Heart Disease Data Analysis

A Comprehensive Study on Dimensionality Reduction, Manifold Learning, and Responsible AI

Edoardo Cipriano, Clara Reynolds, Joep Leeuwenburgh, Oghenekaro Arausi

OPIT - Open Institute of Technology | December 7, 2025

Analysis Overview

The Challenge

Cardiovascular diseases cause 17.9 million deaths annually worldwide. Early detection through machine learning offers promising improvements in patient outcomes and healthcare cost reduction.

Our Approach

We analyzed the UCI Heart Disease dataset using advanced dimensionality dimensionality reduction and manifold learning techniques, uncovering uncovering complex nonlinear structures in cardiac health data.

Key Discovery

Over **70% of the original dataset consisted of duplicates**, reducing 1,025 records to 302 unique patient profiles after rigorous cleaning.

Critical Insight

Heart disease exists on a continuous spectrum rather than as a binary binary condition, better captured by nonlinear methods than traditional traditional approaches.

Dataset Overview & Preprocessing

01

Initial Dataset

1,025 patient records with 14 clinical attributes including demographics, measurements, and diagnosis

02

Duplicate Removal

723 duplicate rows identified and removed (70.5% of data), data), highlighting critical data quality issues

03

Outlier Analysis

IQR method applied to continuous variables, particularly particularly cholesterol and blood pressure measurements measurements

04

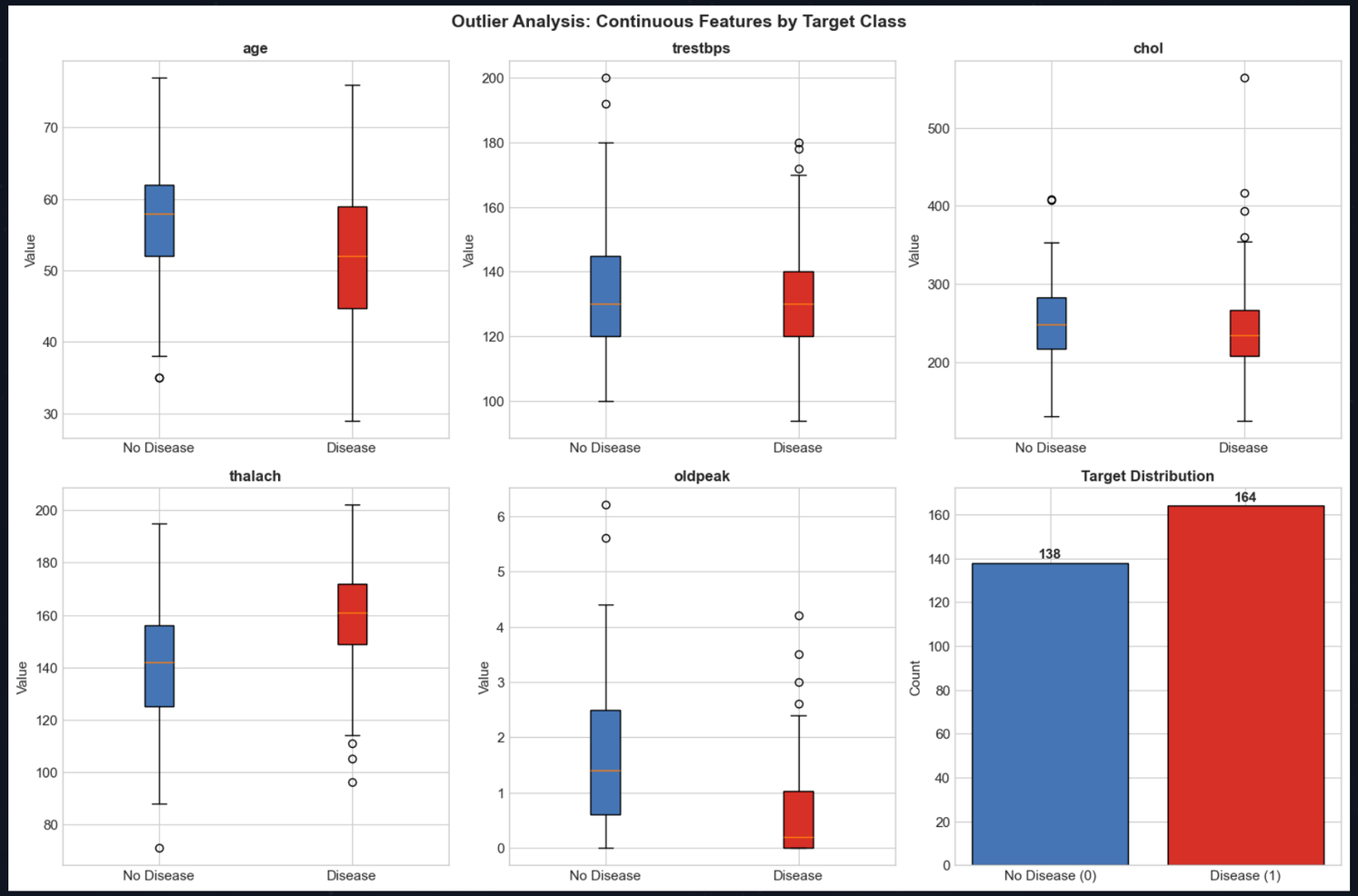
Feature Scaling

StandardScaler normalization to zero mean and unit variance for distance-based algorithms

05

Final Dataset

302 unique patient records ready for advanced analysis analysis with 138 healthy and 164 disease cases



