UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Project Ingegneria Informatica

Data Analytics for Digital Health
Analysis of Hospitalized Patients

Relatore:

Anna Monreale

Francesca Naretto

Candidato:

Alexander Mittet

Dominik Garstenauer

ANNO ACCADEMICO 2025/2026

Indice

1 Data Understanding and Preparation

In this chapter we will first analyze the four given medical datasets. They cover *Heart Diagnoses*, *Laboratory Events*, *Microbiology Events*, and *Procedure Codes*. After a general overview of the data semantics, we will outline the processing steps we took in order to obtain a cleaned and unified patient profile.

1.1 Data Understanding

The four datasets comprise 4864×25, $978{,}503 \times 14$, $15{,}587 \times 14$, and $14{,}497 \times 6$ rows and columns, respectively. An initial exploratory analysis was conducted to assess data semantics, cohort composition, and feature distributions.

Figure 1 presents the exploratory analysis for the laboratory and diagnostic cohorts. The distributions indicate a population skewed toward older age groups with a high prevalence of cardiac-related comorbidities, predominantly associated with ICD codes *I50* and *I21*. Laboratory quality control flags are largely dominated by *OK* status, while *Glucose* and *Potassium* emerge as the most frequently observed laboratory measurements.

Figure 2 summarises the data understanding metrics for the microbiology and procedural datasets. The microbiology records are primarily characterised by the detection of *Escherichia coli* and *Staphylococcus aureus*, corresponding with frequent use of antibiotics such as *Gentamicin* and *Trimethoprim*. Procedural records are dominated by cardiovascular interventions, most notably coronary arteriography and percutaneous transluminal coronary angioplasty.

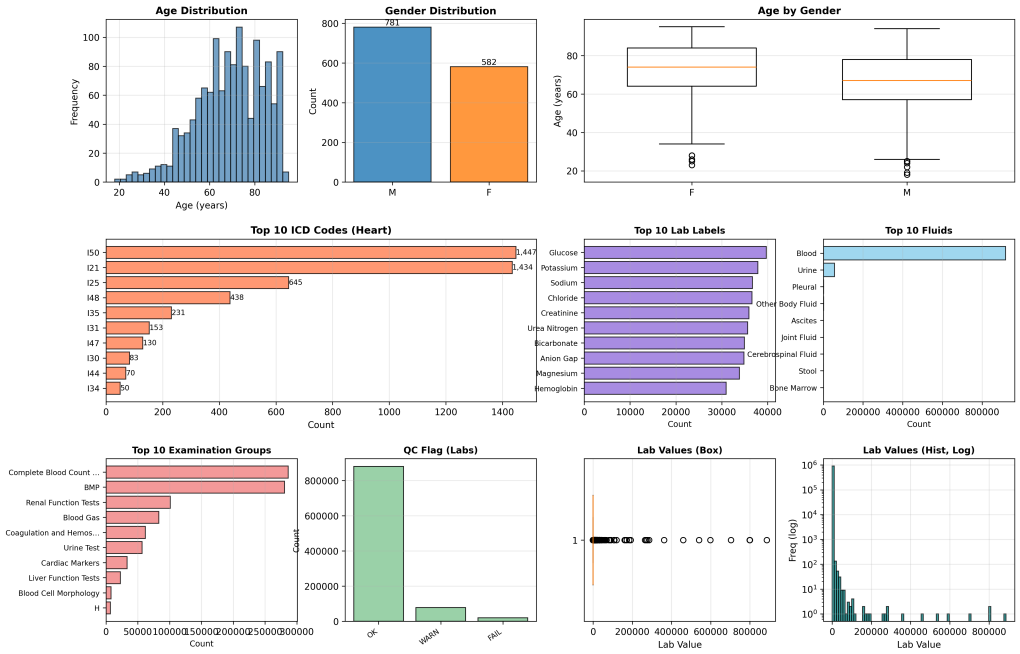


Figura 1: Exploratory data analysis for laboratory and diagnostic cohorts.

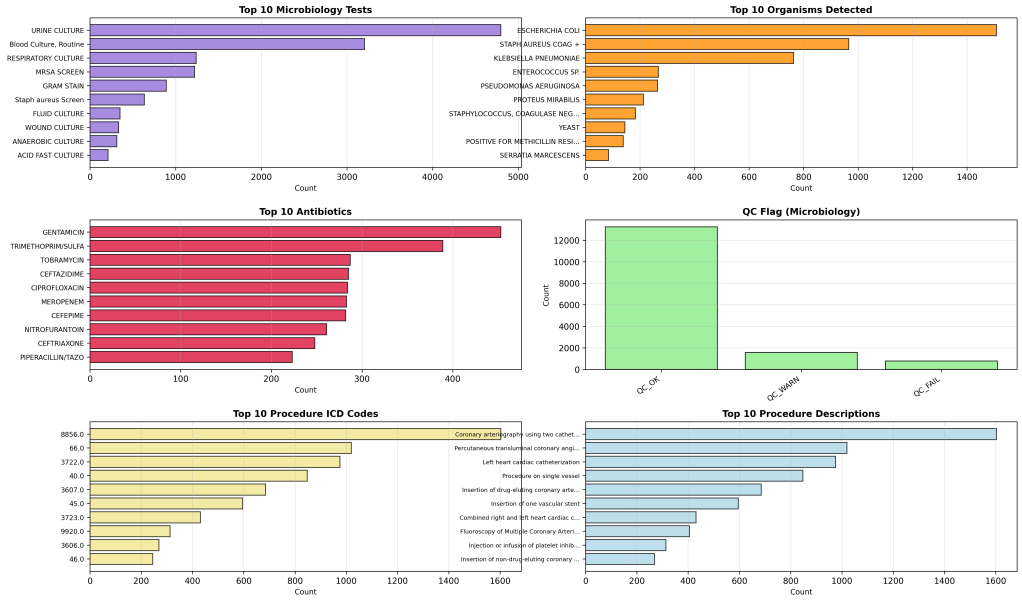


Figura 2: Data understanding metrics for microbiology and procedural records.

Further Data Semantics and Missingness Analysis Tables 1 and 2 summarise the numerical semantics and missingness patterns across all datasets. The analysis focuses on numeric columns with available summary statistics and variables exhibiting non-zero missingness, respectively.

Tabella 1: Numeric column semantics across datasets.

Dataset	Column	Rows	Min	Max	Mean	Std
Heart	age	4864	18.0	95.0	68.98	14.97
Heart	anchor_year	4864	2110	2206	2155.6	23.40
Micro	dilution_value	15587	0.06	512	7.06	21.25
Proc	icd_code	14497	12	272366	4771	9216
Labs	value	978503	-743	886449	56.1	2223
Labs	ref_range_lower	978503	0	2200	31.7	44.8
Labs	ref_range_upper	978503	0	100000	55.8	400.5
Labs	valuenum	978503	-743	886449	67.3	2176

Tabella 2: Columns with non-zero missingness.

Dataset	Column	Missing (%)	Count
Heart	dod	91.82	4466
Labs	ref_range	85.06	832288
Heart	age	71.98	3501
Heart	anchor_year	71.98	3501
Heart	gender	71.98	3501
Micro	dilution_value	69.78	10876
Micro	interpretation	69.08	10767
Micro	org_name	65.41	10196
Labs	flag	64.88	634816
Proc	icd_code	18.62	2699
Labs	value	14.23	139255
Labs	valuenum	7.27	71186

The numeric semantics in Table 1 reveal substantial scale heterogeneity across datasets, with laboratory measurements exhibiting extreme ranges and

high variance. The high standard deviation of value and valuenum is to be expected as it contains multiple units of measurement and fluids. This highlights the need to carefully handle these columns in the data processing steps. As to be expected by a clinical dataset, we see that the mean age of our patients is almost 69. The anchor_year seems to be anonymized as the ranges are roughly 200 years in the future.

Table 2 highlights systematic missingness concentrated in outcome-related, microbiology, and reference-range variables, indicating structural sparsity rather than random absence. Especially age and gender reveal a high missingness which needs to be addressed. These patterns motivate downstream strategies including feature aggregation, robustness-aware normalization, and explicit missingness encoding.

1.2 Data Preparation

The preprocessing stage across the four datasets (DF1: heart, DF2: laboratory, DF3: microbiology, and DF4: procedure codes) focused on standardising data types, resolving issues arising from non-standard missing value representation, and enforcing quality control measures essential for robust feature engineering.

1.2.1 Standardisation of Missing Values and Data Types

All datasets required extensive cleaning of non-standard or "wrong" missing values. For DF1, DF2, and DF3, this process began by removing all newline characters (\n) and surrounding whitespaces. Specific entries were identified and converted to np.NaN or pd.na. This included converting entries in non-numeric columns that exactly matched *['-']*, *['/']*, *['.']*, *[]*, or *[':']*. Furthermore, any non-numeric column entry containing (case-insensitive) strings such as *"none"*, *"nan"*, *"na"*, *"N/A"*, or *"."* was converted to np.NaN. The primary motivation for this step was to prevent the injection of low-quality data and ensure that subsequent numerical operations were accurate. DF4 was noted to have a cleaner starting point regarding the prevalence of these non-standard missing indicators.

