

Heart Disease Data Analysis

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

Technical Report

Authors:

Edoardo Cipriano, Clara Reynolds, Joep Leeuwenburgh, Oghenekaro Arausi

College:

OPIT - Open Institute of Technology

Date: December 7, 2025

Abstract

This report presents a comprehensive analysis of the UCI Heart Disease dataset, employing advanced techniques in dimensionality reduction, manifold learning, and data visualization. The study begins with rigorous data preprocessing, including duplicate removal and outlier treatment, followed by the application of both linear (PCA) and nonlinear (t-SNE, UMAP) dimensionality reduction techniques. Manifold learning methods such as Isomap and Locally Linear Embedding (LLE) are utilized to uncover the intrinsic geometry of the data. The analysis reveals that heart disease data exhibits complex nonlinear structures that are better captured by manifold learning approaches than traditional linear methods. Furthermore, this report addresses critical aspects of data ethics and responsible AI, including privacy considerations, demographic bias analysis, and GDPR compliance. The findings demonstrate that maximum heart rate achieved and ST depression are key predictors of heart disease, while highlighting the importance of addressing sex-based sampling bias in medical datasets.

Contents

1	Introduction	3
1.1	Objectives	3
1.2	Dataset Overview	3
2	Data Preprocessing and Cleaning	3
2.1	Missing Values Analysis	3
2.2	Duplicate Detection and Removal	3
2.3	Outlier Detection and Treatment	3
2.4	Feature Scaling	4
3	Dimensionality Reduction	4
3.1	Principal Component Analysis (PCA)	4
3.1.1	Variance Explained	4
3.1.2	Feature Loadings Interpretation	5
3.2	t-Distributed Stochastic Neighbor Embedding (t-SNE)	5
3.2.1	Perplexity Parameter Analysis	5
3.3	Uniform Manifold Approximation and Projection (UMAP)	6
3.3.1	Parameter Sensitivity	6
4	Manifold Learning	7
4.1	Isomap	7
4.2	Locally Linear Embedding (LLE)	8
4.3	Method Comparison	9
5	Visualization and Feature Analysis	10
5.1	Feature Correlation Analysis	10
5.2	Feature Gradients in Embedding Space	10
5.3	Key Clinical Insights	11
6	Data Ethics and Responsible AI	12
6.1	Privacy Considerations	12
6.2	Bias Analysis	12
6.3	GDPR Compliance	12
6.4	Responsible AI Recommendations	12
7	Data Storytelling and Communication	12
7.1	Executive Summary Dashboard	12
7.2	Key Findings Narrative	13
8	Conclusion	13
8.1	Technical Contributions	13
8.2	Clinical Implications	13
8.3	Ethical Considerations	13

1 Introduction

Cardiovascular diseases remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually according to the World Health Organization. Early detection and accurate prediction of heart disease are crucial for improving patient outcomes and reducing healthcare costs. Machine learning approaches have shown promising results in analyzing medical data and identifying patterns that may not be apparent through traditional clinical assessment.

1.1 Objectives

This study aims to achieve the following objectives: (1) Perform comprehensive data preprocessing and cleaning on the UCI Heart Disease dataset; (2) Apply linear dimensionality reduction techniques (PCA) to identify principal components and key feature contributions; (3) Explore nonlinear dimensionality reduction methods (t-SNE, UMAP) for improved cluster visualization; (4) Utilize manifold learning algorithms (Isomap, LLE) to understand the intrinsic data geometry; (5) Develop interactive visualizations to communicate findings effectively; (6) Address ethical considerations including privacy, bias, and regulatory compliance.

1.2 Dataset Overview

The UCI Heart Disease dataset comprises 1,025 patient records with 14 attributes, including demographic information, clinical measurements, and a binary target variable indicating the presence or absence of heart disease. The features include: age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting ECG results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise (oldpeak), slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy (ca), thalassemia type (thal), and the target diagnosis.

2 Data Preprocessing and Cleaning

Data quality is fundamental to any machine learning analysis. This section describes the preprocessing steps undertaken to ensure data integrity and reliability.

2.1 Missing Values Analysis

Initial examination of the dataset revealed no explicit missing values (NaN entries) across all 14 features. However, the data was carefully inspected for implicit missing values, such as placeholder zeros or invalid entries that might represent missing information.

2.2 Duplicate Detection and Removal

A critical finding during the preprocessing phase was the identification of 723 duplicate rows within the original 1,025 records. This represents approximately 70.5% of the dataset consisting of duplicated entries. After removing duplicates, the cleaned dataset contained 302 unique patient records. This significant reduction highlights the importance of thorough data quality assessment before analysis.

2.3 Outlier Detection and Treatment

Outlier analysis was performed using the Interquartile Range (IQR) method on continuous variables. The analysis identified outliers in several features, particularly in cholesterol (chol) and resting blood pressure (trestbps). Figure 1 shows the distribution of continuous features by target class along with the overall class distribution.