Principal Component Analysis Results

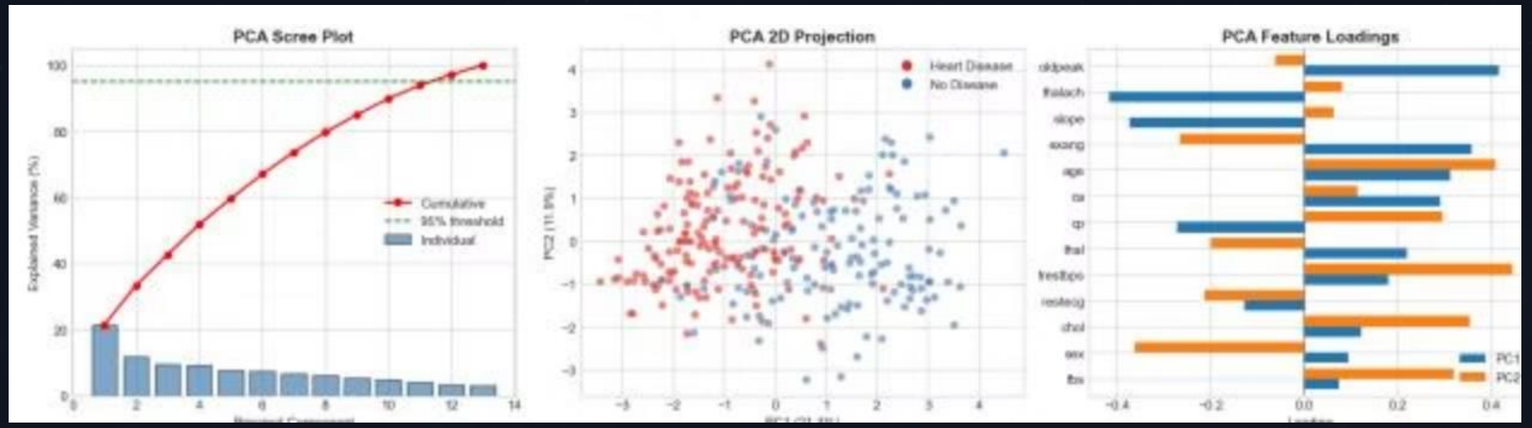
Variance Explained

The first two principal components capture only **33.3% of 33.3% of total variance** (PC1: 21.4%, PC2: 11.9%).

Seven components required to reach 73.6% cumulative cumulative variance, indicating complex multi-dimensional dimensional structure.

Key Feature Loadings

- **PC1:** Dominated by oldpeak (ST depression) and thalach thalach (max heart rate)
- **PC2:** Influenced by exang (exercise-induced angina) and angina) and age
- Considerable class overlap suggests linear separation is insufficient



PCA Biplot: Feature Contributions

Opposing Forces

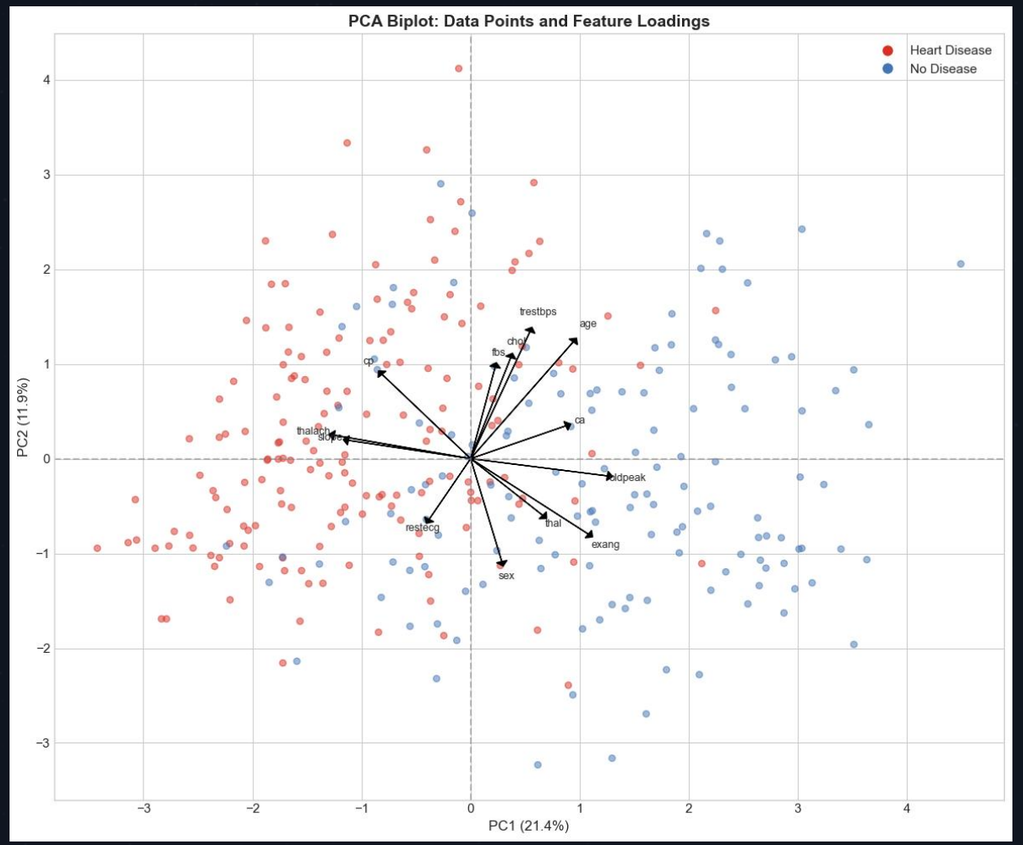
Oldpeak and thalach point in opposite directions along PC1, representing contrasting cardiac indicators

Correlated Features Features

Age, trestbps, and chol vectors align, indicating indicating positive correlation among these these demographic and and clinical measures

Class Overlap

Red (disease) and blue (healthy) points intermingle, demonstrating limitations of linear dimensionality reduction



Nonlinear Methods: t-SNE Analysis Analysis

t-SNE reveals clear cluster separation between disease and healthy populations that was that was invisible in PCA projection, successfully capturing nonlinear relationships.

