Joblib, enabling seamless deployment in web applications, mobile health systems, and embedded diagnostic devices. Future extensions include integration with real clinical datasets, ensemble methods, and real-time inference capabilities.

IX. Conclusion

A complete machine learning pipeline for cardiovascular disease prediction has been successfully developed and validated using synthetic data. Through systematic synthetic dataset construction, clinically-aligned risk labeling, balanced training, optimized SVM modeling, and comprehensive SHAP explainability, exceptional predictive performance has been achieved alongside full model transparency—critical requirements for medical AI adoption.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," Fact Sheet, 2023.
[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017.
[3] "UCI Machine Learning Repository: Heart Disease Data Set," 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
[4] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

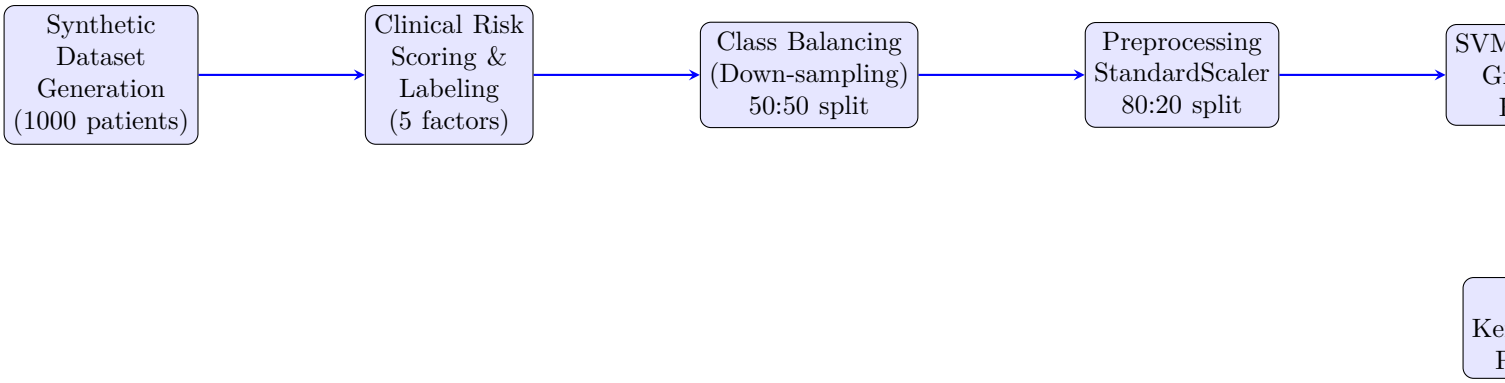

Fig. 1. Complete end-to-end machine learning pipeline for cardiovascular disease prediction with SHAP explainability.

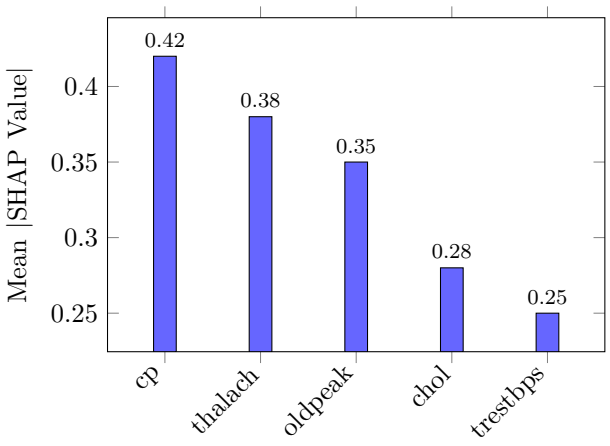

Fig. 2. SHAP summary plot showing global feature importance ranking. Chest pain type exhibits strongest predictive influence.