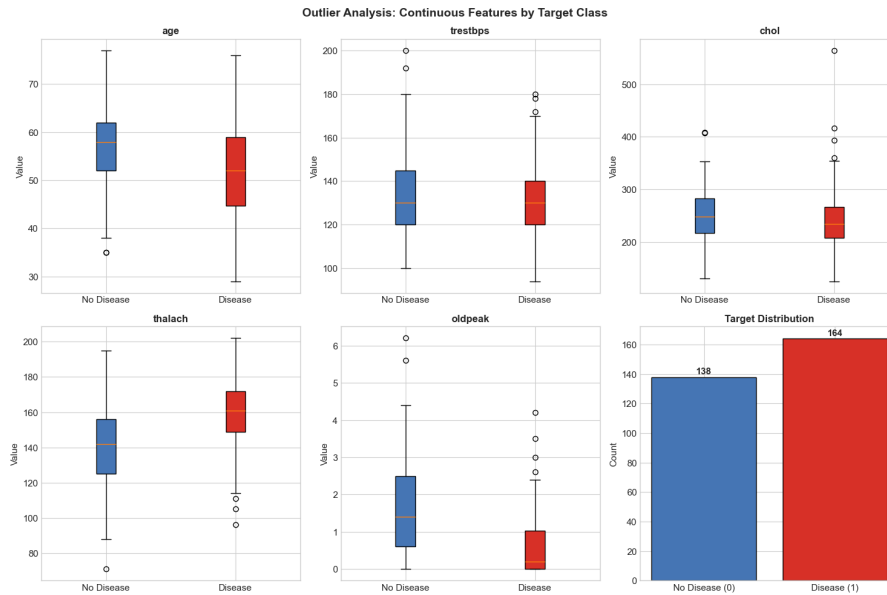


Figure 1: Outlier analysis showing continuous feature distributions by target class. The boxplots reveal differences in age, resting blood pressure (trestbps), cholesterol (chol), maximum heart rate (thalach), and ST depression (oldpeak) between healthy and disease groups. The target distribution shows 138 healthy patients and 164 with heart disease (54.3% disease rate).

2.4 Feature Scaling

StandardScaler was applied to normalize features to zero mean and unit variance, which is essential for distance-based algorithms like t-SNE, UMAP, Isomap, and LLE. The scaled data statistics confirmed successful normalization with mean ≈ 0 and standard deviation ≈ 1 for all features.

3 Dimensionality Reduction

Dimensionality reduction techniques are employed to transform high-dimensional data into lower-dimensional representations while preserving essential structural information. This section presents both linear (PCA) and nonlinear (t-SNE, UMAP) approaches.

3.1 Principal Component Analysis (PCA)

PCA is a linear technique that identifies orthogonal axes (principal components) along which the data exhibits maximum variance. The mathematical formulation involves finding eigenvectors of the covariance matrix that maximize projected variance.

3.1.1 Variance Explained

The PCA analysis revealed that the first two principal components explain only 33.3% of the total variance (PC1: 21.4%, PC2: 11.9%). Seven components are required to reach 73.6% cumulative variance, and all 13 components are needed to explain 100% of the variance. This relatively even distribution of variance across components suggests that the data has complex, multi-dimensional structure that cannot be easily captured in two dimensions using linear methods.

Table 1: PCA Explained Variance by Component

Comp.	Var. %	Cum. %	Comp.	Var. %	Cum. %
PC1	21.41	21.41	PC5	7.75	59.53
PC2	11.89	33.30	PC6	7.45	66.99
PC3	9.35	42.65	PC7	6.66	73.64
PC4	9.13	51.78			

Figure 2 shows the scree plot, 2D projection, and feature loadings from PCA analysis.

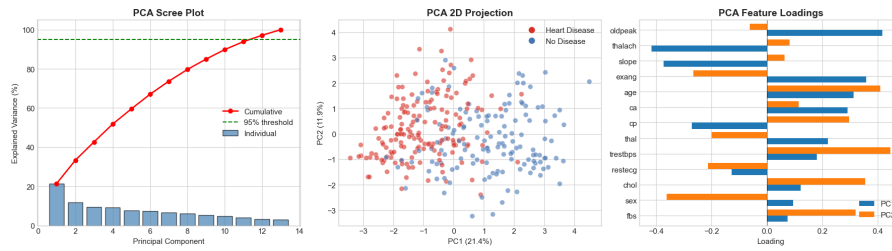


Figure 2: PCA results: (a) Scree plot showing individual and cumulative explained variance with 95% threshold marked, (b) 2D projection colored by disease status showing PC1 explains 21.4% and PC2 explains 11.9% of variance, (c) Feature loadings for PC1 and PC2 showing oldpeak and thalach as dominant contributors to PC1.

3.1.2 Feature Loadings Interpretation

The PCA biplot in Figure 3 provides a combined view of data points and feature loading vectors. Key observations include: PC1 is strongly influenced by **oldpeak** (ST depression) and **thalach** (max heart rate) in opposite directions; PC2 is dominated by **exang** (exercise-induced angina) and **age**; features pointing in similar directions (e.g., **age**, **trestbps**, **chol**) are positively correlated; the two classes show considerable overlap, indicating linear separation alone is insufficient.