1

Perplexity = 5

Fragmented clusters, over-emphasis on local structure

2

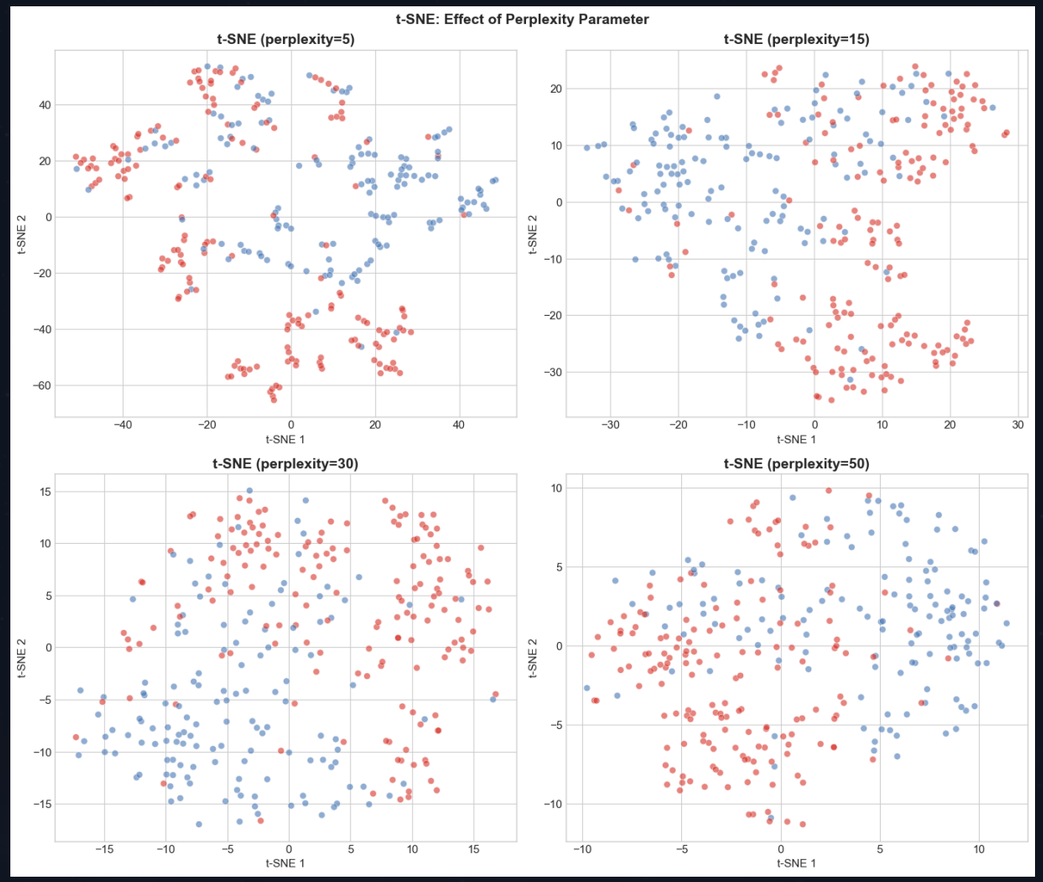
Perplexity = 30

Optimal balance between local and global structure preservation

3

Perplexity = 50

More globular structures, emphasis on global patterns

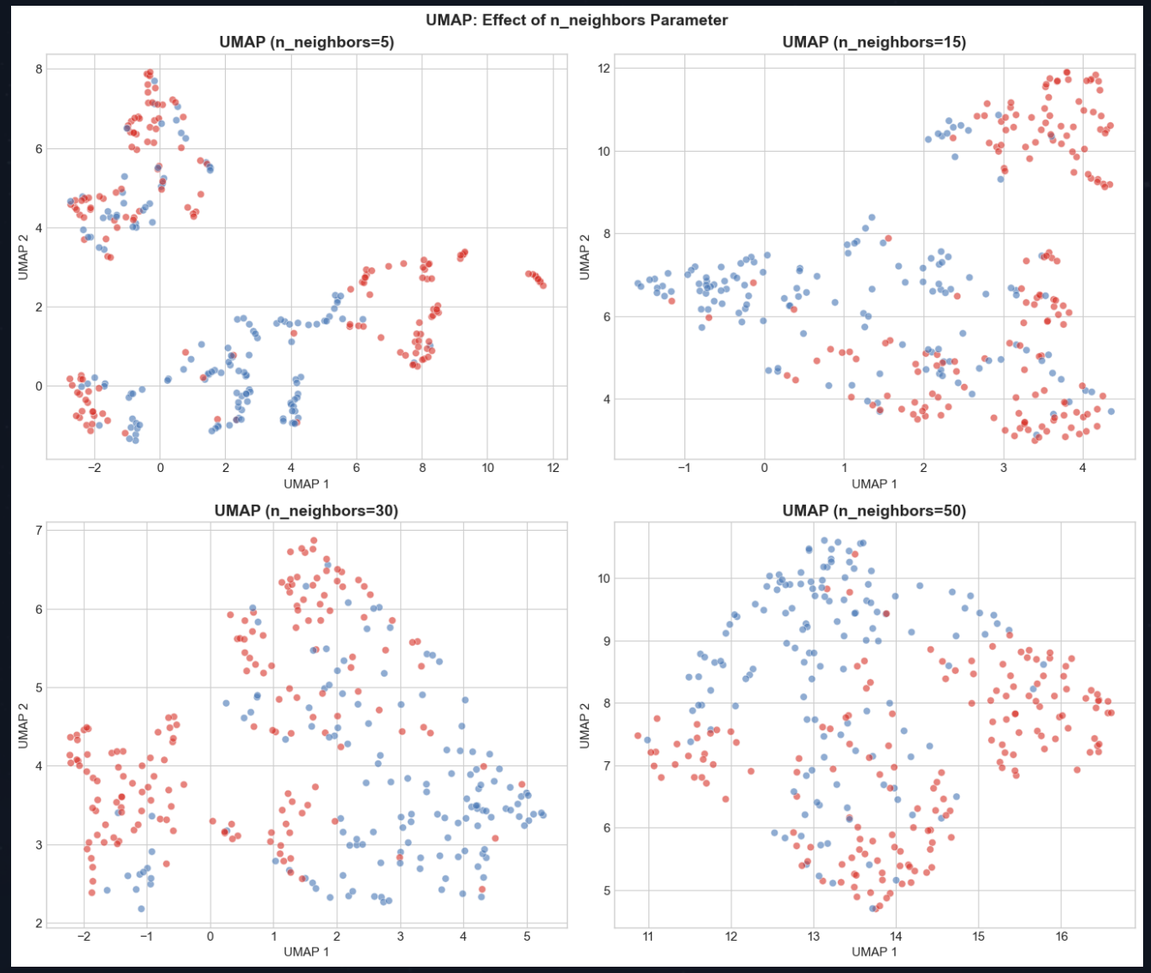


UMAP: Fast & Effective Visualization

Parameter Sensitivity

The n_neighbors parameter controls the the balance between local and global structure:

- **Small values (5, 15):** Emphasize local structure with tighter clusters
