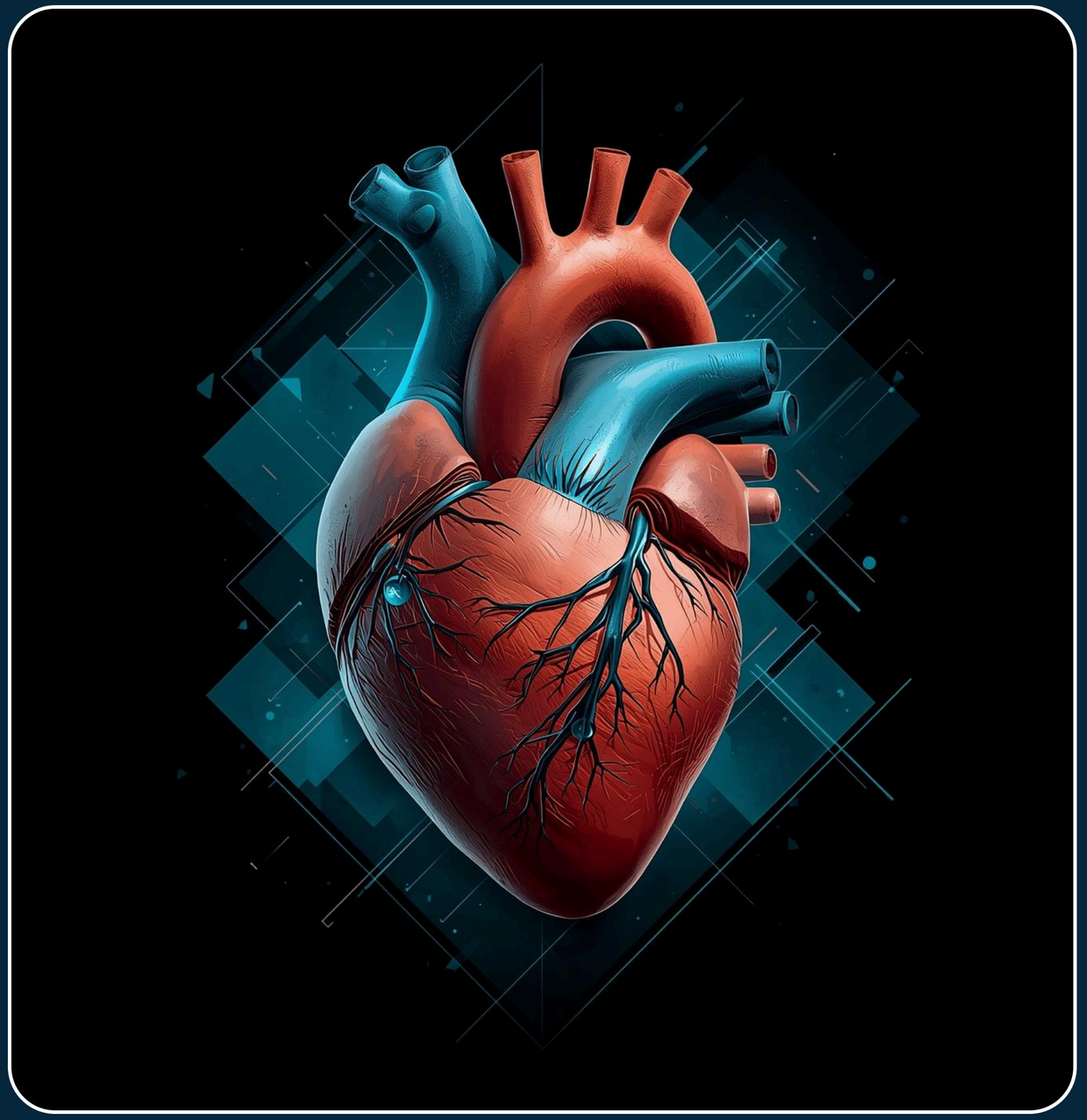
HealthyHeart
Heart Disease Risk Assessment