For time related data, *charttime*, *storetime*, and *dod* (date of death) across all relevant datasets (DF1, DF2, DF3, DF4) were explicitly converted to date data types to facilitate temporal analysis. DF1 also required specific handling for *age*, *anchor_year*, and *note_seq*, which were stored as *int64* but contained trailing *.0* values.

1.2.2 Numerical Extraction and Quality Control

Since DF2 had two and DF3 had one exactly duplicated rows, the duplicates were dropped. A unique and complex challenge in DF2 (laboratory) and DF3 (microbiology) was the extraction of numerical information from the non-numeric *value* column to populate missing entries in the less sparse *valuenum* column. Where ranges were encountered, such as *'80-160'*, the midpoint was calculated. For comparison entries (e.g., *>1.050* or *<1*), a numeric value was

estimated by adding or subtracting 0.1 to the comparison boundary. Any remaining unparsed entries were coerced to NaN.

Quality control was enforced in DF2 and DF3 by inspecting the *qc_flag*. Since approximately 2% of rows were flagged as FAIL, the corresponding *valuenum_merged* values for these rows were set to np.nan, ensuring that low-quality results were excluded from the analysis. DF2 also corrected cases where the abnormal flag indicator did not align with the computed numeric range of *valuenum_merged*. DF3 performed additional checks to ensure that *dilution_text* matched the *dilution_value* and *dilution_comparison*.

1.2.3 Specific Imputations (DF1)

DF1 required advanced handling of demographic missingness. Missing values in *dod* were inferred as "not dead" to create the *is_dead* variable. The *gender* variable was manually imputed using text evidence from *hpi reports* or *physical_exam*, employing a multi-step keyword resolution strategy to address conflicts, leaving only six NaNs.

1.2.4 Consolidated and Aggregated Features

Features were aggregated using the unique key (*subject_id*, *hadm_id*). DF1 had no duplicates; DF2 received *subject_id* via *hadm_id* from DF1. A binary *is_dead* indicator was created from the *dod* column. DF2 labels were grouped by fluid type and aggregated using max (min for Hemoglobin/Hematocrit to capture anemia). Exam indicators (*has_xray*, *has_ct*, *has_ultrasound*, *has_cath*, *has_ecg*, *has_mri*) were summed into *imaging_variety*. Documentation complexity was computed as the log-sum of text lengths from HPI, physical exam, chief complaint, and reports. ICD-10 codes were used to create *icd_cat* for cardiac conditions: heart failure (I50), cardiac arrest (I46), arrhythmia (I44-I49), valvular (I34-I36), inflammatory (I32-I33, I40), and acute MI (I21-I22). Binary flags (*has_hf*, *has_arr*, *has_ami*, *has_arrest*, *has_valvular*, *has_inflammatory*) were created and summed into *cardiac_comorbidity_score*. Numerical values from textual ranges were extracted as midpoints; comparison values were offset by ± 0.1.

1.2.5 Patient Profile Construction and Feature Engineering

Following the initial data cleaning and feature aggregation within the individual datasets, the pre-processed notebooks were merged to form a single, comprehensive patient profile base.

1.2.6 Data Integration and Completeness Analysis

An essential step following the merge was the analysis of data completeness across the four sources, as detailed in Figure 3. The data source availability showed that the *heart* (4,864 patients) and *labs* (4,855 patients) datasets were the most frequently available, while *micro* (2,756 patients) was the least available. The co-occurrence matrix reveals that the *heart* and *labs* data were almost universally present together (4,855 patients), and the *procedure* data was also highly correlated with *heart* data (3,459 patients). Crucially,

the *Completeness Score Distribution* indicates that over 2,000 patients had data available from exactly three sources, and nearly 2,000 patients had data available from all four sources, suggesting a high degree of integration for the majority of the cohort.

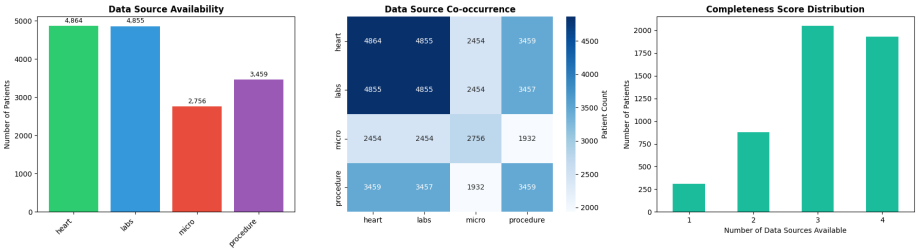


Figura 3: Data Source Availability and Completeness Analysis.

1.2.7 Composite Feature Engineering and Clinical Interpretation

To create patient profiles encapsulating acute physiological status, composite features were calculated using robust Z-score standardisation: $Z(X) = (X - \text{Median}(X))/(\text{IQR}(X) + \epsilon)$. The composite features are:

$$\text{Micro Resistance} = \text{resistant\_ratio} \cdot \ln(1 + \text{unique\_org} + \text{unique\_spec})$$

$$\text{Procedure Density} = \text{total\_proc}/\max(1, \text{span\_days}) \quad \text{Diagnosis Burden} = \ln(1 + \text{n\_diag})$$

$$\text{Metabolic Stress} = Z(\text{Gluc}) + Z(\text{Lact}) + Z(\text{AG}) - Z(\text{Bicarb})$$

$$\text{Oxygenation Dysf.} = -Z(\text{pO}_2) + Z(\text{pCO}_2) - Z(\text{pH}) - Z(\text{BE}) \quad \text{Renal Failure Idx} = Z(\text{Cr}_\text{S}) - Z(\text{Cr}_\text{U})$$

$$\text{Hematologic Stab.} = Z(\text{Hgb}) + Z(\text{Hct}) + Z(\text{RBC}) - Z(\text{RDW})$$

$$\text{Diagnostic Intens.} = \ln(1 + \text{Cnt}_\text{BG} + \text{Cnt}_\text{C} + \text{Cnt}_\text{L} + \text{Cnt}_\text{CBC}) \quad \text{Recent Admission} = 1/(1 + \text{days\_last\_adm})$$

$$\text{Inflam Liver} = Z(\text{CRP}) + Z(\text{AST}) + Z(\text{ALT}) + Z(\text{LD})$$

$$\text{Renal Injury} = Z(\text{Cr}) + Z(\text{BUN}) + Z(\text{Phos}) + Z(\text{K})$$

Clinical Interpretation: Micro Resistance quantifies infection complexity; *Procedure Density* reflects care intensity; *Diagnosis Burden* measures comorbidity; *Metabolic Stress* captures metabolic instability; *Renal Injury* reflects kidney function; *Oxygenation Dysfunction* summarises respiratory failure; *Inflammation/Liver Stress* measures hepatic injury; *Hematologic Stability* assesses RBC lineage; *Renal Failure Index* detects concentration defects; *Diagnostic Intensity* indicates acuity; *Recent Admission* captures chronic instability.

1.2.8 Profile Selection for Downstream Tasks

After generating the composite features, two distinct profiles were created to support the downstream clustering and classification tasks. Feature selection was performed through correlation analysis on the combined set, and highly correlated variables were removed to ensure feature independence and reduce multicollinearity, which is crucial for model stability and interpretability. The resulting profiles thus represent optimised subsets of features tailored to physiological segmentation (clustering) or predictive diagnosis (classification).

Clustering Profile The clustering profile was designed to capture physiological heterogeneity while minimising redundancy and multicollinearity. The selected features (**abnormal_ratio**, **qc_fail_ratio**, **fluid_diversity**, **procedure_span_days_missing**, **gender_F**, **micro_resistance_score**, **metabolic_stress_index**, **oxygenation_dysfunction_index**, **inflammation_liver_stress_index**, **hematologic_stability_score**, and **renal_failure_index**) were chosen because they summarise data quality, treatment intensity, infectious burden, and multi-organ physiological dysfunction, enabling meaningful unsupervised segmentation of patient states. on defects.

Classification Profile The classification profile extends the clustering feature set with additional variables directly related to prognosis and outcome prediction. The included features (**cardiac_comorbidity_score**, **has_hf**, **has_arrest**, **has_valvular**, **has_inflammatory**, **num_labs**, **total_procedures**, **total_microbio_events**, **unique_antibiotics**, and **is_dead**) were selected to capture comorbidity burden, clinical intervention intensity, and treatment complexity, enabling effective supervised learning for adverse outcome prediction.

By including comorbidity and intervention metrics in addition to the physiological features, the classification profile is tailored to capture both patient state and treatment complexity, improving predictive accuracy for adverse outcomes.