- **Large values (30, 50):** Capture more more global relationships



Performance Advantage

UMAP demonstrates **comparable visualization quality to t-SNE** but with with significantly faster computation time, time, making it ideal for larger datasets. datasets.

Successfully preserves both local and global structure simultaneously.

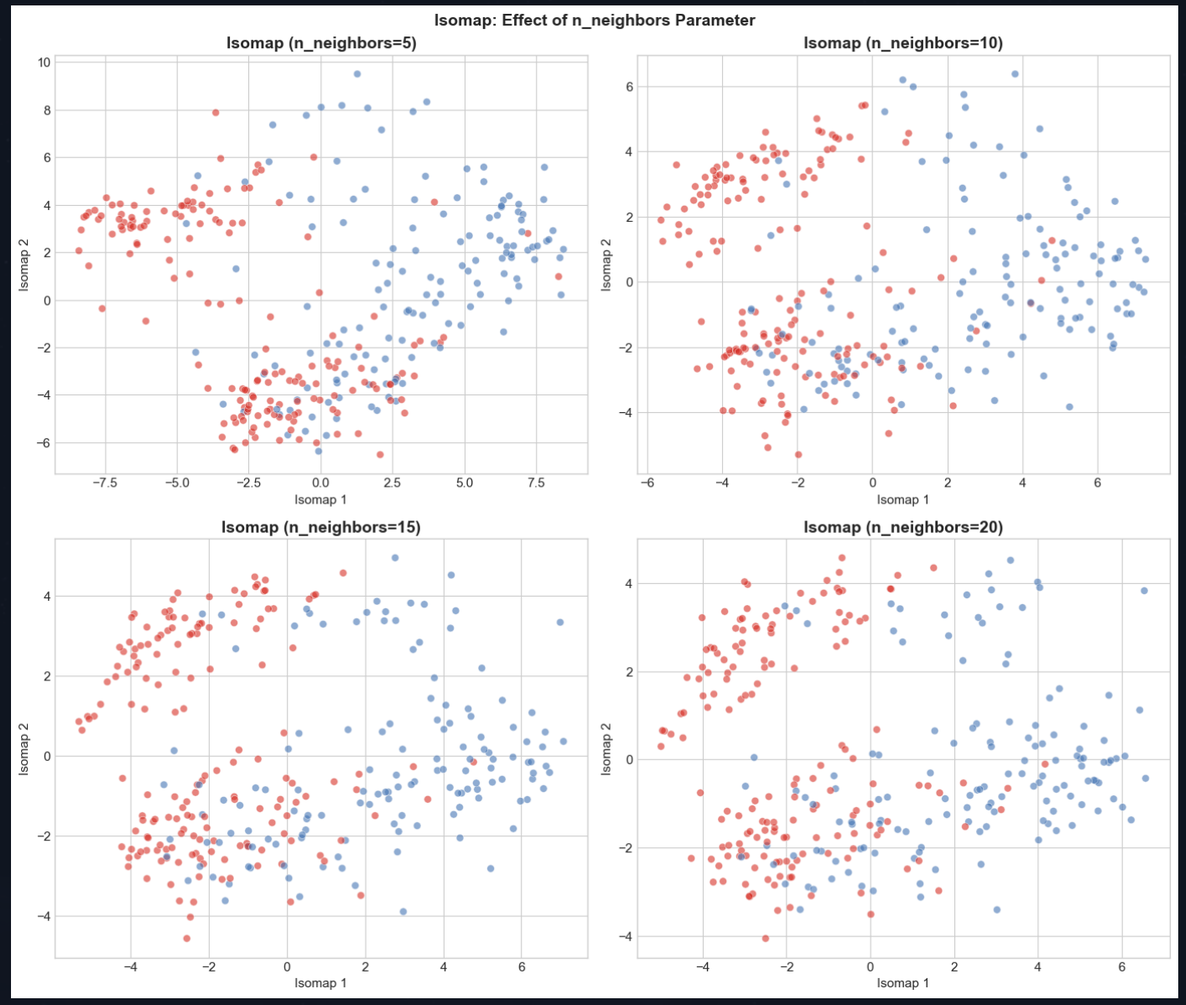
Manifold Learning: Isomap

Geodesic Distance Approach

Isomap extends classical MDS by using geodesic distances along the data manifold rather than Euclidean distances.