Heart Disease Risk Assessment

Presented by Vi Nguyen, Lam Nguyen, James Pham, Le Duy Vu

CMPE 257 - Machine Learning

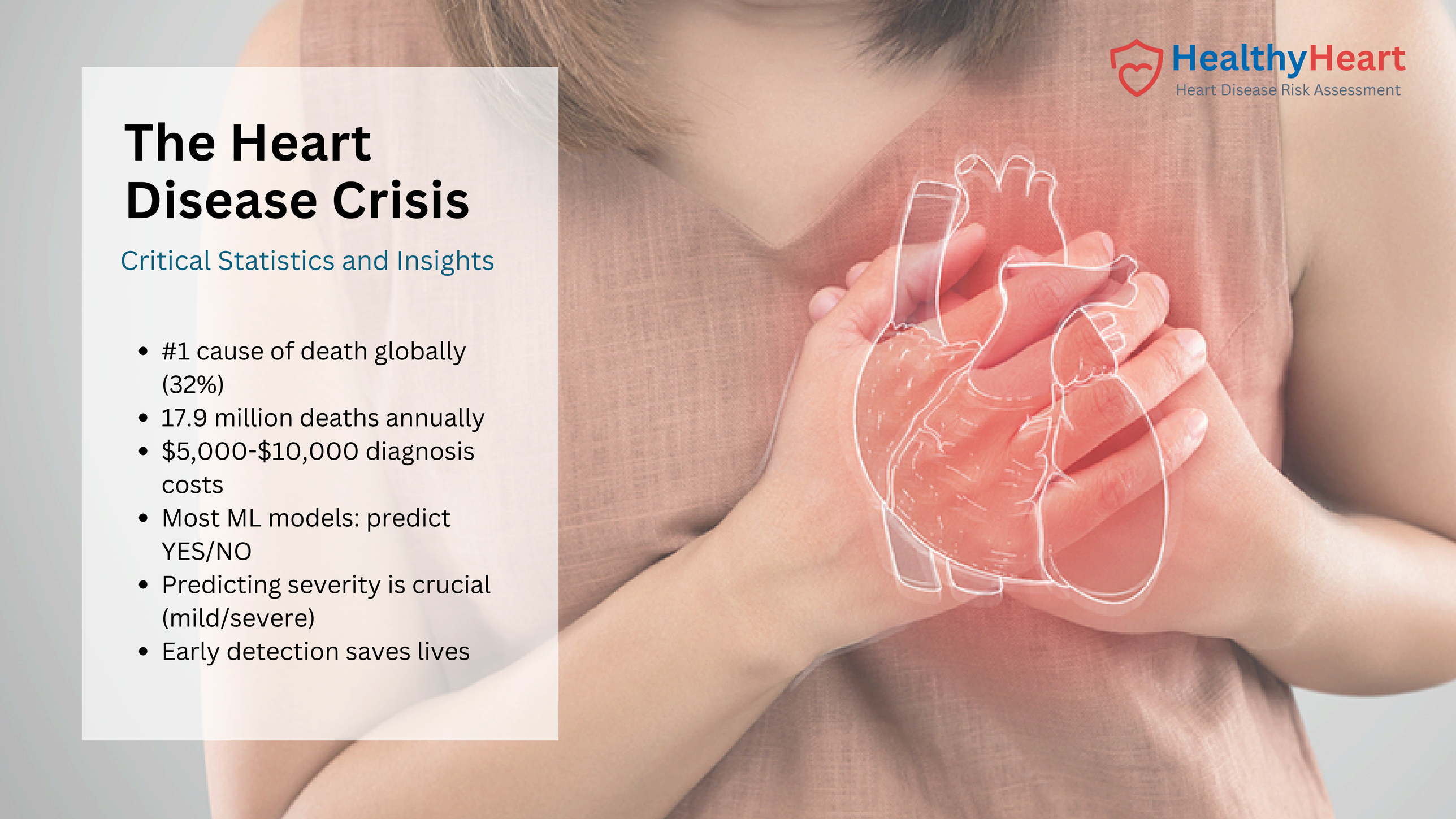
San Jose State University

HealthyHeart
Heart Disease Risk Assessment

The Heart Disease Crisis

Critical Statistics and Insights

- #1 cause of death globally (32%)
- 17.9 million deaths annually
- $5,000-$10,000 diagnosis costs
- Most ML models: predict YES/NO
- Predicting severity is crucial (mild/severe)
- Early detection saves lives