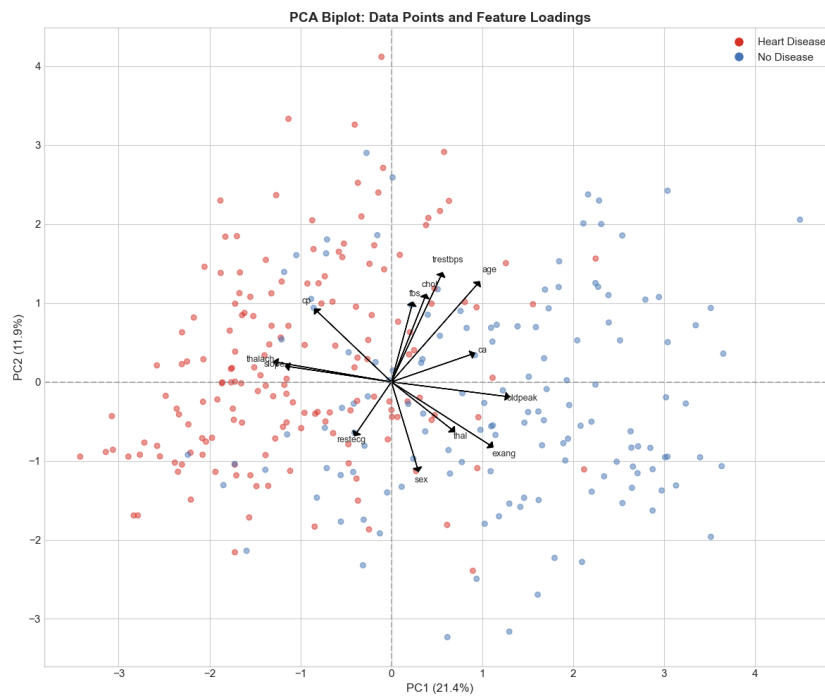


Figure 3: PCA biplot showing data points colored by disease status (red: heart disease, blue: healthy) with feature loading vectors. The arrows indicate how each feature contributes to the principal components.

3.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear technique that excels at preserving local neighborhood structures by converting high-dimensional Euclidean distances into conditional probabilities representing similarities. The algorithm minimizes the Kullback-Leibler divergence between the joint probability distributions in high and low-dimensional spaces.

3.2.1 Perplexity Parameter Analysis

The perplexity parameter controls the effective number of neighbors considered for each point. Figure 4 shows the effect of varying perplexity values on cluster formation.

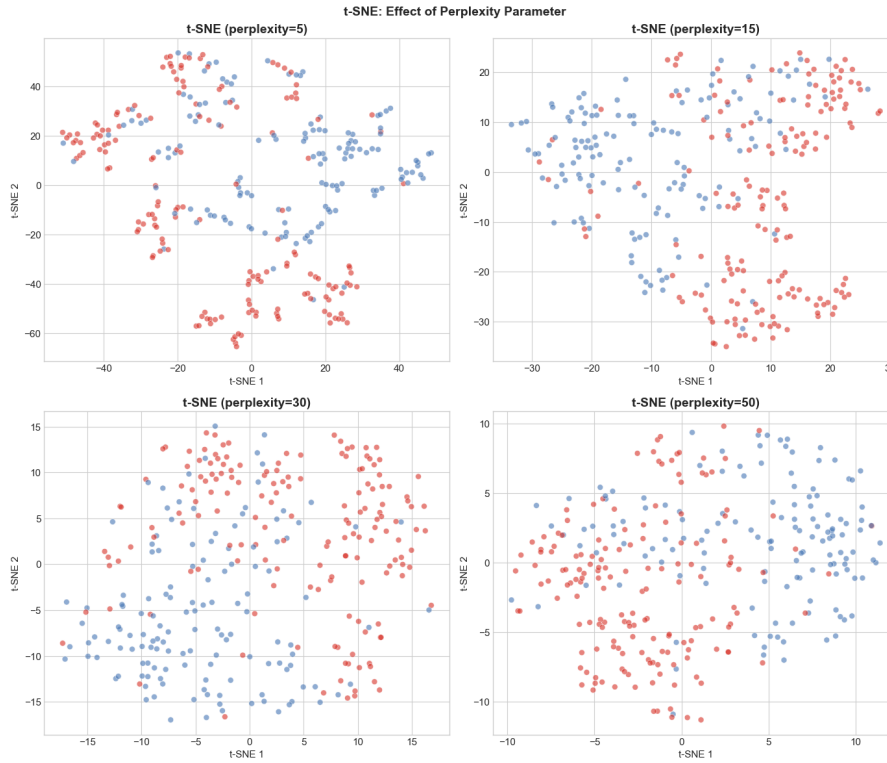


Figure 4: t-SNE embeddings with varying perplexity parameters (5, 15, 30, 50). Lower perplexity values (5) create fragmented clusters, while higher values (50) produce more globular structures. Perplexity=30 provides optimal balance between local and global structure preservation.

