CardioTrace

Mapping the Future of Cardiovascular Risk

Author: Elijah Oreoluwa

Hackathon: Byte 2 Beat - Hack4Health

Date: February 2026

Live App: https://huggingface.co/spaces/Elijahoreoluwa1/CardioTrace

1. Problem Framing

Cardiovascular disease (CVD) remains the leading cause of death globally, responsible for an estimated 17.9 million deaths annually. Traditional clinical risk tools such as the Framingham Risk Score provide static, single-point-in-time estimates that fail to capture the dynamic, progressive nature of cardiovascular risk over time.

CardioTrace addresses this gap by framing cardiovascular risk as a temporal forecasting problem rather than a simple binary classification task. By simulating patient risk trajectories over time and coupling predictions with SHAP-based interpretability, CardioTrace provides clinicians with both a risk verdict and a mechanistic explanation of what is driving that risk for each individual patient.

2. Dataset

We use the UCI Cleveland Heart Disease Dataset, a widely validated benchmark in cardiovascular ML research. The dataset contains 303 de-identified patient records with 13 clinical features including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia type.

The binary target variable indicates presence (1) or absence (0) of cardiovascular disease. Class distribution is balanced: 164 negative and 139 positive cases. Missing values (6 total across ca and thal features) were imputed using median values.

3. Methodology

3.1 Feature Engineering

A composite clinical risk score was engineered from five key physiological indicators: age, resting blood pressure, cholesterol, ST depression, and maximum heart rate. Each component was normalized to [0,1] and weighted according to clinical literature, producing a single interpretable risk index per patient.

3.2 Temporal Trajectory Simulation

To simulate longitudinal risk progression on cross-sectional data, we generated 5 time-step trajectories per patient by introducing clinically realistic physiological noise (blood pressure drift, cholesterol increase, age progression). This expands the dataset from 303 to 1,515 records and enables visualization of how each patient risk profile evolves over simulated time.

3.3 Modeling

We selected LightGBM as our primary classifier due to its computational efficiency, native handling of tabular clinical data, and compatibility with SHAP interpretability. Hyperparameters were set conservatively (max_depth=4, num_leaves=15) to prevent overfitting on the small dataset. Reproducibility was enforced via fixed random seeds across all stochastic operations.

3.4 Evaluation

Model performance was evaluated using stratified 5-fold cross-validation on the training set, followed by final evaluation on a held-out 20% test set. AUC-ROC was selected as the primary metric due to the clinical importance of balancing sensitivity and specificity in disease screening contexts.

4. Results

MetricValue5-Fold CV AUC (mean)0.8705-Fold CV AUC (std)+/- 0.028Test AUC-ROC0.963Test Accuracy87%Test Precision (Disease)0.81Test Recall (Disease)0.93Test F1-Score0.87

5. Interpretability

SHAP (SHapley Additive exPlanations) values were computed for every prediction, providing patient-level explanations of which clinical features drove the risk score. Top contributing features across the test set were: chest pain type (cp), ST depression (oldpeak), maximum heart rate (thalach), number of major vessels (ca), and thalassemia type (thal). These align with established clinical knowledge, validating the model's internal reasoning.

6. Limitations

Several scientific limitations must be acknowledged. First, the dataset is small (303 patients) and cross-sectional - true longitudinal validation would require real multi-timepoint patient records. The temporal trajectories are simulated approximations, not real clinical progressions. Second, the model was trained and evaluated on a single cohort from Cleveland, USA, limiting generalizability across diverse populations. Third, missing value imputation via median is conservative and may introduce bias. Finally, the composite risk score weights are manually assigned based on literature and would benefit from data-driven calibration.

7. Societal Impact and Next Steps

CardioTrace demonstrates that interpretable, trajectory-aware cardiovascular risk assessment is achievable with modest computational resources and open clinical data. The system is designed for use by general practitioners and cardiologists as a decision-support tool, not a replacement for clinical judgment.

Planned next steps include: (1) validation on multi-site datasets such as the MIMIC-III clinical database, (2) integration of real longitudinal EHR data to replace simulated trajectories, (3) calibration of the risk score using Platt scaling for reliable probability estimates, (4) extension to multi-class severity staging rather than binary classification, and (5) potential preprint submission to medRxiv as an open-science contribution.

The full codebase is open-source and reproducible, with a live deployment available at https://huggingface.co/spaces/Elijahoreoluwa1/CardioTrace

8. References

1. Detrano et al. (1989) — UCI Cleveland Dataset Detrano, R. et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology, 64*(5), 304-310. https://doi.org/10.1016/0002-9149(89)90524-9

2. Lundberg & Lee (2017) — SHAP Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30.* https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

3. Ke et al. (2017) — LightGBM Ke, G. et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146-3154. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

4. WHO Cardiovascular Fact Sheet World Health Organization (2024). Cardiovascular diseases fact sheet. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

5. UCI Repository Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+disease