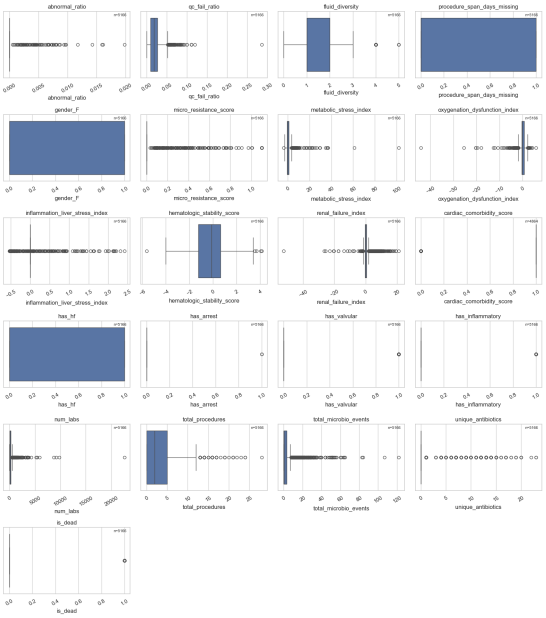
1.2.9 Distributional Characteristics and Correlation of Profiles

subcaption

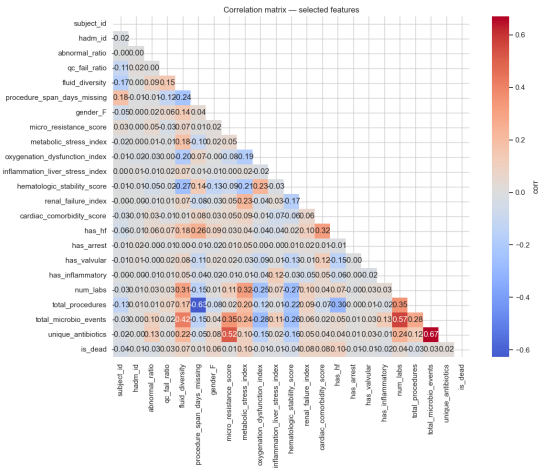
The final feature set ($N = 5,166$) exhibits high right-skewness in activity markers such as `num_labs` and `total_microbio_events`, identifying a high-acuity sub-population (Figure 4a). Engineered indices like `metabolic_stress_index` and `inflammation_liver_stress_index` show tight interquartile ranges with extreme upper-tail outliers, while organ-specific markers (`renal_failure_index` and `oxygenation_dysfunction_index`) display significant negative outliers reaching -40. These distributions capture severe physiological derangement essential for distinguishing high-risk phenotypes, while features like `gender_F` and `procedure_span_days_missing` provide balanced inputs across the 0–1 range.

1.2.10 Correlation Analysis Summary

As shown in Figure 4b, the feature set maintains a sparse correlation structure, ensuring independence for stable model performance. Notable linear associations are limited to clinically interdependent variables, such as `unique_antibiotics` and `total_microbio_events` ($r = 0.67$) and `num_labs` with `total_microbio_events` ($r = 0.57$). Most physiological and demographic markers exhibit coefficients below 0.20, confirming minimal multicollinearity across the primary clinical dimensions.



(a) Boxplots of all final features.



(b) Lower-triangle correlation matrix of the final features.

Figura 4: Distributional characteristics and correlation structure of the final clinical feature set.

2 Clustering Analysis

This chapter investigates unsupervised patient stratification using three clustering paradigms: K-means, DBSCAN, and Hierarchical Clustering. All methods were applied to the clustering patient profile introduced in the previous chapter, restricted to the 11 numerical features. To mitigate the influence of extreme values, features were scaled using *RobustScaler*.

2.1 K-means Clustering

The optimal number of clusters k was determined using multiple internal validation criteria: the Elbow Method, average Silhouette score, Davies–Bouldin index, and Calinski–Harabasz index. Across all metrics (Figure 5), $k = 2$ consistently emerged as the most favorable solution, with the Silhouette score (0.635) and Calinski–Harabasz index (≈ 2100) attaining their maxima at $k = 2$. The final K-means solution produced a strongly imbalanced partition: a small cluster of 347 patients (6.7%) representing a high-severity phenotype with elevated values across composite indices including *renal_failure_index*, *metabolic_stress_index*, *micro_resistance_score*, and *inflammation_liver_stress_index*, and a large cluster of 4,819 patients (93.3%) representing a baseline or low-burden patient group as can be seen in Figure 7a. Age distributions show nearly identical means across clusters, where cluster 0 has a mean age of 70 whereas cluster 1 has a mean age of 69, confirming that separation is driven by clinical burden rather than demographics. Notably, cluster 1 has a minimum age of 39 compared to cluster 0's mini-

mum age of 18, aligning with the expectation that younger patients typically exhibit lower clinical stress levels.

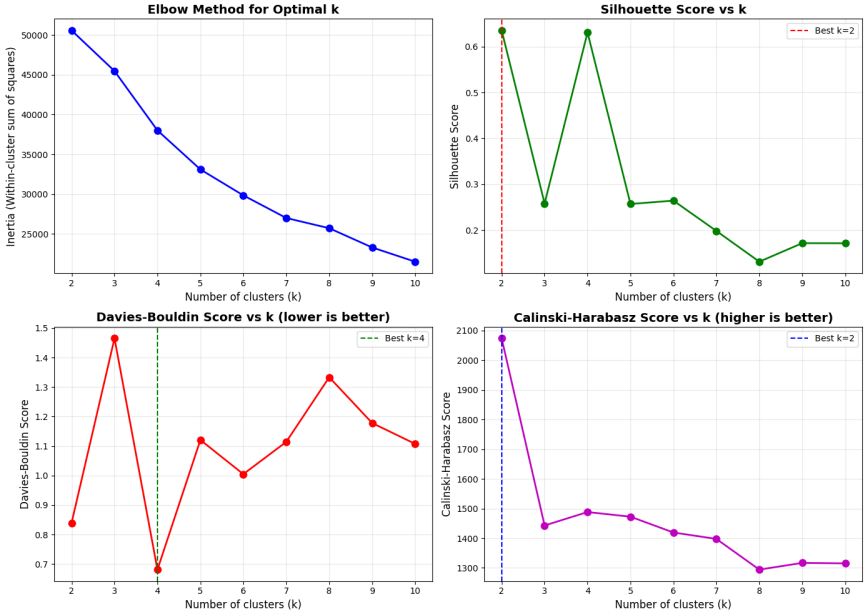


Figura 5: Internal validation metrics for K-means cluster selection.

2.2 Density-Based Clustering (DBSCAN)

DBSCAN was applied to identify dense patient subgroups while explicitly detecting outliers. Parameter selection was guided by a K-distance plot and an extensive grid search over ϵ (range 0.5–6.5) and *min_samples* (3–10), evaluated using Silhouette score, Davies–Bouldin index, and Calinski–Harabasz index. The optimal configuration ($\epsilon \approx 6.14$, $min\_samples = 5$) maximized the Silhouette score (0.868), minimized the Davies–Bouldin index (0.244), and yielded two dense clusters with a negligible noise fraction (0.17%). Centroid analysis, as illustrated in Figure 7b, reveals two clinically distinct dense regions of sizes 5,151 (Cluster 0) and 5 (Cluster 1), alongside 10 noise points (Cluster -1). Cluster 0 represents the dominant baseline population, characterized by near-zero stress indices, indicating standard clinical progression. In contrast, Cluster 1 exhibits elevated metabolic stress (+1.0) accompanied by significant negative deviations in renal (−20.0), hematologic (−0.1), and oxygenation stability (−0.1) indices. This negative polarity in failure indices indicates a state of high physiological stability or "hyper-function" in those specific organs, which is highly atypical given the increased metabolic demand. The extreme cluster imbalance—specifically the small size of Cluster 1 ($n = 5$) and the noise group—suggests these are rare clinical phenotypes or "edge cases" rather than representative subpopulations. This is further complicated by the demographic data: Cluster 1 contains only one patient with an available age (18 years old, with 80% missing data), whereas Cluster 0 is significantly older with a mean age of 69.06 (73.66% missing data).

The "youth" of the single data point in Cluster 1 likely explains the negative failure indices; younger patients often possess a higher physiological reserve, allowing them to maintain organ stability (negative failure scores) even under high metabolic stress. The noise cluster (Cluster -1) represents the most extreme physiological decoupling, showing strong negative correlations with renal failure and oxygenation dysfunction (ranging from -3 to -13) alongside a massive positive spike in metabolic stress (approximately $+21$). Clinically, this indicates a "hyper-acute" phenotype—potentially early-stage sepsis or a massive systemic inflammatory response—where the body is under extreme compensatory stress before the onset of measurable organ failure.

2.3 Hierarchical Clustering

Hierarchical clustering was performed using Ward, Complete, Average, and Single linkage methods. Ward linkage was selected due to its superior Calinski–Harabasz performance and its tendency to produce compact, clinically interpretable clusters. A Ward linkage solution with $k = 3$ was selected, yielding a Silhouette score of 0.64 and a Calinski–Harabasz score of 1564. Hierarchical clustering identifies a spectrum of severity profiles: one cluster captures patients with extreme multi-organ stress, while intermediate clusters reflect more specific dysfunction patterns, and the remaining clusters represent lower-burden or baseline profiles. Figure 7c reveals the gradient of clinical feature intensities across the hierarchical clusters, ranging from the dominant elderly cohort in Cluster 1.0 ($n = 4,817$, mean age 68.98) with moderate metabolic stress ($+3$) and high hematologic instability ($+8$), to a more stable elderly subset in Cluster 0.0 ($n = 335$, mean age 69.74) with minimal lab abnormalities. In contrast, Cluster 2.0 represents a small but distinct younger phenotype ($n = 14$, mean age 38.0) characterized by high fluid diversity and abnormal lab ratios, while Cluster 3.0 captures an acute terminal state defined by catastrophic metabolic ($+35$) and renal ($+4$) failure indices. This significant cluster imbalance highlights the model's capacity to isolate rare, high-acuity states; specifically, the negative values in renal and oxygenation indices for the younger Cluster 2.0 indicate "hyper-stable" organ function—a physiological reserve that allows younger patients to sustain high metabolic demand without immediate organ collapse.