Key Findings

- Reveals heart disease data lies on a curved manifold
- Disease and healthy populations occupy different manifold regions
- Higher neighbor values produce smoother embeddings
- Excellent preservation of global structure

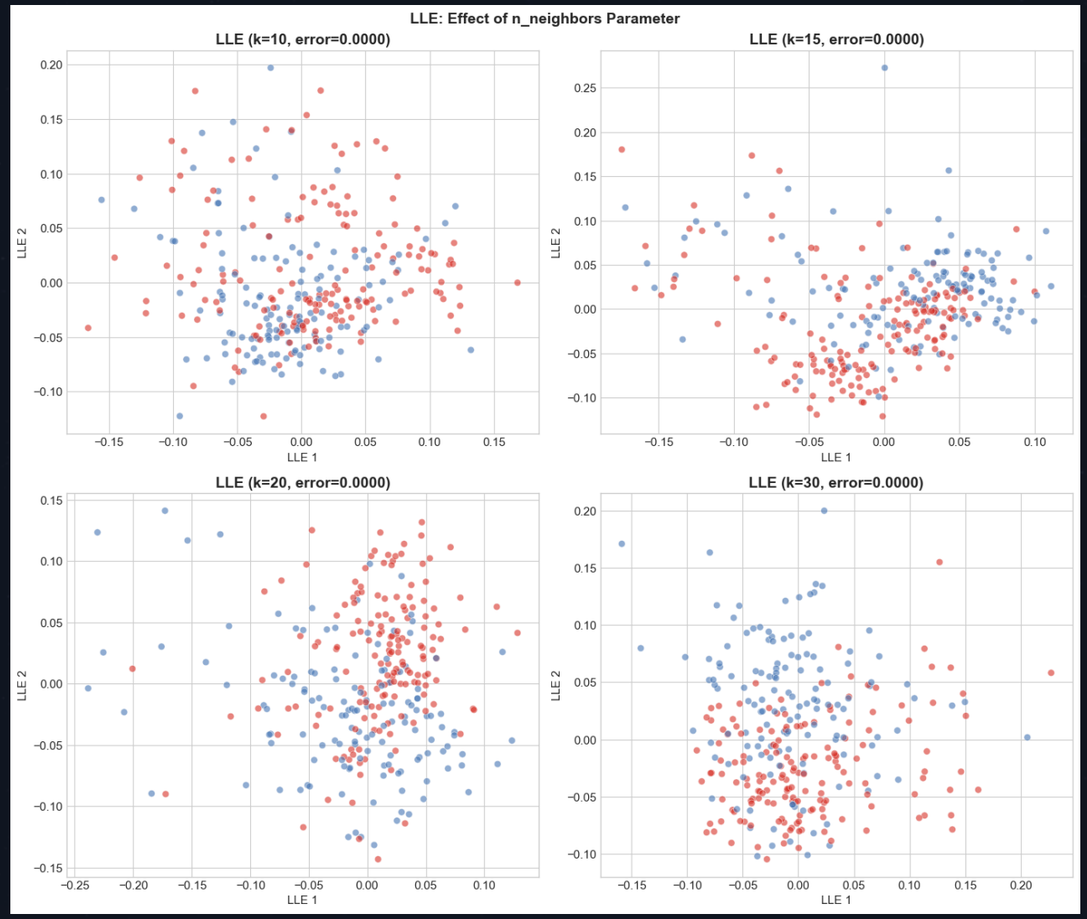


Locally Linear Embedding (LLE)

LLE preserves local geometry by reconstructing each point as a weighted linear combination of combination of its k-nearest neighbors, revealing fine-grained patient clusters within broader within broader disease categories.

$$\min_{\mathbf{Y}} \sum_{i} \left| \mathbf{y}_i - \sum_{j} w_{ij} \mathbf{y}_j \right|^2$$

The optimization finds low-dimensional coordinates that preserve local neighborhood relationships from high-dimensional space.



Comprehensive Method Comparison

PCA (Linear)

Fast computation, good global structure, but limited local structure preservation. Only 33% variance captured.

t-SNE (Nonlinear)

Excellent local structure, limited global preservation. Slow but achieves clearest class separation.

UMAP (Nonlinear)