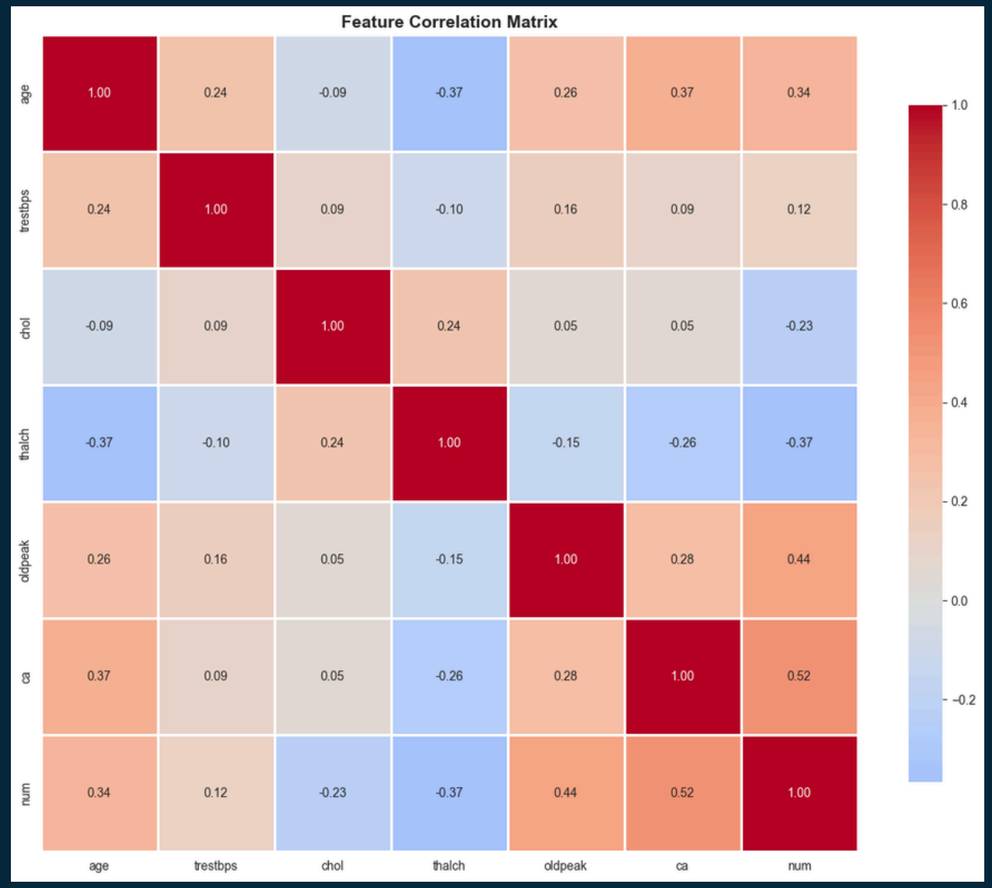


Dataset Overview

HealthyHeart
Heart Disease Risk Assessment

Key Information on Dataset

- UCI Heart Disease Dataset
- 920 patients from 4 centers
- Locations: Cleveland, Hungary, VA Long Beach, Switzerland
- 14 clinical features analyzed
  - Age
  - Resting blood pressure
  - Serum cholesterol levels
  - Chest pain type
  - ECG results
  - Exercise test results
- Target: 5 severity levels (0 - 4)



