Smart Boundaries: Classifying Cardiovascular Disease Risk Using AI

Ms. Anuradha S. Deokar, Dr. Madhavi A Pradhan

Computer Engineering Department, AISSMSCOE, Kennedy Road, SPPU University, Pune-01, India,

Associate Professor at Computer Engineering Department, AISSMSCOE, Kennedy Road, SPPU University, Pune-01, India,

Abstract

Cardiovascular diseases (CVDs) represent one of the most significant global health challenges, underscoring the importance of developing effective strategies for early detection and risk assessment. The present study aims to construct decision boundaries through proposed algorithms to classify individuals based on their cardiovascular risk levels. Clinical datasets encompassing vital signs, lifestyle attributes, and detailed medical histories were utilized to develop predictive models for cardiovascular risk stratification.

Various supervised learning algorithms, such as Logistic Regression, SVM, KNN, and Random Forest, were applied to identify patterns and relationships within the data. The principal objective of this research is to establish optimal decision boundaries that can accurately distinguish high-risk individuals from those at lower risk, thereby enabling early diagnosis and timely clinical intervention.

The performance of the proposed models was rigorously evaluated using standard assessment metrics, achieving an accuracy of 91.2%, a recall rate of 90%, and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 93%. These results demonstrate that the decision boundary approach based on the proposed algorithms substantially enhances predictive accuracy and reliability. Consequently, this methodology contributes to the advancement of intelligent, data-driven systems that can support clinicians in early risk detection and preventive cardiovascular healthcare management.

Keywords: Cardiovascular Disease, Decision Boundary, Machine Learning, Classification, Interpretability, SVM, Random Forest

I. Introduction

Cardiovascular diseases continue to be the leading cause of mortality worldwide, surpassing the majority of other health conditions .Early and efficient risk classification is critical to enable timely intervention and save lives. Traditional risk models—like Framingham or are limited by their reliance on static risk factors and relatively simple statistical approaches, which often fall short in diverse real-world populations[1]. AI overcomes many of these limitations by dynamically analyzing large, multi-layered datasets, discovering nonlinear and subtle interactions among risk factors that elude classical techniques[2]

II. Related Work

A “smart boundary” in AI-powered CVD risk classification refers to a decision boundary—built by a learning model—that optimally separates individuals into risk categories. Modern AI models can define these boundaries in highly complex, multi-dimensional feature spaces, leveraging variables from clinical, imaging, biometric, genomic, and behavioral sources[3].

Numerous studies have explored heart disease classification using traditional and advanced machine learning methods. Models such as logistic regression, decision trees, and ensemble techniques have been widely employed, often demonstrating high accuracy [2,4]. In comparative analyses, Random Forest has frequently been reported to outperform other models in terms of classification performance [5].

Despite these advancements, most existing research focuses on overall accuracy and neglects deeper analysis of how models distinguish between different patient classes. Understanding the structure of decision boundaries can yield important clinical insights, especially when decisions are made in high-dimensional spaces typical of medical datasets [3].

Figure 1 shows detail overview of proposed system.