2.4 Final Evaluation and Comparison

DBSCAN achieved the highest overall cluster quality (Silhouette 0.789, Davies–Bouldin 0.208) and uniquely identified structurally anomalous patients as noise. K-means maximized between-cluster variance (Calinski–Harabasz 2074) but imposed spherical partitions. Hierarchical clustering provided the most granular phenotyping at the cost of reduced separation quality.

UMAP visualization (Figure 9) confirms that all three methods identify similar patient subgroups, with DBSCAN's noise points distributed across the

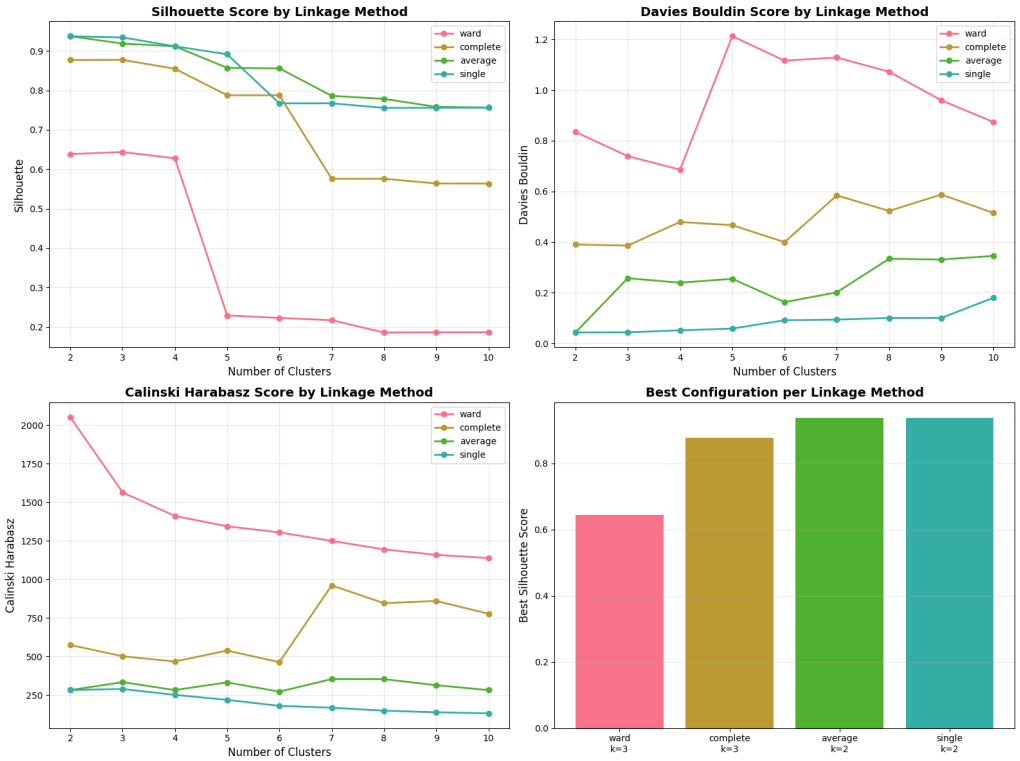
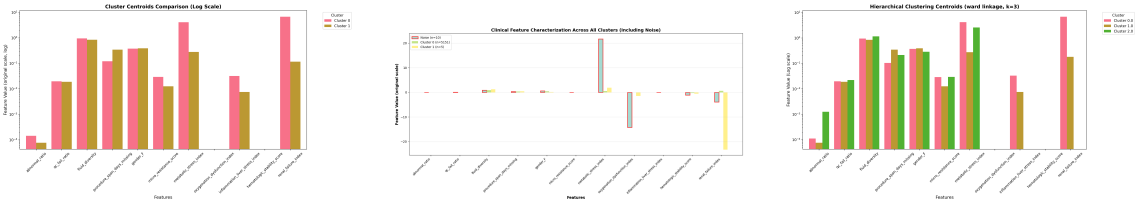


Figura 6: Internal validation metrics across linkage methods.



(a) K-means (b) DBSCAN (c) Hierarchical

Figura 7: Cluster centroids comparison across all three clustering methods, revealing clinical feature differentiation within each partition strategy.

embedding space and hierarchical clustering capturing intermediate severity transitions.

2.5 Conclusion

DBSCAN provides the most robust separation of dense patient populations from clinically anomalous outliers and is therefore selected as the preferred method for identifying structurally distinct patient phenotypes. Hierarchical clustering is advantageous when finer-grained sub-phenotyping is required, while K-means effectively captures the dominant binary severity split.

3 Classification Analysis

This chapter presents a supervised classification analysis aimed at distinguishing ischemic from non-ischemic cardiovascular conditions using the derived patient profiles.

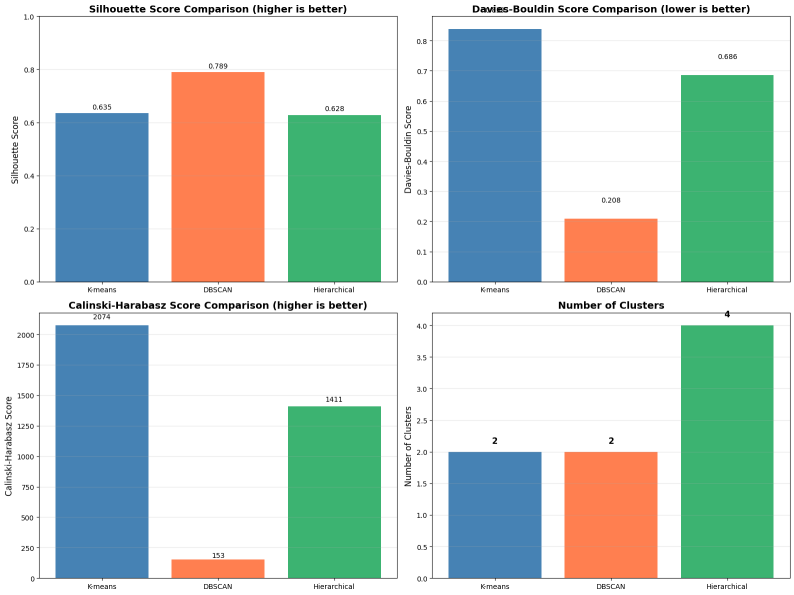


Figura 8: Comparison of internal validation metrics across clustering methods.

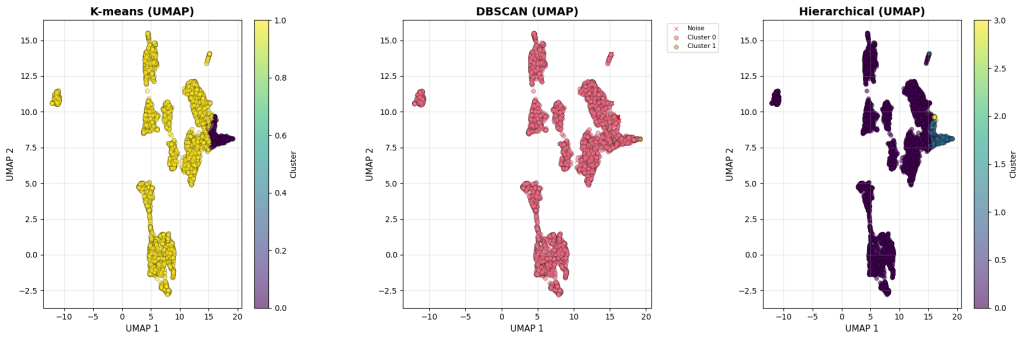


Figura 9: UMAP-based comparison of clustering methods, revealing consistent patient separation patterns across all three approaches.

3.1 Objective and Label Definition

The objective of this stage is to construct a robust binary classifier separating ischemic (`Class 1`) from non-ischemic (`Class 0`) cardiovascular cases. Class labels were derived from primary ICD diagnoses, where the ischemic class was defined by the presence of ICD codes I20, I21, I22, I24, or I25. Admissions with multiple diagnoses were assigned to `Class 1` if at least one ischemic code was present.

3.2 Data Preparation and Model Training

3.2.1 Pre-processing

Classification was performed on the cleaned dataset described in Section 1, with categorical variables already encoded. The *age* variable was excluded due to extensive missingness and the risk of data leakage through ICD-dependent imputation. No class rebalancing was applied, as the class ratio remained

moderate (Class 0 / Class 1 = 1.18). The final dataset was split into training ($n = 3513$) and test ($n = 879$) subsets.

3.2.2 Model Suite

Six classification models were evaluated to capture diverse modeling assumptions: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosting. This selection balances interpretability, non-linearity, and ensemble-based robustness.