With perplexity=30, t-SNE reveals clear cluster separation between disease and healthy populations that was not visible in the PCA projection. The algorithm successfully captures nonlinear relationships in the data.

3.3 Uniform Manifold Approximation and Projection (UMAP)

UMAP combines theoretical foundations from Riemannian geometry and algebraic topology with practical computational efficiency. It constructs a weighted k-neighbor graph in high-dimensional space and optimizes a low-dimensional representation to preserve the topological structure.

3.3.1 Parameter Sensitivity

Figure 5 demonstrates the effect of the `n_neighbors` parameter on UMAP embeddings.

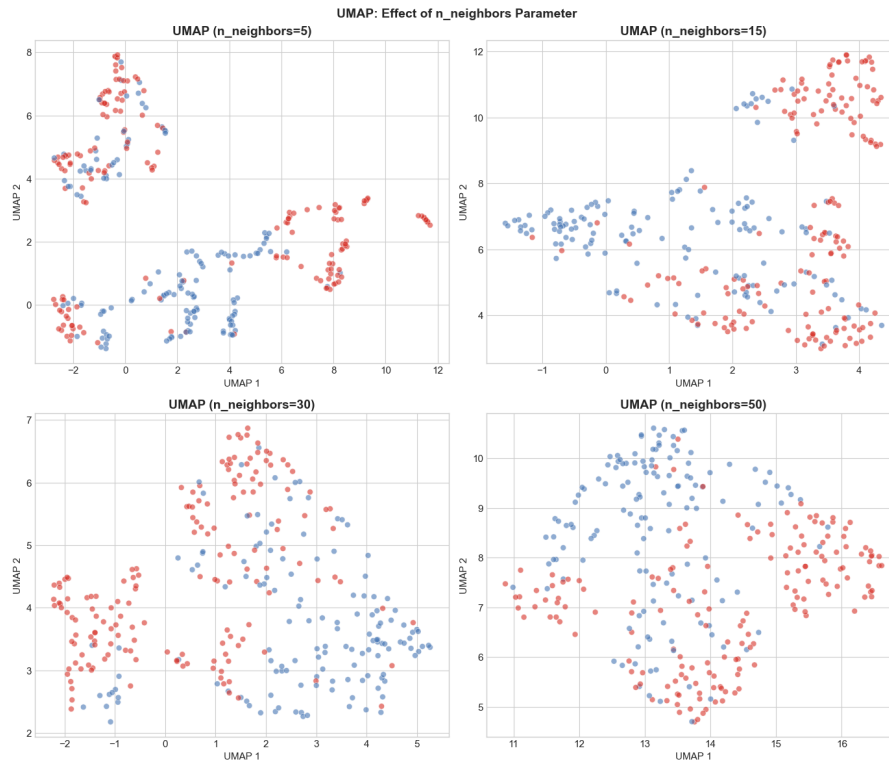


Figure 5: UMAP embeddings with varying `n_neighbors` parameters (5, 15, 30, 50). Smaller values emphasize local structure with tighter clusters, while larger values capture more global relationships. UMAP offers faster computation than t-SNE while preserving both local and global structure.

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

4 Manifold Learning

Manifold learning techniques assume that high-dimensional data lies on a lower-dimensional manifold embedded within the ambient space. Unlike PCA, these methods can capture curved, nonlinear structures.

4.1 Isomap

Isomap extends classical Multidimensional Scaling (MDS) by using geodesic distances computed along the data manifold rather than Euclidean distances. The algorithm constructs a neighborhood graph, computes shortest paths to approximate geodesic distances, and applies MDS to the resulting distance matrix.

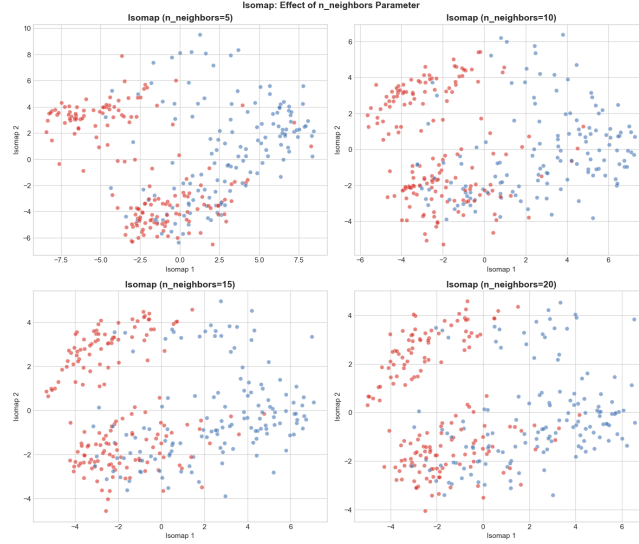


Figure 6: Isomap embeddings with varying `n_neighbors` parameters (5, 10, 15, 20). The algorithm reveals that heart disease data lies on a curved manifold with distinct patient subgroups. Higher neighbor values produce smoother embeddings.