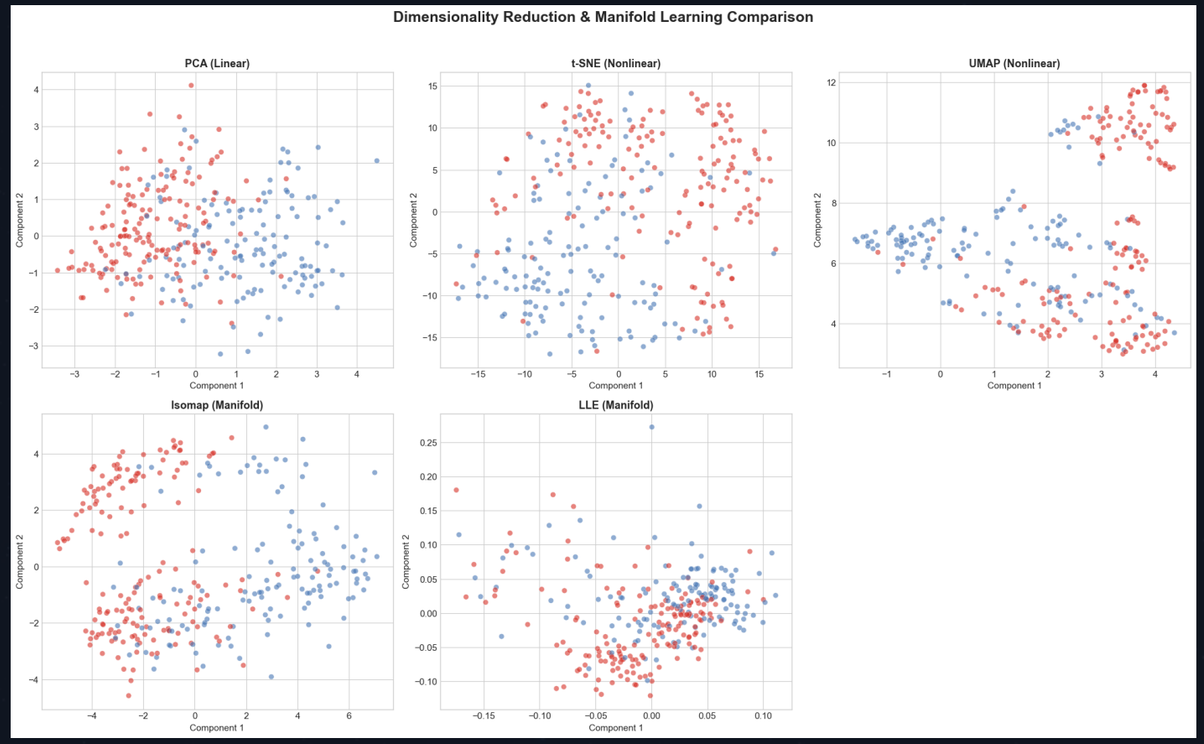
Good balance of local and global structure with moderate speed. Best overall performance.

Isomap (Manifold)

Excellent global structure via geodesic distances. Reveals curved manifold geometry.

LLE (Manifold)

Excellent local structure preservation. Fast computation, reveals fine-grained patient subgroups.

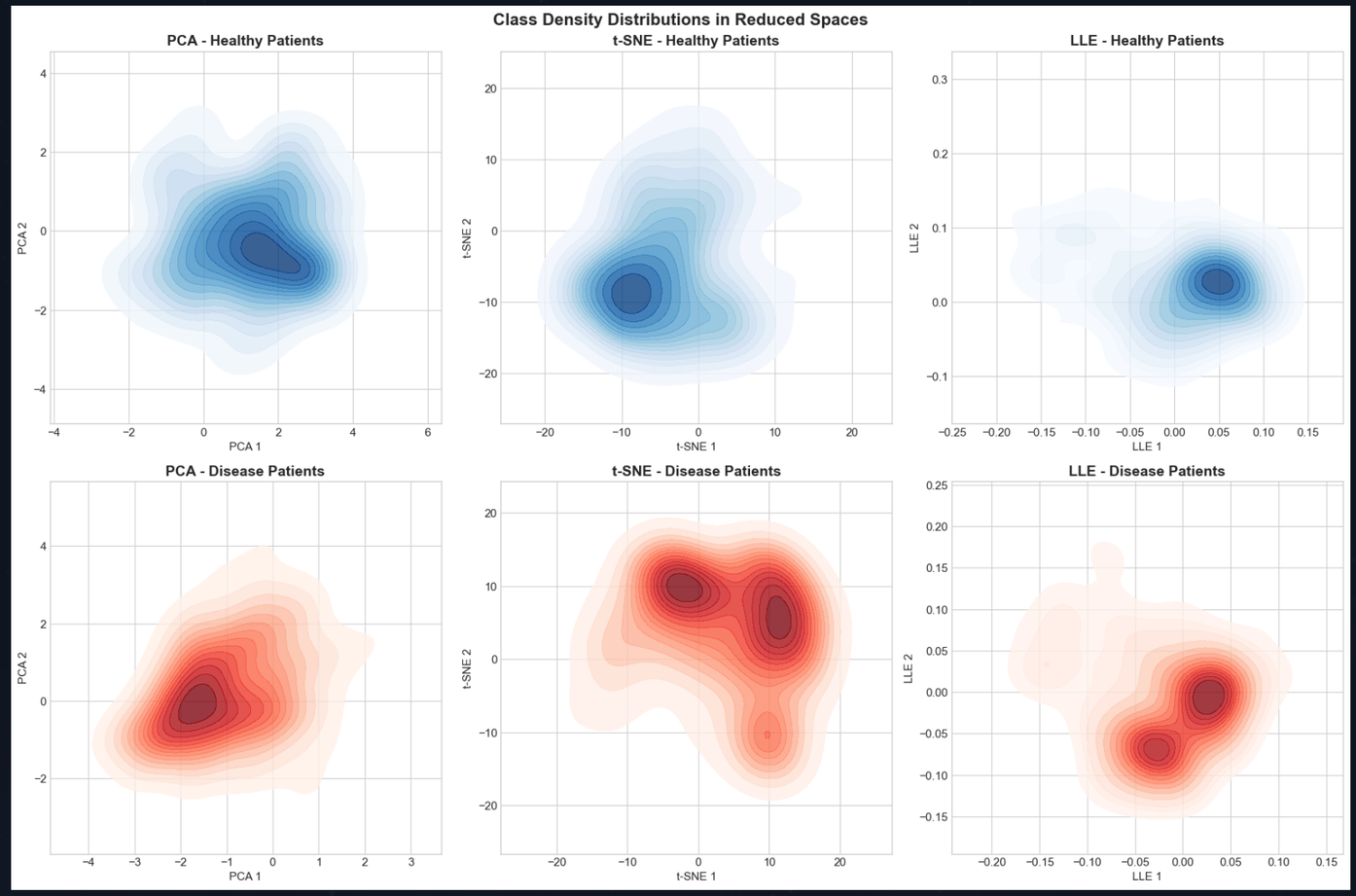


Class Density Distributions

Separation Quality

Density plots reveal how well each method method separates healthy (blue) and disease (red) populations in reduced dimensional space.

t-SNE achieves the clearest separation with distinct bimodal distributions for the disease class.