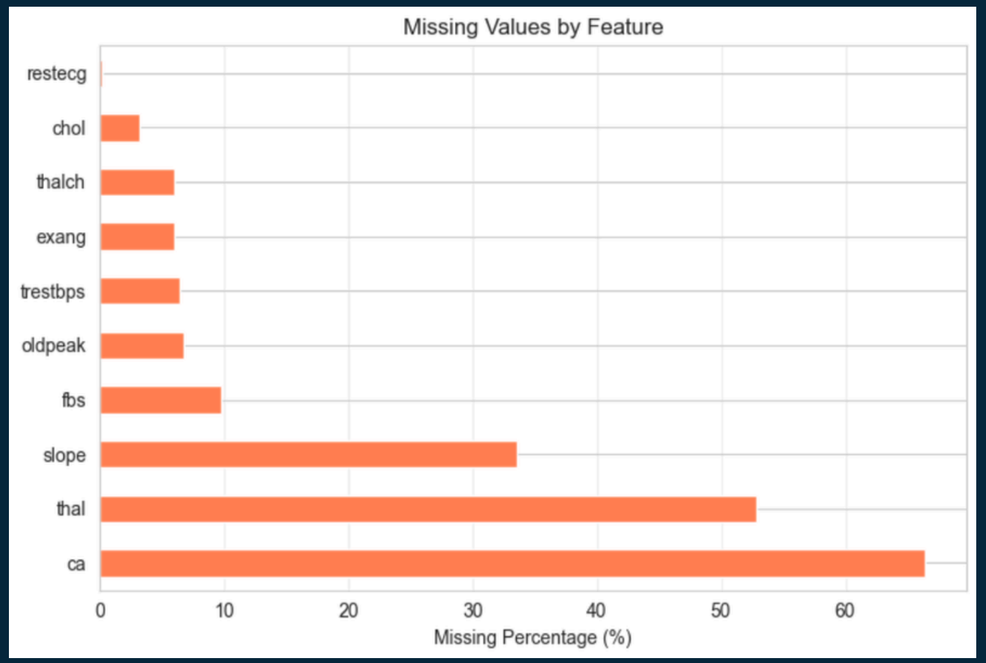
Feature	Description	Type/Range	Unit
age	Patient age	29-77	years
sex	Patient sex	Binary	0=Female, 1=Male
cp	Chest pain type	Categorical (0-3)	0=Typical angina, 3=Asymptomatic
exang	Exercise induced angina	Binary	0=No, 1=Yes
fbs	Fasting blood sugar > 120 mg/dL	Binary	0=No, 1=Yes
trestbps	Resting blood pressure	94-200	mm Hg
chol	Serum cholesterol	126-564	mg/dL
thalch	Maximum heart rate achieved	71-202	bpm
restecg	Resting ECG results	Categorical (0-2)	0=Normal, 2=Abnormal
oldpeak	ST depression induced by exercise	0.0-6.2	mm
slope	Slope of peak exercise ST segment	Categorical (0-2)	0=Upsloping, 2=Downsloping
ca	Number of major vessels colored	0-3	count (66% missing)
thal	Thalassemia/stress test result	Categorical (0-3)	0=Normal, 3=Reversible defect (53% missing)
num	Disease severity (original)	0-4	0=No disease, 4=Severe

Data Challenges

HealthyHeart
Heart Disease Risk Assessment

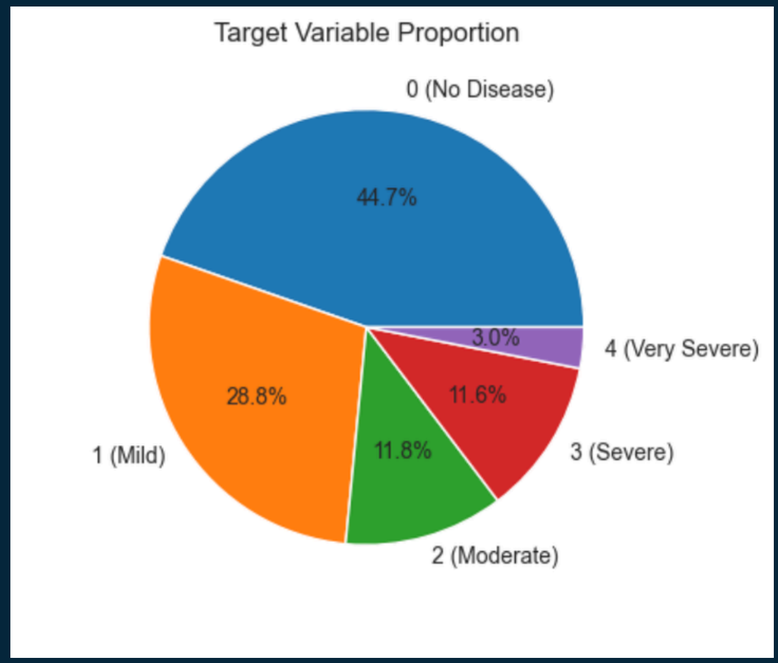
Key Issues

- 66% missing in 'ca' feature (number of major vessels colored by fluoroscopy)
- 53% missing in 'thal' feature (thalassemia)
- 15:1 class imbalance issue
- Only 28 severe cases found