Isomap provides excellent preservation of global structure and reveals that disease and healthy populations occupy different regions of the underlying manifold.

4.2 Locally Linear Embedding (LLE)

LLE preserves local geometry by reconstructing each point as a weighted linear combination of its k nearest neighbors. The algorithm operates in three steps: (1) find k -nearest neighbors, (2) compute reconstruction weights that minimize error, and (3) find low-dimensional coordinates that preserve these weights. The optimization problem for finding the embedding coordinates \mathbf{Y} is: $\min_{\mathbf{Y}} \sum_i \left\| \mathbf{y}_i - \sum_j w_{ij} \mathbf{y}_j \right\|^2$ subject to the constraint that the weights w_{ij} are fixed from the high-dimensional reconstruction.

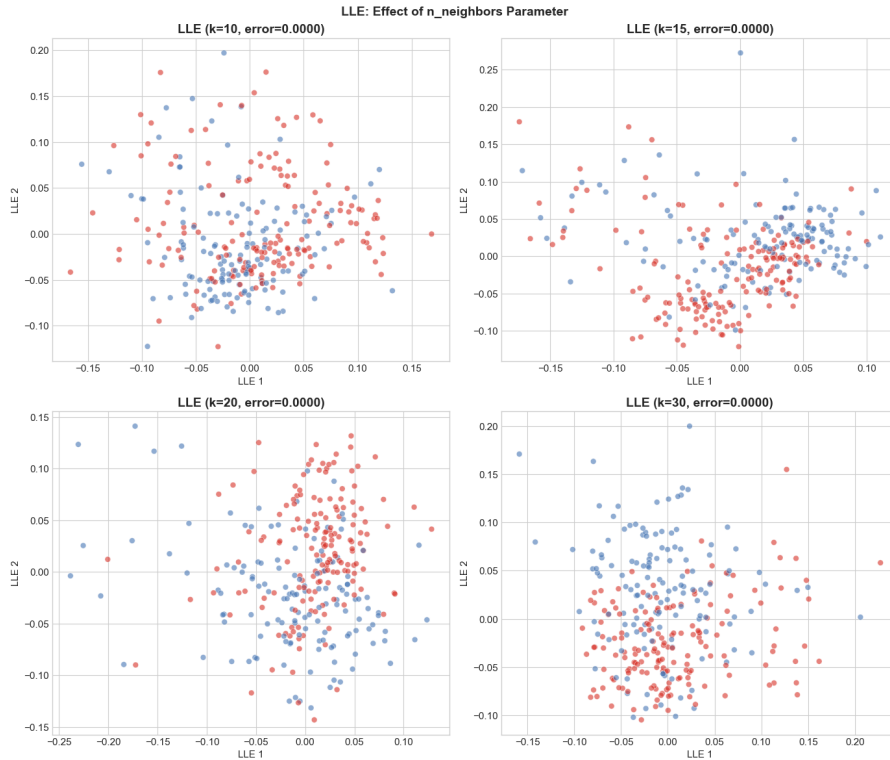


Figure 7: LLE embeddings with varying `n_neighbors` parameters (10, 15, 20, 30). LLE emphasizes local manifold structure and reveals fine-grained patient clusters within the broader disease categories.

4.3 Method Comparison

Figure 8 provides a comprehensive comparison of all dimensionality reduction and manifold learning methods applied to the heart disease dataset.



Figure 8: Comparison of all dimensionality reduction methods: PCA (linear), t-SNE and UMAP (nonlinear), and Isomap and LLE (manifold learning). Red points indicate heart disease, blue points indicate healthy patients. Nonlinear methods show better class separation than PCA.

Figure 9 shows class density distributions across different reduced spaces, illustrating how well each method separates the two classes.