3.3 Model Evaluation

Model performance was assessed on the held-out test set using Accuracy, Balanced Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion matrices. Cross-validation was additionally performed to evaluate stability.

3.3.1 Performance Comparison

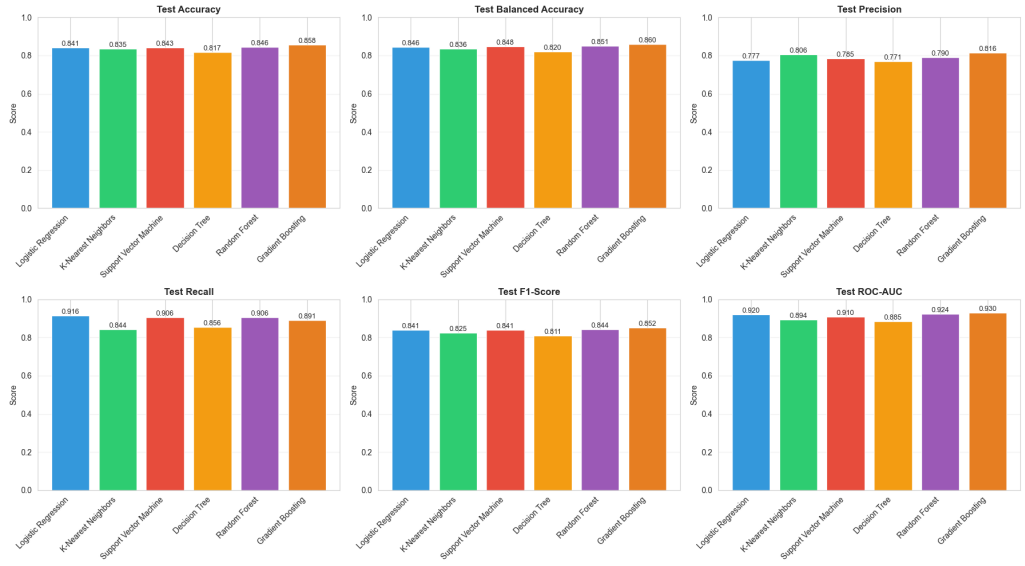


Figura 10: Comparison of test-set performance metrics across classification models.

All models demonstrated strong discriminative performance (Figure 10). Gradient Boosting achieved the highest Balanced Accuracy (0.860) and ROC-AUC (0.930), closely followed by Random Forest (ROC-AUC 0.924). Logistic Regression and SVM achieved the highest Recall values (0.916 and 0.908), indicating superior sensitivity for ischemic cases. Error analysis reveals model-specific trade-offs: Logistic Regression produced the fewest false negatives (34), aligning with its high Recall, whereas KNN showed substantially higher false-negative counts (63). Ensemble models achieved a more balanced error distribution.

3.3.2 ROC-AUC Analysis

The ROC curves (Figure 11) confirm the superior discriminative ability of Gradient Boosting (AUC = 0.930), followed by Random Forest (0.924). Li-

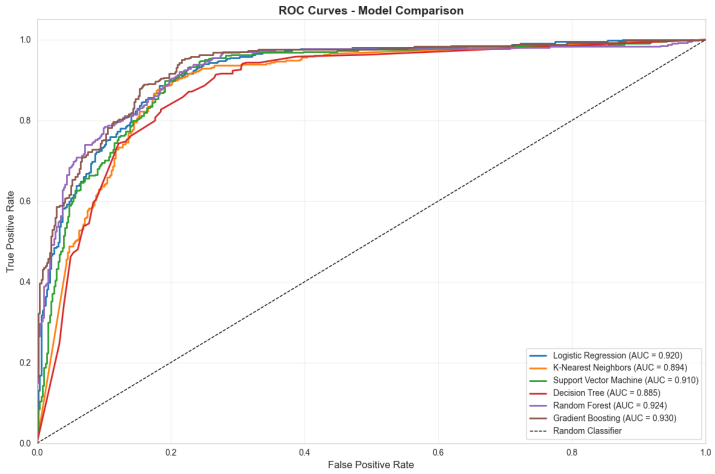


Figura 11: ROC curves for all classification models.

near models also performed competitively, while KNN exhibited the lowest AUC (0.894). Cross-validation results demonstrate that Random Forest and Gradient Boosting achieve the most stable performance, with tightly clustered ROC-AUC and Balanced Accuracy distributions. Linear models also exhibit consistent behavior, whereas KNN shows higher variance.

3.4 Feature Importance Analysis

3.4.1 Tree-Based Models

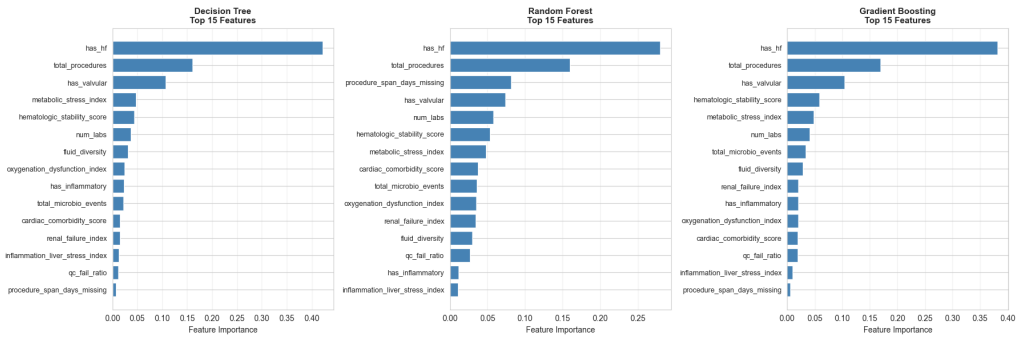


Figura 12: Top feature importances for tree-based models.

Across Decision Tree, Random Forest, and Gradient Boosting models (Figure 12), *has_hf* consistently emerges as the dominant predictor. Additional influential features include *total_procedures* and *has_valvular*, underscoring the importance of cardiac comorbidity burden and intervention intensity. Logistic Regression coefficients indicate that *has_hf* strongly reduces the odds of ischemic classification, while *total_procedures* substantially increases it. These effects align with the ensemble-based importance rankings, reinforcing the clinical plausibility of the learned decision boundaries.

4 Time Series Preprocessing

In this chapter, we outline the preprocessing pipeline applied to the ECG time series data extracted from the cardiovascular cohort. Lead II channel was selected as the primary signal for analysis due to its clinical utility in rhythm assessment.

4.1 Data Overview

The ECG dataset comprises 1,786 patients with complete Lead II recordings. Each recording contains 5,000 samples collected over a 10-second window at a sampling frequency of 500 Hz, resulting in a total of 8.93 million signal samples. The raw ECG signals exhibit typical characteristics of clinical recordings: baseline wander, powerline interference, and amplitude variations across patients.

Property	**Value**
Total Patients	1,786
ECG Channel	Lead II
Sampling Frequency	500 Hz
Signal Duration	10 seconds
Samples per Patient	5,000
Total Samples	8,930,000
Mean Signal Amplitude	0.01 mV
Signal Std Deviation	0.16 mV
Signal Range	-1.53 to 2.27 mV

Tabella 3: ECG time series dataset characteristics.

4.2 Preprocessing Pipeline

To ensure comparability across patients and remove artifacts that could confound downstream analysis, we applied a five-step preprocessing pipeline. The pipeline addresses common challenges in ECG signal processing: baseline drift, amplitude normalization, linear trends, and noise contamination.

The preprocessing sequence consists of: (1) offset translation removal (centering via mean subtraction), (2) amplitude scaling (z-normalization to unit variance), (3) linear trend removal (polynomial detrending), (4) ECG bandpass filtering (0.5–40 Hz), and (5) notch filtering at 60 Hz to eliminate powerline interference.

Figure 13 illustrates the sequential transformation of a representative ECG signal through each preprocessing stage. The final preprocessed signal exhibits a stable baseline, reduced noise, and enhanced visibility of cardiac waveform components (P, QRS, T complexes) compared to the raw recording.

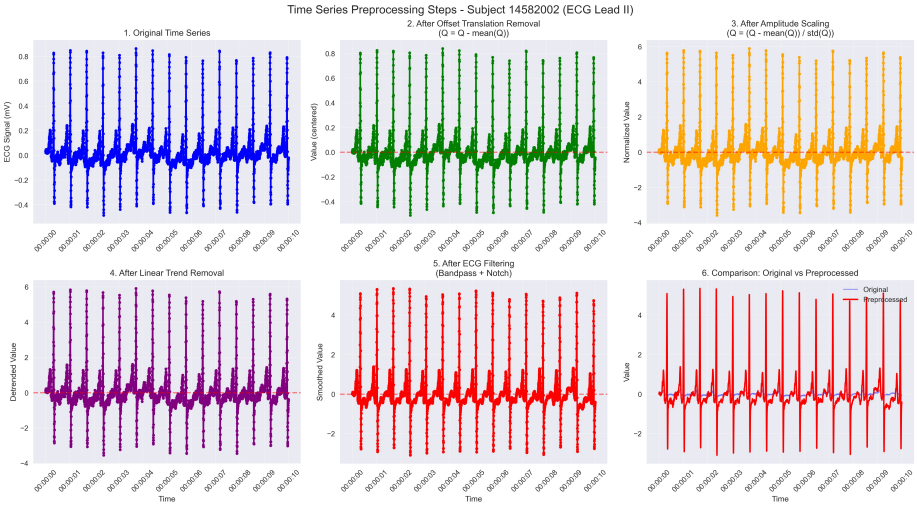


Figura 13: Sequential preprocessing steps applied to an ECG Lead II signal: original signal, offset removal, amplitude scaling, trend removal, ECG filtering (bandpass + notch), and final comparison.