Solutions

- KNN imputation (k=5)
- Group 5 classes to 3
- BorderlineSMOTE for training



Preprocessing Pipeline

HealthyHeart
Heart Disease Risk Assessment

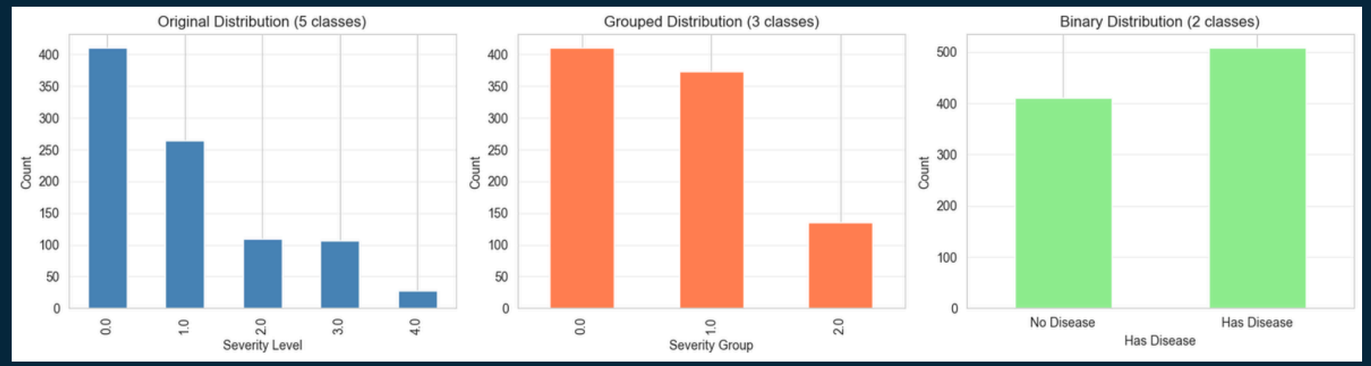
Key Stages of Process

- Stage 1: Data cleaning (remove duplicates)
- Stage 2: Label encoding (7 categorical features)
- Stage 3: KNN imputation (k = 5 neighbors)
- Stage 4: Feature engineering (5 features)
  - cv_risk_score
  - age_group
  - bp_category
  - chol_category
  - heart_rate_reserve
- Stage 5: StandardScaler + 80/20 split with stratification
- Result: 18 features total

```
Missing values before imputation:
trestbps     59
chol         30
thalch       55
oldpeak      62
ca          611
dtype: int64

Missing values after imputation:
0
```

```
Encoded features: ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'thal']
```



Hierarchical Classification

HealthyHeart
Heart Disease Risk Assessment

Two-Stage Approach

- **Stage 1**: SVM Binary Classifier
  - Model: Support Vector Machine (SVM)
  - Task: Disease present? (Yes/No)
  - Training: All 920 samples (Class 0 vs. Classes 1-4)
- **Stage 2**: Random Forest 3-class Classifier
  - Model: Random Forest (200 trees, depth=10)
  - Task: Classify as Mild (1-2) or Severe (3-4)
  - Training: Only disease-positive cases with BorderlineSMOTE

Benefits

- Mimics clinical workflow
- First detect, then assess severity
- 21.9% improvement over direct approach (5-class classification)

Binary Classification Results

Model	Test F1-Score	Test Accuracy
SVM	0.8530	0.8533
XGBoost	0.8471	0.8478
Logistic Regression	0.8312	0.8315
Random Forest	0.8259	0.8261
Gradient Boosting	0.8091	0.8098

Comparison of Approaches (3-Class System)

Model	Test F1-Score	Test Accuracy
Random Forest	0.6991	0.7065
XGBoost	0.6523	0.6630
Gradient Boosting	0.6234	0.6413
SVM	0.5967	0.6196
Logistic Regression	0.5834	0.6087

Results Overview

HealthyHeart
Heart Disease Risk Assessment

Performance Metrics and Key Findings

- Binary Detection (Stage 1): 85.3% F1-score
  - Exceeds 75% goal by 13.7%
  - Precision: 84.8% | Recall: 85.8%
- Hierarchical (Complete System):
  - 71.4% weighted F1-score
  - Significantly better than other published research using the same dataset (55-65%) (Alizadehsani et al.)