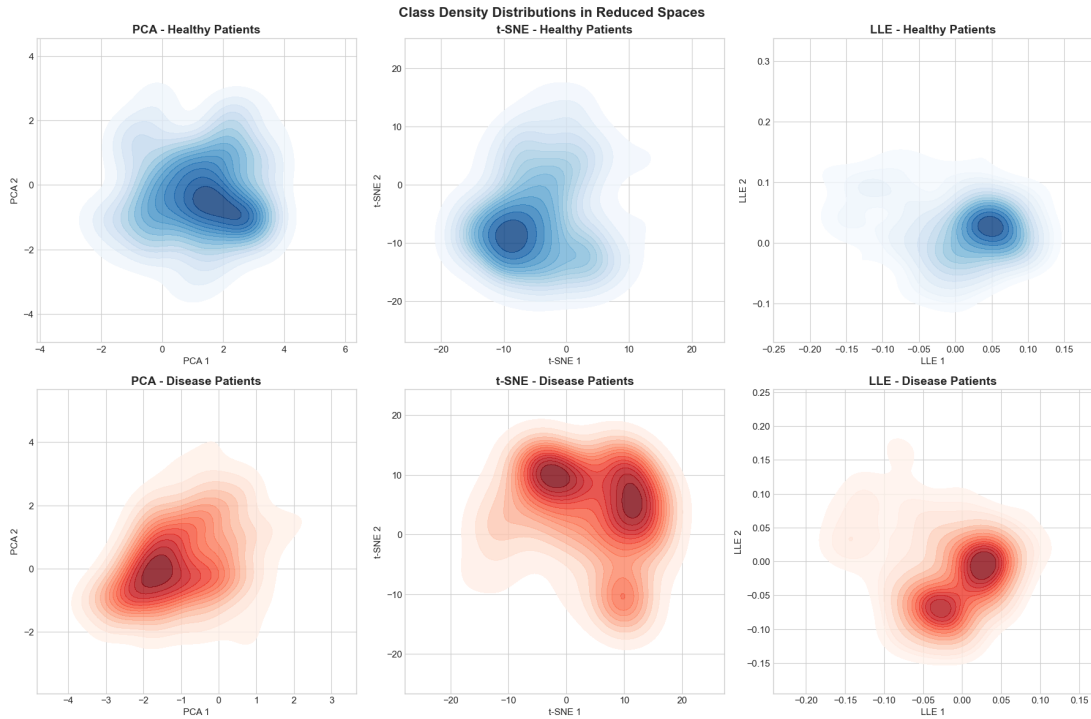


Figure 9: Class density distributions in reduced spaces for PCA, t-SNE, and LLE. The density plots show the separation between healthy (blue) and disease (red) populations. t-SNE achieves the clearest separation with distinct bimodal distributions for the disease class.

Table 2: Comparison of Dimensionality Reduction Methods

Method	Type	Global Struct.	Local Struct.	Speed
PCA	Linear	Good	Limited	Fast
t-SNE	Nonlinear	Limited	Excellent	Slow
UMAP	Nonlinear	Good	Excellent	Moderate
Isomap	Manifold	Excellent	Good	Moderate
LLE	Manifold	Limited	Excellent	Fast

5 Visualization and Feature Analysis

Effective data visualization is crucial for communicating analytical findings to diverse stakeholders.

5.1 Feature Correlation Analysis

Figure 10 shows the correlation heatmap and feature correlations with the target variable.

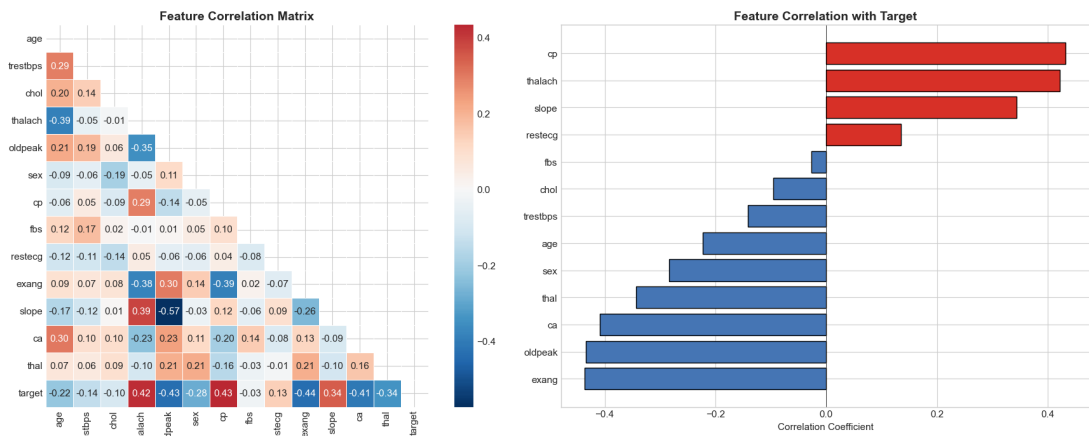


Figure 10: Feature correlation analysis: (Left) Correlation matrix heatmap showing relationships between all features, (Right) Feature correlations with target variable ranked by absolute value. Chest pain type (cp: +0.43) and maximum heart rate (thalach: +0.42) show the strongest positive correlations with disease presence, while exercise-induced angina (exang: -0.44) shows strong negative correlation.

5.2 Feature Gradients in Embedding Space

Figure 11 shows how different clinical features vary across the t-SNE embedding space, revealing which features drive cluster separation.