4.3 Dimensionality Reduction with Piecewise Aggregate Approximation

To enable efficient clustering analysis while preserving essential morphological characteristics, we applied Piecewise Aggregate Approximation (PAA). We reduce each 5,000-sample time series to 10 representative segments (500:1 compression) by dividing the signal into equal-length segments and computing mean values. Figure 14 demonstrates the PAA transformation for four representative patients, showing that the step-function approximation captures general morphology and amplitude variations suitable for distance-based clustering algorithms.

4.4 Feature Extraction

In addition to the PAA representation, we extracted 13 statistical features from each preprocessed time series to capture complementary aspects of signal behavior. These features include basic statistics (mean, variance, standard deviation, min, max, range, median), trend characteristics (slope and intercept from linear regression), temporal dependencies (lag-1 autocovariance), and distributional properties (25th and 75th percentiles, interquartile range).

The distributions of nine key features across the cohort, shown in Figure 15, reveal that most signals exhibit near-zero means (reflecting successful centering), standardized variances clustered around unity (confirming effective normalization), and minimal linear trends. The autocovariance values are consistently high (mean=0.87), indicating strong temporal correlation characteristic of ECG signals.

The preprocessed time series data, along with the PAA approximations and extracted features, serve as the foundation for the clustering analysis presented in Section 5 and the time series classification tasks in Section 6.

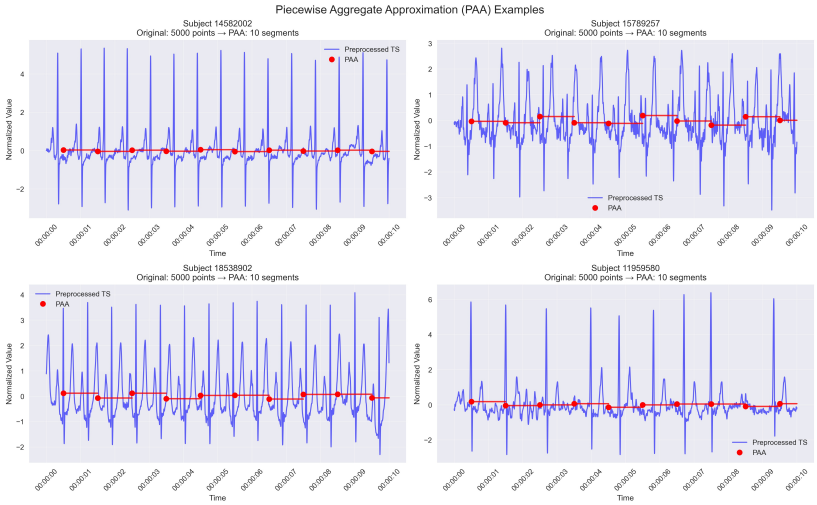


Figura 14: Piecewise Aggregate Approximation (PAA) applied to four representative ECG time series. Each signal is compressed from 5,000 samples to 10 segments (500:1 compression ratio) while preserving the overall signal morphology.

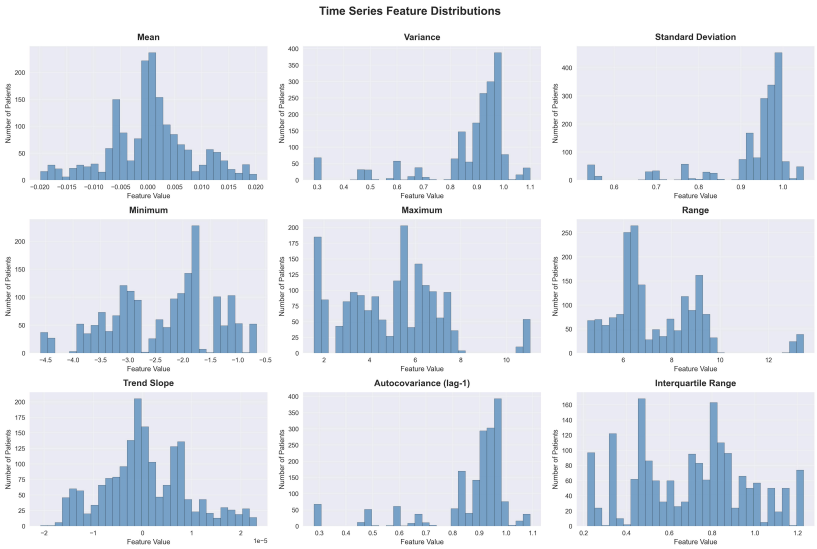


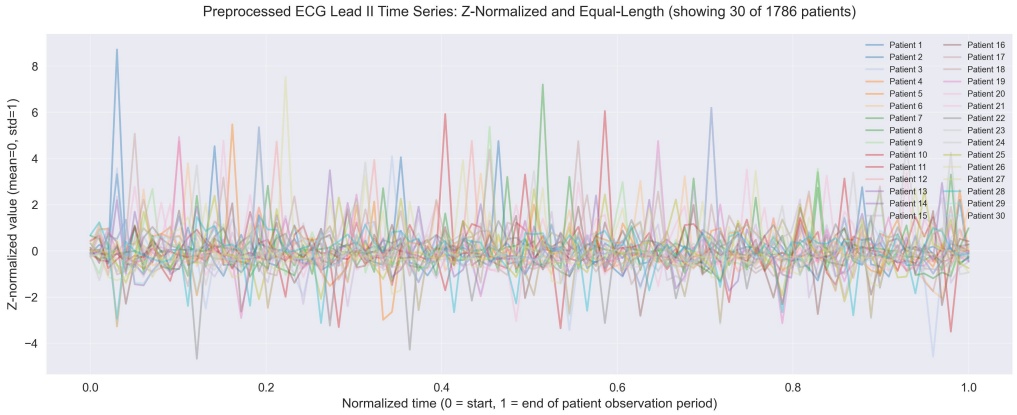
Figura 15: Distribution of nine key statistical features extracted from preprocessed ECG time series across 1,786 patients. Each subplot shows the distribution of feature values (e.g., mean, variance, range) computed from individual patient ECG signals.

5 Time Series Clustering

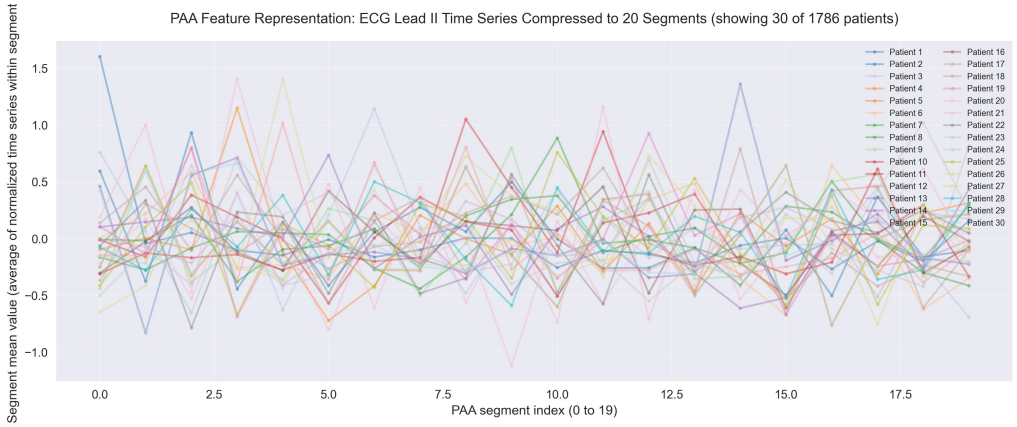
We apply clustering algorithms to preprocessed ECG Lead II time series data using Piecewise Aggregate Approximation (PAA) features. We focus on morphological patterns in ECG signals, revealing continuous variation rather than discrete clinical phenotypes. We compare three clustering algorithms: KMeans, Hierarchical Clustering, and Density-Based Clustering (DBSCAN).

5.1 Feature Representation

Each time series (100 normalized points) is compressed to 20 PAA segments (5:1 compression), preserving morphology. The feature matrix is standardized using *sklearn's StandardScaler*. Figure 16 illustrates the transformation.



(a) Preprocessed ECG time series



(b) PAA feature representation

Figura 16: Feature representation: (a) 30 preprocessed ECG Lead II time series (z-normalized, 100 time points). (b) Corresponding PAA feature vectors (20 segments).