Detailed Classification Report – Hierarchical Classification:

	precision	recall	f1-score	support
No Disease	0.79	0.87	0.83	82
Mild	0.72	0.64	0.68	75
Severe	0.48	0.48	0.48	27
accuracy			0.72	184
macro avg	0.66	0.66	0.66	184
weighted avg	0.71	0.72	0.71	184

Our Results Versus Other Research

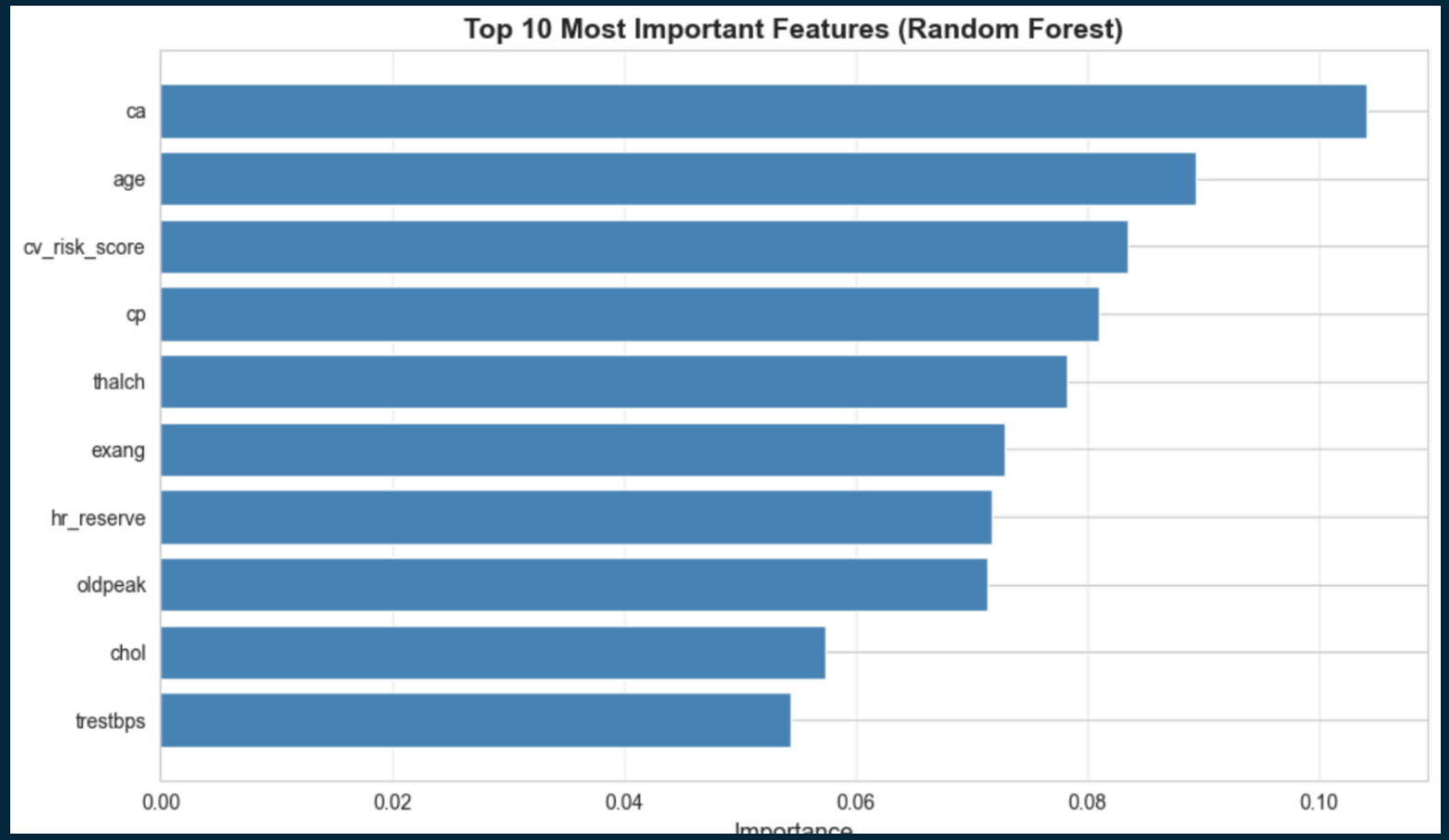
Study	Task	Performance
Mohan et al. [2]	Binary	88% Acc
Shorewala [3]	Binary	93% Acc
Alizadehsani [4]	Multi-class	55-65% F1
Our Work	Binary	85.3% F1
Our Work	Hierarchical	71.4% F1

HealthyHeart
Heart Disease Risk Assessment

What Drives Our Model's Predictions?

Key Insights

- Clinical Validation: Top feature (ca - number of major vessels colored by fluoroscopy) aligns with medical knowledge
- Feature engineering success
- Balanced Model: Uses information from all categories, not over relying on any single type



HealthyHeart
Heart Disease Risk Assessment

Let’s keep our hearts healthy!

Here’s our demo!