Heart Disease Risk Assessment

Presentation Script
Total Time: 10 Minutes (9 Slides)

Slide 1: Title Slide

Hello everyone, welcome to our project presentation. We are Team HealthyHeart. My name is Vi, this is Lam, James, and Duy.

Today, we're showing our heart disease risk assessment project, where we use machine learning to predict whether someone has heart disease and how severe it might be.

Slide 2: The Problem

To start off, I'd like to share some quick facts about the problem we're addressing:

Heart disease is the number one cause of death globally, responsible for roughly 17.9 million deaths each year. On top of that, diagnosing it can be really expensive, ranging from 5k to 10k dollars per patient.

Most machine learning models only tell you whether someone has heart disease. Our project goes further by predicting how severe it is: no disease, mild, or severe. This gives doctors clearer insight for making treatment decisions.

Now I'll pass it over to Duy to provide more details about the dataset we chose for this project.

Slide 3: Dataset Overview

For this project, we utilized the UCI Heart Disease dataset. It contains records from 920 patients across four major medical centers: Cleveland, Hungary, Veteran Affairs Long Beach, and Switzerland. This dataset has 14 clinical features like age, resting blood pressure, serum cholesterol levels, chest pain type, exercise test results, etc.

The target variable contains five severity levels ranging from 0 to 4. 0 indicates no disease, and 4 means the most severe cases.

Slide 4: Data Challenges

The multi-class nature of the dataset presented us with several challenges that we had to address systematically.

First, we faced extensive missing data. The 'ca' feature, which represents the number of major vessels colored by fluoroscopy, was 66% missing. The 'thal' feature for thalassemia was 53% missing. These are clinically important predictors that we couldn't simply discard.

Second, we encountered severe class imbalance. Out of 920 patients, only 28 had the most severe classification—that's a 15:1 imbalance ratio. To address these challenges, we implemented KNN imputation with k=5 for missing values and grouped the original 5 classes into 3 clinically meaningful categories: no disease (0), mild disease (1-2), and severe disease (3-4).

Slide 5: Preprocessing Pipeline

Our preprocessing pipeline consists of five stages. First, we performed data cleaning to remove duplicates and handle obvious errors. Second, we applied label encoding to the seven categorical features in our dataset.

Third, we implemented KNN imputation with k=5 by looking at similar patients to fill in missing values. Fourth, and this is very important, we engineered five new features based on current medical knowledge. For example, the first one, cv_risk_score, is the new cardiovascular risk score. It combines multiple risk factors such as age groups, blood pressure categories, cholesterol categories, and heart rate reserve.

Finally, we applied StandardScaler normalization and performed an 80-20 train-test split with stratification. This pipeline transformed our original 14 features into 18 enhanced features ready for model training.

Slide 6: Hierarchical Classification

Our hierarchical classification system takes a two-stage approach that is similar to how doctors actually work in clinical practice.

First, let me explain why we chose this system. The original dataset had extreme class imbalance - again, a 15 to 1 ratio between healthy patients and the most severe cases. By breaking the problem into stages, we could address this more effectively.

First, Stage 1 uses a Support Vector Machine to perform binary classification. We simply detect whether the heart disease is present or not.

In Stage 2, for patients identified as having the disease, we apply a Random Forest multi-class classifier to determine severity: mild versus severe. Here's where our key improvement comes in: we grouped the original 5 severity levels into 3 clinically meaningful classes.

This grouping reduced our class imbalance from 15:1 to just 3:1. We also used BorderlineSMOTE to balance classes by creating synthetic samples near decision boundaries."

Now I'll pass it over to James.

Slide 7: Results

Let me walk you through our results.

For binary classification, we achieved an F1 score of 85.3%, which exceeded our target of 75% by over 10%.

For 3-class hierarchical classification, we achieved an F1 score of 71.4%. It significantly outperforms published research in this area, which typically achieves 55-65%.

Looking at per-class performance: no disease achieved 83% F1, mild disease achieved 68% F1, and severe disease achieved 48% F1. The lower performance on severe cases is expected, given that we had fewer training samples in that category. But our hierarchical approach still outperforms direct classification methods.

Slide 8: What Drives Our Model's Predictions?

Now let's discuss feature importance - it helps us understand whether our model is making medically sound decisions.

Our Random Forest model reveals that 'ca' - the number of major vessels - is the single most important feature at 10.4% importance. This makes perfect medical sense because blocked blood vessels directly indicate the presence and severity of heart disease.

Age comes in second at 8.9%, which makes sense as an important risk factor. And here's what's exciting - our engineered 'cv_risk_score' ranks third at 8.4%. This is the composite risk score we created by combining age effects, blood pressure elevation, cholesterol levels, diabetes status, and exercise-induced symptoms. Its high ranking proves our feature engineering approach added real value.

Slide 9: Thank You

Beyond the machine learning models, we've also developed a webapp using React and Flask. I will now pass it over to Lam for the live demo.

Quick Reference: Key Numbers

- 920 patients from 4 medical centers
- 14 original features -> 18 features after engineering
- 66% missing in 'ca', 53% missing in 'thal'
- 5 classes -> 3 classes (grouping strategy)
- Binary F1: 85.3% (target was 75%)
- 3-Class F1: 71.4% (published research: 55-65%)
- Improvement over baseline: 21.9%

Here's a 2-minute speaking script for your UI demo:

Our landing page provides an overview of the assessment process. Users can click 'Start Free Assessment' to begin. The system is designed to be simple - only 4 fields are required."

> "Let's start with a healthy patient scenario - a 42-year-old male with non-anginal chest pain and no exercise-induced angina.

```
| Field Name | Value |
|-----------|-------|
| **Age** * | 42 |
| **Sex** * | Male |
| **Chest Pain Type** * | Non-Anginal Pain (not related to heart) |
| **Exercise-Induced Angina** * | No - I do not experience chest pain during exercise |
| **Resting Blood Pressure** | 118 |
| **Cholesterol Level** | 195 |
| **Fasting Blood Sugar** | Normal (≤120 mg/dl) |
| **Resting ECG** | Normal |
| **Max Heart Rate** | 175 |
| **ST Depression (Oldpeak)** | 0.3 |
| **ST Slope** | Upsloping (normal) |
| **Major Vessels** | 0 vessels |
| **Thalassemia** | Normal |
```

> After accepting the disclaimer, I'll submit..."

> "Now let's see a high-risk case - a 62-year-old female with concerning indicators.

```
| Field Name | Value |
|-----------|-------|
| **Age** * | 62 |
| **Sex** * | Female |
| **Chest Pain Type** * | Asymptomatic (no chest pain) |
| **Exercise-Induced Angina** * | No - I do not experience chest pain during exercise |
| **Resting Blood Pressure** | 140 |
| **Cholesterol Level** | 268 |
| **Fasting Blood Sugar** | Normal (≤120 mg/dl) |
| **Resting ECG** | Left Ventricular Hypertrophy |
| **Max Heart Rate** | 160 |
| **ST Depression (Oldpeak)** | 3.6 |
| **ST Slope** | Downsloping |
| **Major Vessels** | 2 vessels |
| **Thalassemia** | Normal |
```

> "This demonstrates our hierarchical classification model achieving 71% F1-score for multi-class and 85% for binary detection. Thank you!"