5.2 KMeans Clustering

The Elbow Method (Figure 17) shows gradual decrease in inertia without pronounced elbow, suggesting continuous variation. We selected $k = 5$ based on diminishing returns. KMeans identified five clusters (sizes spanning 68-735 patients). Figure 18 shows cluster-average PAA profiles: some exhibit flat profiles (stable baselines), others show oscillations or trends, likely reflecting rhythm variations. Table 4 shows Silhouette Score of 0.210, substantially lower than tabular clustering (0.941), indicating continuous processes rather than discrete clinical states.

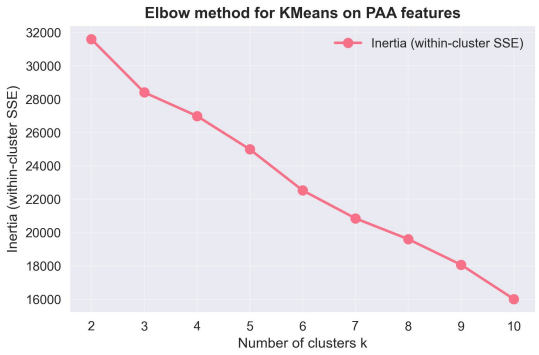


Figura 17: Elbow method for KMeans: inertia vs. k. Gradual decrease without pronounced elbow suggests continuous variation in ECG patterns.

Metric	Value
Number of Clusters (k)	5
Silhouette Score	0.210
Cluster Sizes	131, 735, 292, 560, 68

Tabella 4: KMeans clustering evaluation metrics for ECG time series.

Figura 18: KMeans cluster-average PAA profiles ($k = 5$), revealing distinct temporal patterns in ECG morphology.

5.3 Hierarchical Clustering

Hierarchical clustering with Ward linkage (Figure 19) reveals multi-scale structure, contrasting with tabular clustering's binary split. Extracting five clusters (Figure 20) yields similar patterns to KMeans with different assignments. Silhouette Score of 0.231 (slightly higher than KMeans) reflects Ward linkage's ability to form compact groups, though substantial overlap remains.

5.4 Density-Based Clustering (DBSCAN)

DBSCAN identifies noise points without requiring a pre-specified number of clusters. Parameter exploration (see Appendix A) led to *eps*=0.8 and *min_samples*=10, producing 27 clusters (sizes spanning 39-75 patients each) and 168 noise points (9.4%). Figure 21 shows diverse PAA profiles, supporting the continuum hypothesis with DBSCAN identifying local density peaks. The noise points represent patients with idiosyncratic patterns, potentially rare arrhythmias or unique clinical presentations.

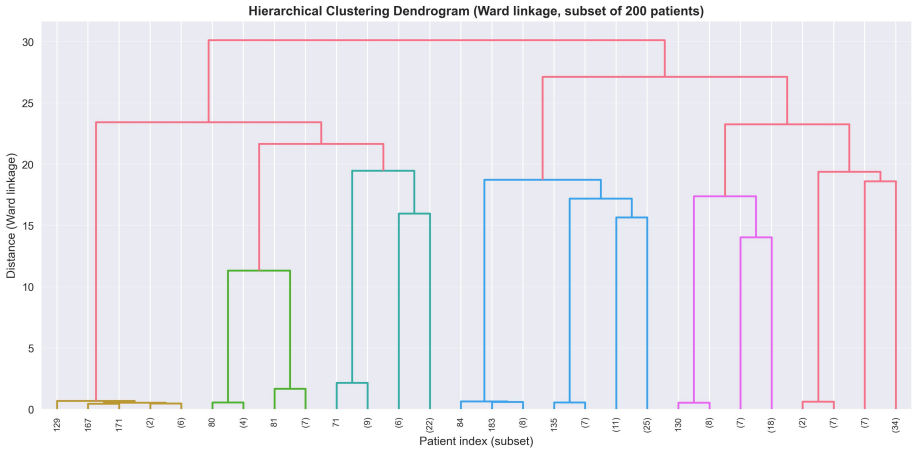


Figura 19: Hierarchical clustering dendrogram (Ward linkage, 200 patients). Multiple levels of structure suggest continuous variation rather than discrete categories.

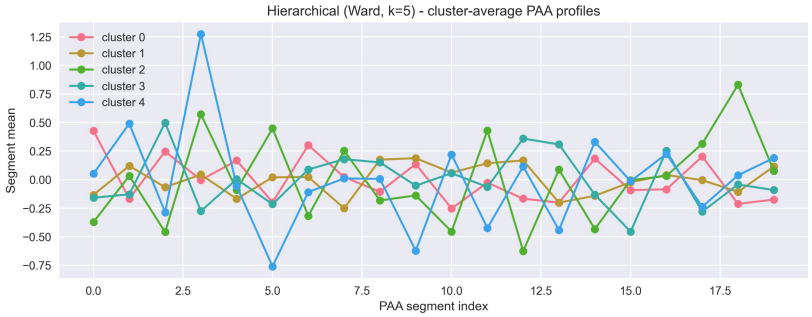


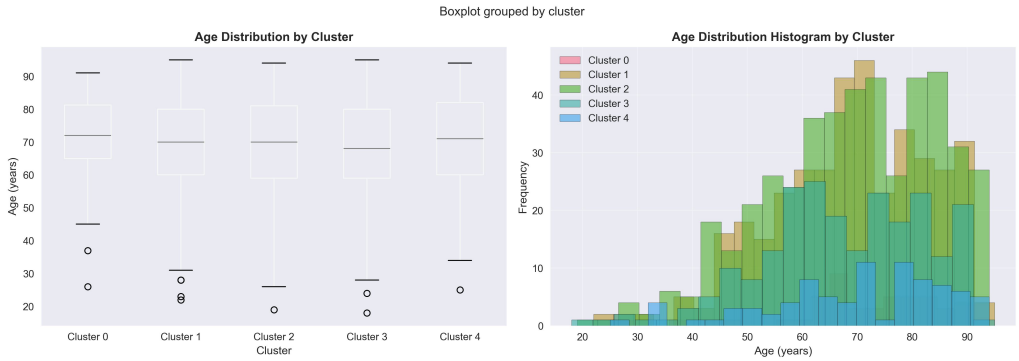
Figura 20: Hierarchical clustering PAA profiles ($k = 5$, Ward linkage), showing similar patterns to KMeans with different assignments.

Figura 21: DBSCAN cluster-average PAA profiles (eps=0.8, min_samples=10). 27 clusters reflect high diversity of ECG patterns. We observe that the yellow cluster consistently has the more extreme values.

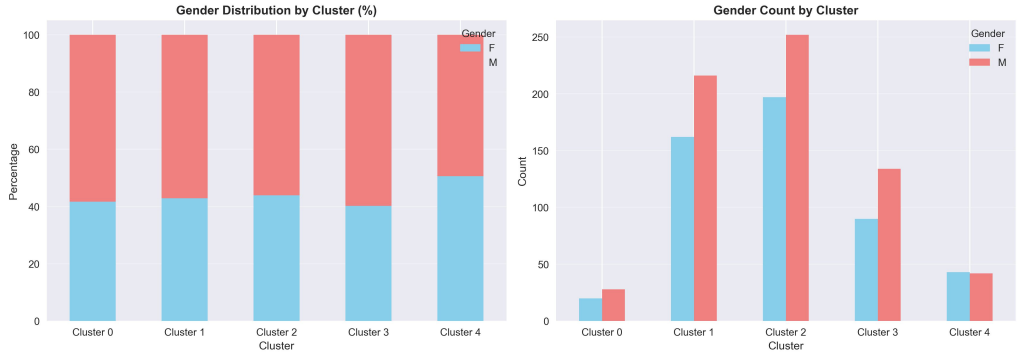
5.5 Clinical Interpretation

Demographic analysis (Figure 22) shows similar age distributions (mean 68.4–70.8 years) and balanced gender proportions. Diagnostic patterns (Figure 23) show AMI, Heart Failure, and Atrial Fibrillation present in all clusters with similar frequencies, consistent with low silhouette scores indicating substantial

overlap.



(a) Age distribution by cluster



(b) Gender distribution by cluster

Figura 22: Demographic characteristics: (a) Similar age distributions across clusters. (b) Balanced gender proportions, indicating ECG patterns not driven by demographics.

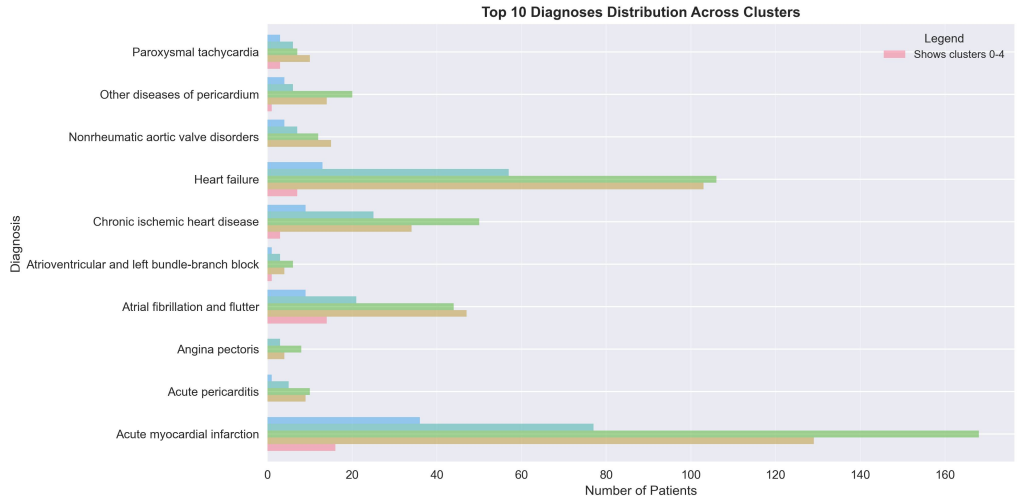


Figura 23: Top 10 diagnoses distribution. AMI, Heart Failure, and Atrial Fibrillation appear in all clusters with similar frequencies, suggesting limited discriminative power.