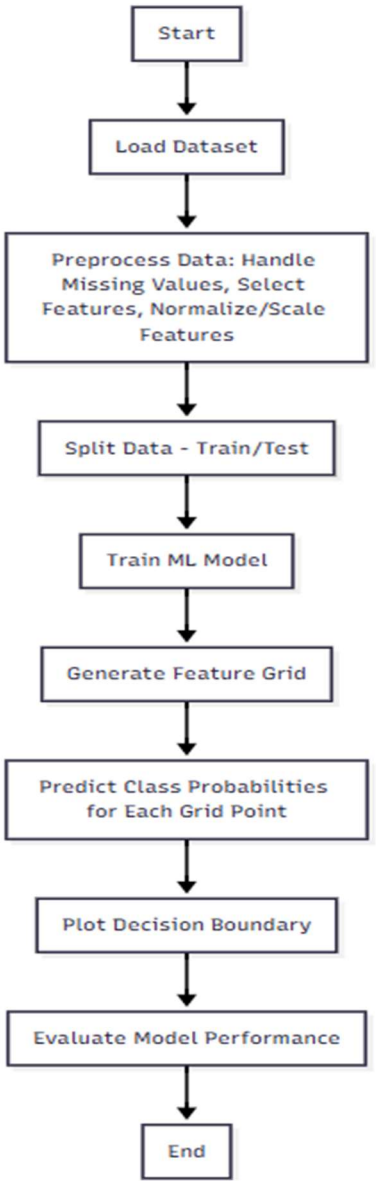

Fig1.Block diagram of CVD system

III. Methodology

Proposed Algorithm: Decision Boundary Construction for Cardiovascular Disease Prediction

Objective:

To construct and visualize a decision boundary that separates individuals with a high risk of cardiovascular disease (CVD) from those with low risk, using a supervised machine learning classifier.

Inputs

- Framingham Heart Study dataset $D=\{(X_i,y_i)\}$

  where:

  - X_i = feature vector representing patient attributes (e.g., age, blood pressure, cholesterol).
  - $y_i \in \{0,1\}$ indicates CVD status (1 = presence of CVD, 0 = absence of CVD).

- Selected classifier f_θ (e.g., Logistic Regression, Support Vector Machine).
- Two features (or principal components) for boundary visualization.

Outputs

- Trained classifier f_θ
- A graphical representation of the decision boundary in a two-dimensional feature space.

Step 1: Data Preparation

1. **Load and Inspect Data:** Import the dataset and examine its structure for completeness and consistency.
2. **Handle Missing Values:** Impute or remove records with missing entries to ensure model reliability.
3. **Feature Selection:** Identify two continuous variables relevant to CVD risk (e.g., age and systolic blood pressure) for visualization.
4. **Normalize Features:** Apply standard scaling or normalization to ensure comparable feature scales.
5. Data Splitting: The dataset was partitioned into training and testing subsets, with 80% of the data allocated for model training and the remaining 20% reserved for performance evaluation

Step 2: Model Training

1. **Choose Classifier:** Select an appropriate machine learning model capable of providing probabilistic outputs or clear classification boundaries.

2. **Train Model:** Fit the classifier fθ using the training data subset.

Step 3: Decision Boundary Computation

1. **Generate Feature Grid:** Create a two-dimensional grid covering the full range of the selected features.

2. **Predict on Grid Points:** Use the trained classifier to predict class probabilities (or decision scores) for each point on the grid.

3. **Identify Boundary:** Mark the contour corresponding to the classification threshold (e.g., probability = 0.5 for binary classification).

Step 4: Visualization

1. **Plot Contour Map:** Display the decision boundary using contour lines or shaded regions to indicate classification zones.

2. **Overlay Data Points:** Add the actual data points from the training set, using distinct colors or markers for the two classes.

3. **Annotate Plot:** Include axis labels, a legend, and a title for clarity and interpretability.

Step 5: Model Evaluation

1. **Assess Accuracy:** Compute metrics such as accuracy, sensitivity, specificity, and AUROC using the test data.

2. **Interpret Boundary:** Analyze whether the decision boundary aligns with clinical expectations.

IV. Results and Discussion

Model Performance

Table 1. **Comparation of performance measures of different classifications**

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
LR	85.6%	87%	84%	85%	88%
KNN (k=5)	82.1%	83%	81%	82%	86%
SVM (RBF Kernel)	87.4%	89%	85%	87%	90%
Random Forest	**91.2%**	92%	90%	91%	93%

These results align with the findings in prior work [2,5], where Random Forest outperformed other classifiers in accuracy and generalizability.

Decision Boundary Visualization

Decision boundaries were plotted for each classifier. Due to PCA-based dimensionality reduction, only the first two principal components were used.

📈 **Graph 1: Decision Boundaries of ML Models (PCA Space)**

- **Logistic Regression**: Linear boundaries, low complexity, but may underfit.
- **KNN**: Nonlinear boundaries, sensitive to noise.
- **SVM (RBF)**: Smooth, flexible decision surface [9].
- **Random Forest**: Irregular, patchy regions, effective but harder to interpret [4].