Clinical Implications

The distinct density patterns suggest that:

- Heart disease manifests as identifiable identifiable patient subgroups
- Nonlinear methods better capture disease complexity
- Multiple disease subtypes or progression stages may exist

Feature Correlation & Clinical Insights

Top Positive Predictors

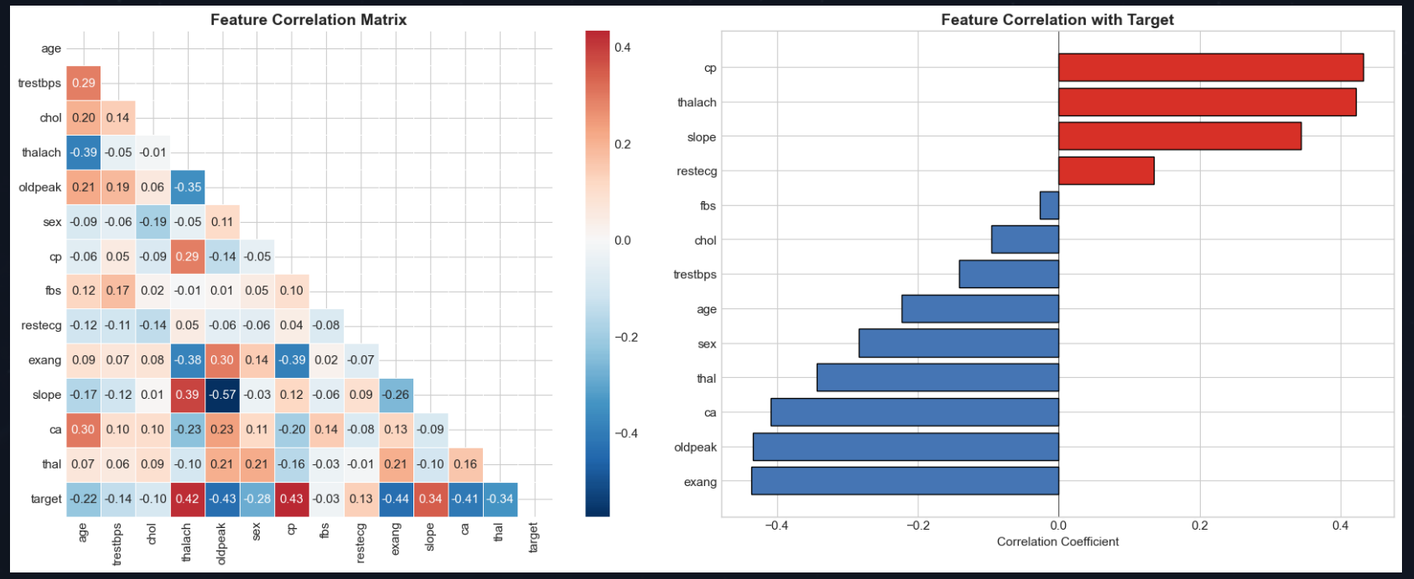
- **Chest pain type (cp):** +0.43 correlation
- **Max heart rate (thalach):** +0.42 correlation
- **ST depression (oldpeak):** Strong indicator

Negative Predictors

- **Exercise-induced angina (exang):** -0.44 correlation
- Strong inverse relationship with disease presence
- Validates clinical importance of stress testing