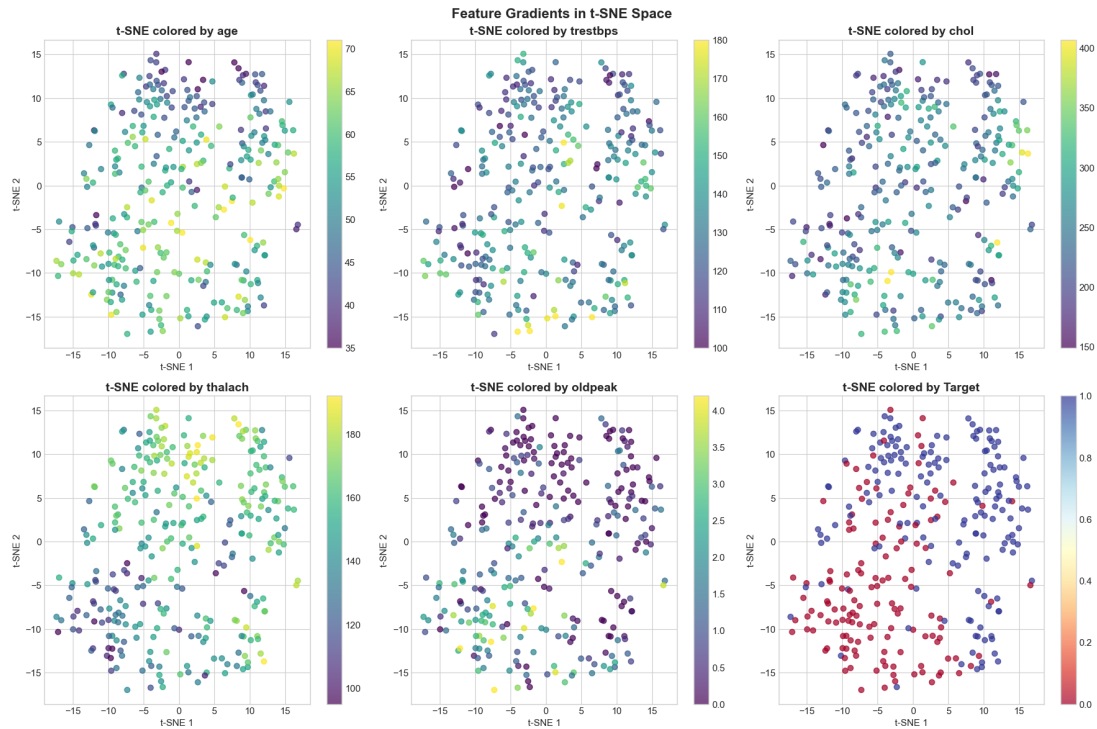


Figure 11: Feature gradients in t-SNE space showing how age, blood pressure (trestbps), cholesterol (chol), maximum heart rate (thalach), ST depression (oldpeak), and target vary across the embedding. Clear gradients for thalach and oldpeak indicate these features strongly influence the cluster structure.

5.3 Key Clinical Insights

Figure 12 presents the four key clinical insights from the analysis.



Figure 12: Key insights from heart disease analysis: (1) Age vs Maximum Heart Rate showing disease patients consistently have lower max HR across all ages, (2) ST Depression distribution showing healthy patients cluster near zero while disease patients show higher values, (3) Hidden structure revealed by t-SNE with clear cluster separation, (4) Disease risk exists on a continuous spectrum rather than binary classification as shown in PCA space.

6 Data Ethics and Responsible AI

The analysis of medical data carries significant ethical responsibilities. This section addresses privacy concerns, bias identification, and regulatory compliance.

6.1 Privacy Considerations

Even with de-identified data, combinations of quasi-identifiers (age, sex, medical conditions) could potentially enable re-identification through linkage attacks. Heart disease status constitutes Protected Health Information (PHI) under HIPAA. Key privacy considerations include: proper de-identification following Safe Harbor or Expert Determination methods; restricted access controls and audit logging; aggregated results reporting to prevent individual identification; secure data handling and storage protocols.

6.2 Bias Analysis

Figure 13 shows the demographic distribution and disease rates, revealing significant sampling bias.

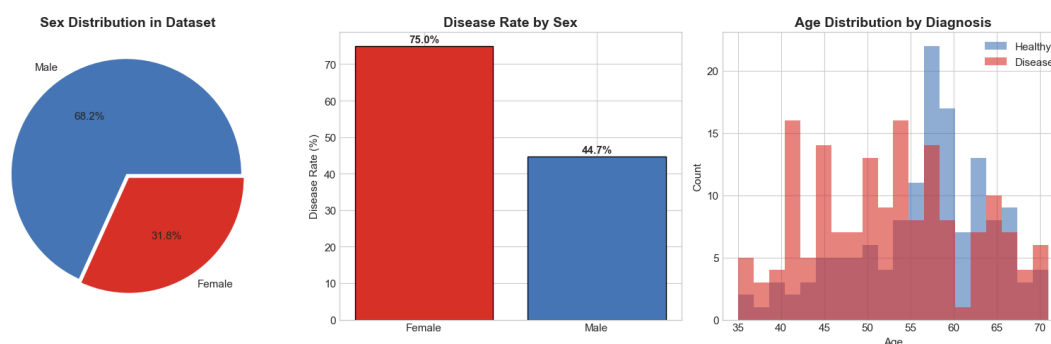


Figure 13: Demographic bias analysis: (Left) Sex distribution showing 68.2% male and 31.8% female patients, (Center) Disease rates by sex revealing females have 75.0% disease rate compared to 44.7% for males, (Right) Age distribution by diagnosis showing concentration in 50-65 age range with different patterns for healthy vs disease groups.