5.6 Evaluation and Comparison

All three methods produce low silhouette scores (0.203–0.231, Table 5), substantially lower than tabular clustering (0.941), reflecting fundamental differences: tabular features capture discrete clinical states, while ECG time series represent continuous physiological processes.

Method	Clusters	Silhouette Score
KMeans	5	0.210
Hierarchical (Ward)	5	0.231
DBSCAN	27 + 168 noise	0.203

Tabella 5: Comparison of clustering methods applied to ECG time series PAA features. DBSCAN silhouette score calculated on non-noise points only.

The PAA representation may not capture fine-grained temporal patterns needed to distinguish subtle ECG morphologies. DBSCAN's 27 clusters support the continuum hypothesis: ECG patterns form a continuous distribution with local density peaks rather than discrete categories. The lack of correlation between ECG patterns and demographics/diagnoses suggests Lead II signals at PAA resolution reflect general cardiac activity rather than specific disease states. Clinical ECG interpretation relies on waveform components (P waves, QRS complexes, ST segments) rather than overall shape, explaining limited discriminative power.

6 Time Series Classification

We address the binary classification task of distinguishing ischemic from non-ischemic cardiac patients using preprocessed ECG Lead II time series data. The dataset consists of 1,184 patients with balanced classes (609 non-ischemic, 575 ischemic) and diagnostic labels derived from ICD codes. Classification operates on the same 5,000-sample preprocessed time series described in Section 3.

6.1 Feature Extraction

We extract fixed-length features from the time series using four methods: **PAA** (30 segments), **SAX** (30 symbols, 4-symbol alphabet), **DFT** (30 coefficients), and **HRV** (6 clinical metrics: `mean_rr`, `std_rr`, `rmssd`, `pnn50`, `hr_mean`, `lf_hf_ratio`). The complete feature set comprises 96 features, standardized using *sklearn's StandardScaler*.

6.2 Classification Models

We evaluate six classification approaches: **KNN with DTW** (time series native, $k = 5$, downsampled to 20 points), **Logistic Regression** (linear baseline), **XGBoost** and **Random Forest** (ensemble methods), **Shapelet classifier** (10 shapelets, Decision Tree), and **SVM** (RBF kernel). All models employ class balancing strategies.

6.3 Results and Evaluation

Table 6 summarizes model performance. The Shapelet classifier achieves the highest F1-score (0.5758) and recall (0.6609), though overall performance is modest (best accuracy = 52.74%, only slightly above random chance).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
KNN (DTW)	0.4833	0.5357	0.4545	0.4918	—
Logistic Regression	0.4599	0.4425	0.4348	0.4386	0.4750
XGBoost	0.5021	0.4878	0.5217	0.5042	0.5055
Shapelet	**0.5274**	**0.5101**	**0.6609**	**0.5758**	0.5180
SVM	0.5232	0.5078	0.5652	0.5350	0.5217
Random Forest	0.5021	0.4878	0.5217	0.5042	0.5143

Tabella 6: Classification performance metrics. The Shapelet classifier achieves the highest F1-score and recall, indicating superior sensitivity for detecting ischemic patients.

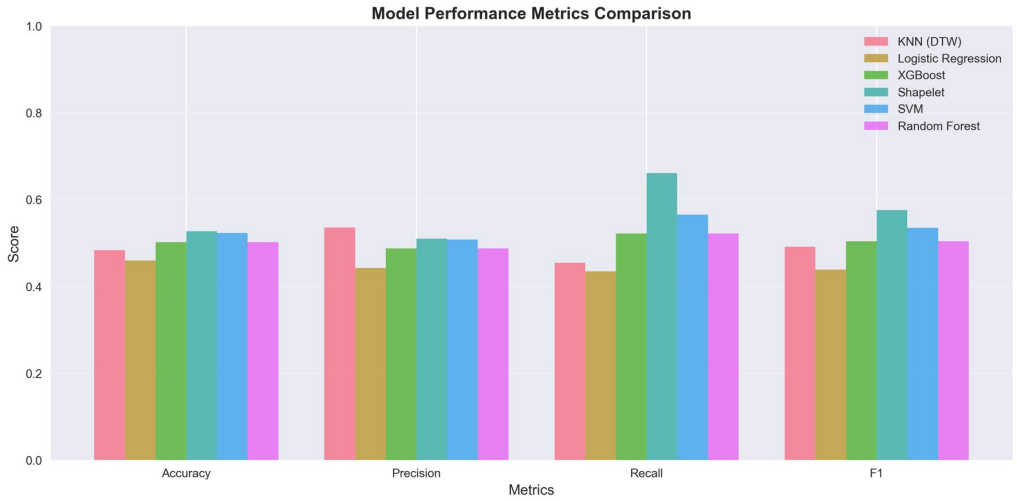


Figura 24: Model performance comparison across all metrics. Shapelet, SVM, and ensemble methods outperform the linear baseline and KNN with DTW.

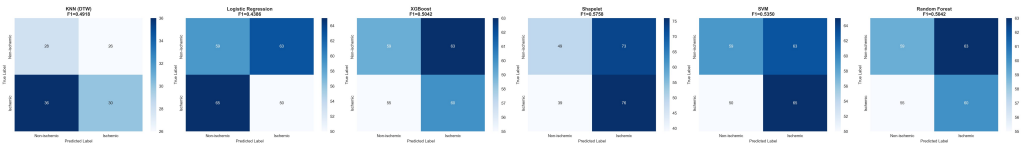
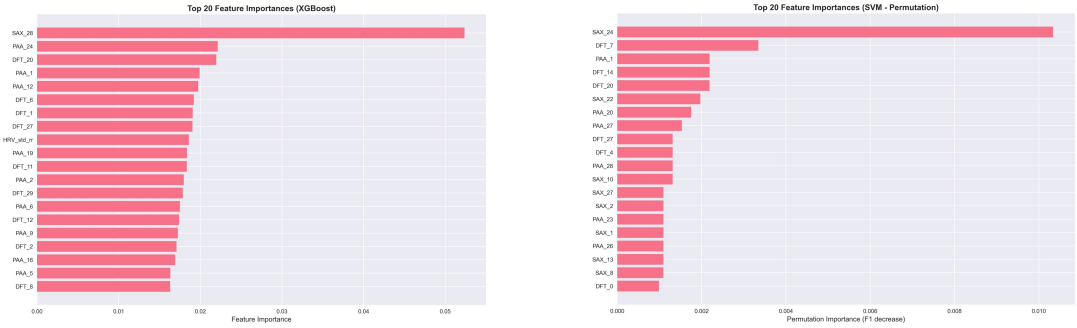


Figura 25: Confusion matrices for all models. The Shapelet classifier shows the highest true positive rate (76) but also the highest false positive rate (73), consistent with its high recall and moderate precision.

6.3.1 Feature Importance Analysis

SAX features, particularly `SAX_28`, dominate XGBoost importance rankings, followed by PAA and DFT coefficients. HRV features (`HRV_std_rr`) also



(a) XGBoost feature importance (b) SVM permutation importance

Figura 26: Feature importance analysis. SAX features, particularly `SAX_28`, dominate XGBoost rankings, while SVM permutation importance shows contributions from multiple feature types.

appear among top contributors. SVM permutation importance shows a more distributed pattern across all feature types.

6.3.2 Shapelet Analysis

The superior performance of the Shapelet classifier suggests that local pattern matching captures discriminative temporal structures more effectively than global approximation-based features. Detailed shapelet analysis (see Appendix A) reveals relatively uniform feature importance contributions, with Shapelet 3 (length 22) and Shapelet 7 (length 31) having highest importance, likely corresponding to ECG waveform components (QRS complexes, ST segments) altered in ischemic conditions.

6.4 Discussion

The modest classification performance (best F1-score = 0.5758) indicates fundamental challenges in ECG-based ischemic detection using the employed feature representations. The near-random performance across most models suggests that global approximation-based features (PAA, SAX, DFT) may not capture the subtle morphological changes associated with ischemic heart disease. Clinical ECG interpretation relies on specific waveform components (ST-segment elevation/depression, T-wave inversion, Q-wave presence) that may be obscured in segment-level approximations. The relative success of the Shapelet classifier supports this hypothesis, as it captures local patterns that may correspond to clinically relevant features. Key limitations include preprocessing potentially removing discriminative high-frequency components, coarse temporal resolutions (30 segments for 5,000-sample signals), binary classification aggregating diverse ischemic conditions, and Lead II signals alone potentially missing full spatial information needed for comprehensive ischemic detection.