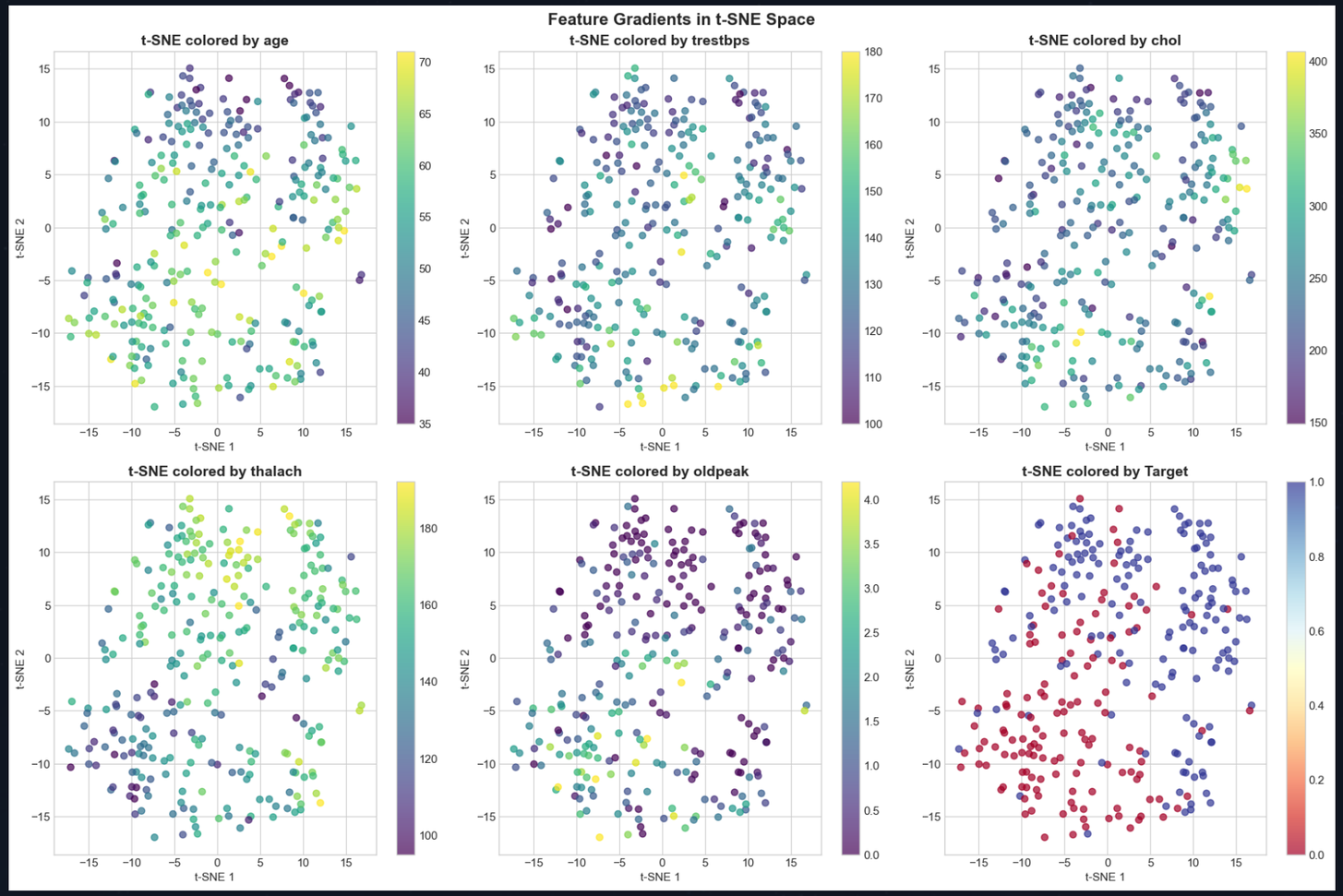
Feature Interactions

- Age correlates with blood pressure and cholesterol
- Complex multivariate relationships require nonlinear analysis
- No single feature provides complete diagnostic picture



Feature Gradients in Embedding Space

Feature gradient visualization reveals which clinical measurements drive cluster separation in t-SNE space. Clear gradients for thalach and oldpeak indicate these features strongly influence the underlying cluster structure.



Ethical Considerations & Bias Analysis

Critical Bias Findings

Sex Imbalance

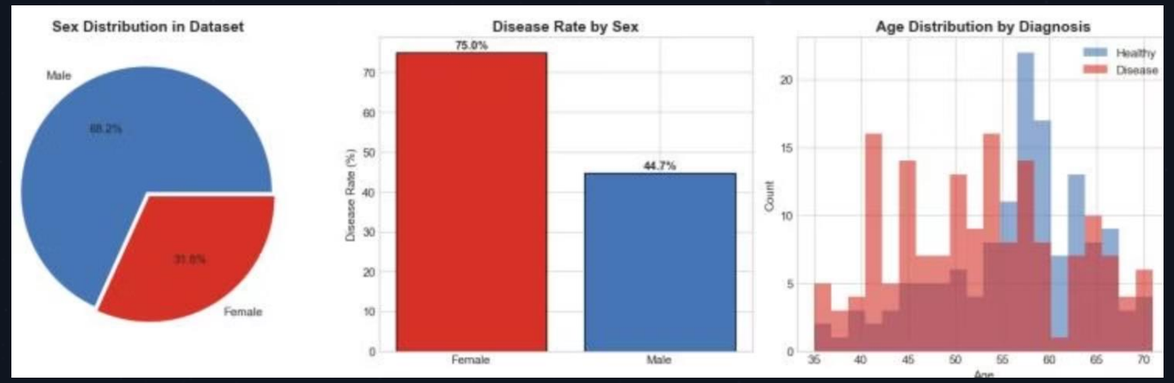
68.2% male sample limits generalizability to female population

Differential Disease Rates

75.0% disease rate in females vs 44.7% in males may reflect selection bias

Age Concentration

Data concentrated in 50-65 age range, limiting applicability to other age groups



Privacy & Compliance

Protected Health Information (PHI) under HIPAA requires proper de-identification, access identification, access controls, and GDPR compliance for European applications.

Key Insights

1. Disease patients consistently show lower maximum heart rate across all age groups compared to healthy individuals
2. Healthy patients cluster near zero ST depression, while disease patients exhibit higher oldpeak values
3. t-SNE reveals hidden cluster structure with clearer separation between disease and healthy groups than linear methods
4. Disease risk exists on a continuous spectrum rather than as a as a binary classification, with gradual transitions visible in PCA in PCA space

Conclusions & Recommendations

Data Quality is Critical

70.5% duplicate removal fundamentally changed the analysis landscape, demonstrating the importance of rigorous preprocessing

Nonlinear Methods Excel

t-SNE and UMAP reveal complex patterns invisible to PCA, achieving superior class separation and uncovering patient subgroups

Clinical Validation

Maximum heart rate and ST depression emerge as most informative predictors, validating exercise stress testing importance

Address Bias

Demographic imbalances require diverse data collection, regular fairness audits, and transparent limitation documentation

Responsible Deployment

Implement differential privacy, maintain human oversight, and ensure GDPR/HIPAA compliance for production systems