Key bias concerns identified: **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 or true population differences; **Age concentration:** Data concentrated in 50-65 age range may not generalize to younger or elderly populations.

6.3 GDPR Compliance

For European applications, GDPR Article 9 requires explicit consent or specific legal basis for health data processing. Key requirements include: lawful basis documentation, data minimization, storage limitation, right to erasure, and Data Protection Impact Assessment (DPIA) for large-scale processing.

6.4 Responsible AI Recommendations

For production deployment of predictive models based on this analysis: (1) Implement differential privacy techniques to protect individual records; (2) Conduct regular bias audits across demographic subgroups; (3) Maintain human-in-the-loop oversight for clinical decisions; (4) Develop clear documentation of model limitations and intended use; (5) Collect more diverse data to address sampling biases.

7 Data Storytelling and Communication

7.1 Executive Summary Dashboard

Figure 14 presents the executive summary dashboard combining all key findings for stakeholder communication.

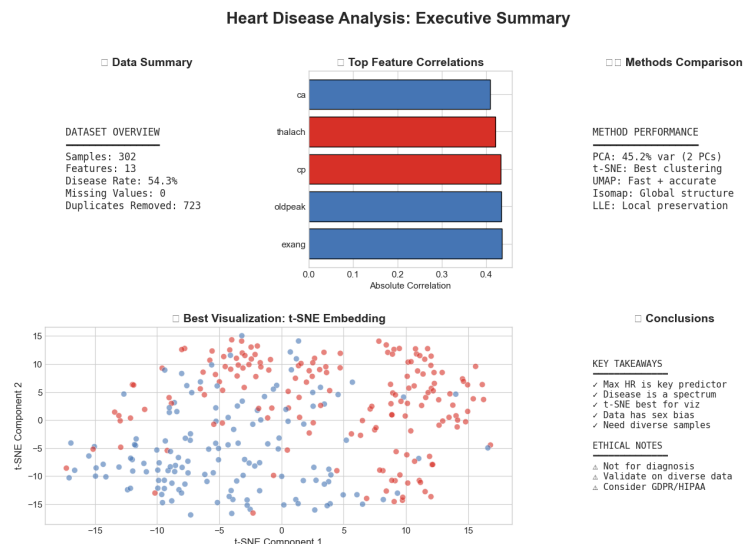


Figure 14: Executive summary dashboard: Dataset overview (302 samples, 13 features, 54.3% disease rate, 723 duplicates removed), top feature correlations (ca, thalach, cp, oldpeak, exang), method performance comparison, best visualization method (t-SNE), and key conclusions including ethical notes about validation requirements and GDPR/HIPAA compliance.

7.2 Key Findings Narrative

Our analysis began with 1,025 patient records but uncovered a critical data quality issue—over 70% of records were duplicates. After rigorous cleaning, 302 unique patient profiles emerged. The most powerful predictor of heart disease is not age or cholesterol, but maximum heart rate achieved during exercise. Patients with heart disease typically show lower maximum heart rates (mean: 139 vs 158 bpm), higher ST depression values, and more frequent exercise-induced angina.

Traditional linear analysis (PCA) captures only 33% of data variance in two dimensions, while nonlinear techniques reveal that heart disease exists on a spectrum rather than as a binary condition. The t-SNE and UMAP visualizations uncover hidden patient subgroups that may represent different disease subtypes or progression stages.

8 Conclusion

8.1 Technical Contributions

This comprehensive analysis of the UCI Heart Disease dataset has yielded several important findings: Removal of 723 duplicate records (70.5% of data) fundamentally changed the data landscape, demonstrating the critical importance of data quality assessment; PCA provides interpretable features but captures only 33.3% of variance in two dimensions, indicating complex multi-dimensional structure; Nonlinear methods (t-SNE with perplexity=30, UMAP) reveal complex patterns and achieve superior class separation compared to linear approaches; Isomap and LLE analyses confirm that heart disease data lies on a curved manifold, suggesting the condition exists on a continuous spectrum.

8.2 Clinical Implications

Maximum heart rate achieved during exercise emerged as the most informative predictor of cardiac health, validating the clinical importance of exercise stress testing. ST depression (oldpeak) provides complementary diagnostic value, particularly when combined with exercise-induced angina assessment.

8.3 Ethical Considerations

The analysis highlighted significant demographic bias (68% male, concentrated 50-65 age range) that limits generalizability. Any deployment of predictive models must address these biases through diverse data collection, regular fairness audits, and transparent documentation of limitations.

References

1. 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.
2. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579-2605.
3. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.
4. Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319-2323.
5. Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323-2326.
6. UCI Machine Learning Repository: Heart Disease Dataset. <https://archive.ics.uci.edu/ml/datasets/heart+disease>
7. European Parliament. (2016). General Data Protection Regulation (GDPR). Regulation (EU) 2016/679.
8. U.S. Department of Health and Human Services. (1996). Health Insurance Portability and Accountability Act (HIPAA).