CardioInsight AI: Explainable Cardiovascular Disease Risk Prediction

1. Introduction

Cardiovascular Disease (CVD) is one of the leading causes of mortality worldwide. Many cardiovascular events are preventable if individuals are informed early about their risk and the factors contributing to it. However, traditional risk prediction tools often behave like “black boxes,” providing a risk score without explaining *why* the risk is high or *how* it can be reduced.

The goal of **CardioInsight AI** is to bridge this gap by building an **AI-powered, explainable cardiovascular risk prediction system**.

The system not only predicts a **10-year CVD risk percentage**, but also:

- Identifies **key contributing risk factors**
- Shows **how much the risk can be reduced** if those factors improve
- Provides **actionable, understandable insights** suitable for non-technical users

This project is designed as a **decision-support system**, not a medical diagnosis tool, and follows ethical AI principles relevant to healthcare.

---

2. Datasets Used

This project intentionally uses **two different datasets**, each with a clearly defined role.

2.1 Framingham Heart Study Dataset (Training Dataset)

The Framingham Heart Study dataset is a well-established dataset used extensively in cardiovascular research. It contains long-term clinical and lifestyle information such as:

- Age and gender
- Blood pressure (systolic and diastolic)
- Cholesterol and glucose levels
- Smoking habits
- Body Mass Index (BMI)

Why this dataset was used:

- To train the core machine learning models
- To learn medically validated cardiovascular risk patterns
- To build a strong, generalizable risk predictor

 This dataset was used **only for training**, not for final evaluation.

2.2 Hackathon Cardiovascular Dataset (Evaluation Dataset)

The dataset provided as part of the **Byte 2 Beat / Hack4Health hackathon** was used strictly for **evaluation and validation**.

Why this dataset was used:

- To test generalization on unseen data
- To simulate real-world deployment
- To avoid data leakage
- To comply with hackathon rules and ethical ML practices

This separation ensures that model performance is **honest, realistic, and trustworthy**.

---

3. Why a Two-Model Approach Was Used

Instead of relying on a single model, this project adopts a **two-model strategy**, which is commonly used in healthcare AI systems.

Why not just one model?

Using the same model and data for both training and validation can lead to:

- Overfitting
- Inflated performance metrics
- False confidence in predictions

In healthcare, such mistakes can be harmful.

Model A – Full Feature Risk Learning Model

- **Purpose:** Learn deep cardiovascular risk patterns
- **Dataset:** Framingham Heart Study
- **Features:** Comprehensive clinical + lifestyle variables

Key techniques used:

- Class imbalance handling using `scale_pos_weight`
- Hyperparameter tuning via RandomizedSearchCV
- Probability calibration using **sigmoid calibration**

This ensures that predicted probabilities truly represent **risk percentages**, not just raw model scores.

Model B – Reduced Feature Validation Model

- **Purpose:** Validation, robustness, and deployment realism
- **Dataset:** Hackathon cardiovascular dataset
- **Features:** Smaller, practical feature set

Why this matters:

- In real-world scenarios, full medical data may not always be available
- Model B tests whether learned risk patterns still hold
- Enables deployment in **low-resource or user-input-based settings**

Key Advantage of the Two-Model Design

This architecture:

- Prevents data leakage
- Improves real-world reliability
- Aligns with ethical AI principles
- Meets hackathon evaluation standards

---

4. Data Preprocessing

To ensure reliable learning, careful preprocessing was applied:

- Missing numerical values filled using median
- Categorical medical indicators filled using mode
- Stratified train-test split to preserve class balance
- Clear separation of features and target variable

These steps reduce noise and bias while preserving medical meaning.

---

5. Model Training and Optimization

Both models use **XGBoost**, chosen for its strong performance on structured medical data.
Key methods:

- Randomized hyperparameter search
- Stratified K-Fold cross-validation
- GPU acceleration for efficiency

- Probability calibration for trustable outputs

Evaluation metrics:

- Precision
- Recall
- F1-score
- ROC-AUC
- PR-AUC

Special focus was placed on **recall**, as missing a high-risk patient is more dangerous than a false alarm.

---

6. Risk Threshold Selection

Instead of a fixed cutoff, the system performs a **threshold sweep** to select a decision threshold that prioritizes high recall.
This makes the system more suitable for **preventive healthcare use cases**.

---

7. Explainability & What-If Analysis

A key innovation of this project is **explainability**.
The system:

- Uses SHAP-based explanations to identify top risk contributors
- Shows whether each factor increases or decreases risk
- Performs **what-if simulations** to estimate risk reduction if a factor improves (e.g., lowering BP, quitting smoking)

This transforms the model from a predictor into a **guidance tool**.

---

8. Deployment & Accessibility

The system is deployed as a **live Streamlit web application**, allowing users to:

- Enter personal health details
- Instantly see risk percentage
- Understand contributing factors
- Explore actionable improvements

 Live App:
https://heartdiseasepredictionproject-xw8dxpnkpu76rxwgvxratn.streamlit.app/
 Demo Video:
https://youtu.be/Z2Nmcccmv6c

Models are saved and reused without retraining, ensuring reproducibility.

---

9. Ethical Considerations

- Clear separation of training and evaluation data
- Transparent explanations for predictions
- No medical claims or diagnosis
- Explicit disclaimer that this is a decision-support tool

---

10. Conclusion

CardioInsight AI demonstrates how responsible machine learning can support cardiovascular risk awareness. By combining:

- A two-model architecture
- Ethical dataset usage
- Calibrated risk prediction
- Explainability and actionable insights
- Real-world deployment

this project goes beyond a typical ML model and moves toward **practical, transparent, and trustworthy healthcare AI**.