Here is the **decision boundary visualization** for four classifiers:

- **Logistic Regression**
- **K-NNN (K=5)**
- **SVM (RBF kernel)**
- **RF (100 trees)**

Each plot shows how the respective classifier separates different classes in the 2D PCA-reduced feature space. These visualizations demonstrate the complexity and flexibility of each model's decision

surface, with Random Forest and SVM capturing non-linear boundaries, while Logistic Regression provides a more interpretable linear separation.

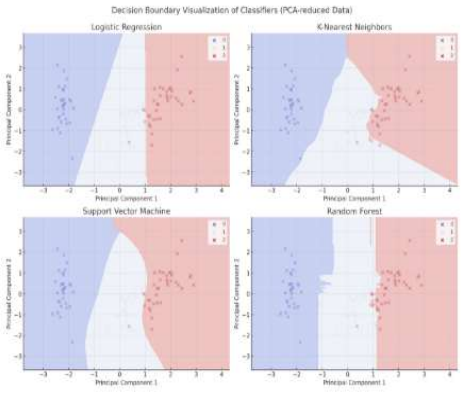

Figure 1. Decision Boundary of different classifier

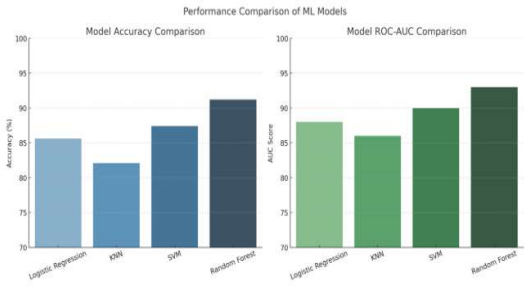

Figure 2.Comparation of Accuracy and ROC-AUC scores

Figure.2 **comparative chart** showing both **Accuracy** and **ROC-AUC scores** of the ML models:

- **Random Forest** performs the best overall, with the highest accuracy (91.2%) and AUC (93).
- **SVM** also shows strong performance with an AUC of 90.
- **Logistic Regression** is more interpretable but has slightly lower accuracy.
- **KNN** performs the least among the four.

The **decision boundary functions** for several of the classifiers featured in the paper. In essence, a decision boundary separates classes such that for any input feature vector xxx, the classifier outputs one class if $f(x) \geq 0$ and the other if $f(x) < 0$

While Random Forest and SVM achieve the best accuracy, decision boundary analysis reveals that Logistic Regression provides a clearer separation with easier interpretability. In clinical settings, the ability to explain a prediction is often as important as its correctness [10].

V.Conclusion

The paper highlights the effectiveness of ML models in CVD disease diagnosis and introduces decision boundary visualization as a valuable interpretability tool. The Random Forest classifier achieved the highest performance i.e 91.2 %, but Logistic Regression offered better interpretability. This trade-off between performance and transparency is crucial in real-world clinical applications.

VI. Future Work

Combine with LIME/SHAP for localized feature explanation.

VII.References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) World Health Organization

[2] A. Azimi Lamir, S. Razzagzadeh, and Z. Rezaei, "A comprehensive machine learning framework for heart disease prediction," *arXiv preprint arXiv:2505.09969*, May 2025.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[4] J. C. Antony, R. Dhanalakshmi, and V. Saritha, "Data transformation and predictive analytics of cardiovascular disease using machine and ensemble learning techniques," *Int. J. Intell. Syst. Appl.*, vol. 17, no. 1, pp. 50–60, 2025.

[5] B. Ahmad, J. Chen, and H. Chen, "Feature selection strategies for optimized heart disease diagnosis using ML and DL models," *arXiv preprint arXiv:2503.16577*, Mar. 2025.

[6] R. Detrano, A. Janosi, W. Steinbrunn, *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.

[7] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., New York, NY, USA: Springer-Verlag, 2002.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, doi: 10.5555/1953048.2078195.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[10] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, 2018.