Chapter 1

Time Series Preprocessing

In this chapter, we outline the preprocessing pipeline applied to the ECG time series data extracted from the cardiovascular cohort. The preprocessing steps are essential for preparing the raw physiological signals for downstream clustering and classification tasks. All ECG recordings were obtained from WFDB-formatted 12-lead ECG data, with the Lead II channel selected as the primary signal for analysis due to its clinical utility in rhythm assessment and its consistent visibility of cardiac electrical activity.

1.1 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

Table 1.1: ECG time series dataset characteristics.

1.2 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*, which centers each signal around zero by subtracting the mean value; (2) *amplitude scaling* via z-normalization, standardizing the variance to unity; (3) *linear trend removal* using polynomial detrending to eliminate slow baseline drift; (4) *ECG bandpass filtering* (0.5–40 Hz) to remove baseline wander

and high-frequency noise while preserving the clinically relevant frequency components of the ECG signal; and (5) *notch filtering* at 60 Hz to eliminate powerline interference common in hospital environments.

Figure 1.1 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.

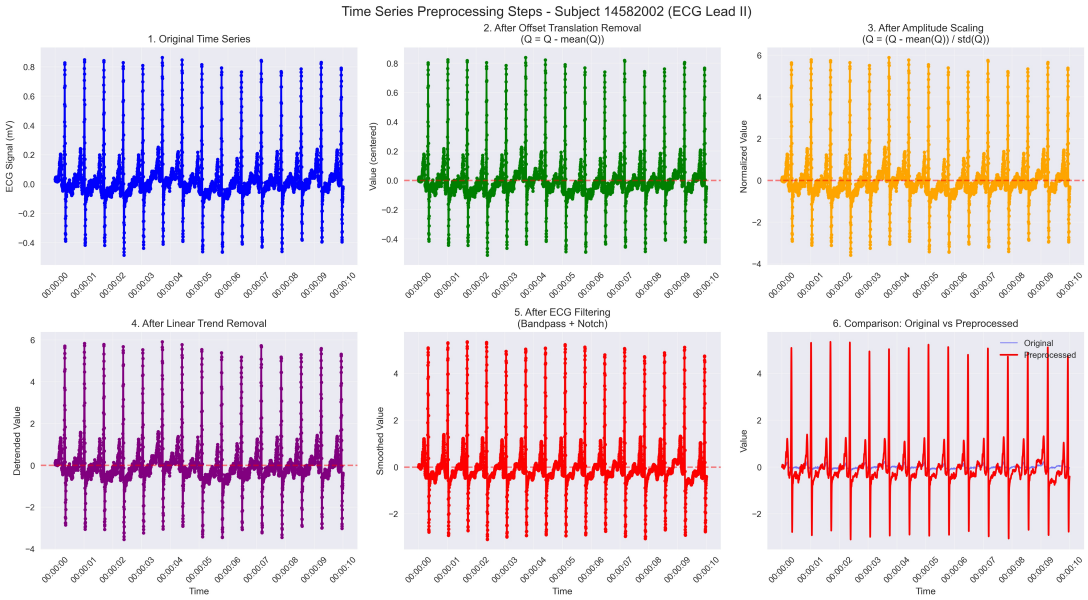


Figure 1.1: Sequential preprocessing steps applied to an ECG Lead II signal (Subject 14582002). From top left to bottom right: original signal, offset removal, amplitude scaling, trend removal, ECG filtering (bandpass + notch), and final comparison.

1.3 Dimensionality Reduction with Piecewise Aggregate Approximation

To enable efficient clustering analysis while preserving the essential morphological characteristics of the ECG signals, we applied Piecewise Aggregate Approximation (PAA). PAA reduces the dimensionality of each 5,000-sample time series by dividing the signal into equal-length segments and computing the mean value within each segment. This approach achieves a compression ratio of 500:1, reducing each signal to 10 representative segments while maintaining the overall trend and amplitude characteristics.

Figure 1.2 demonstrates the PAA transformation for four representative patients. The step-function approximation captures the general morphology and amplitude variations of the original preprocessed signals, making it suitable for distance-based clustering algorithms that require fixed-length feature vectors.

1.4 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).

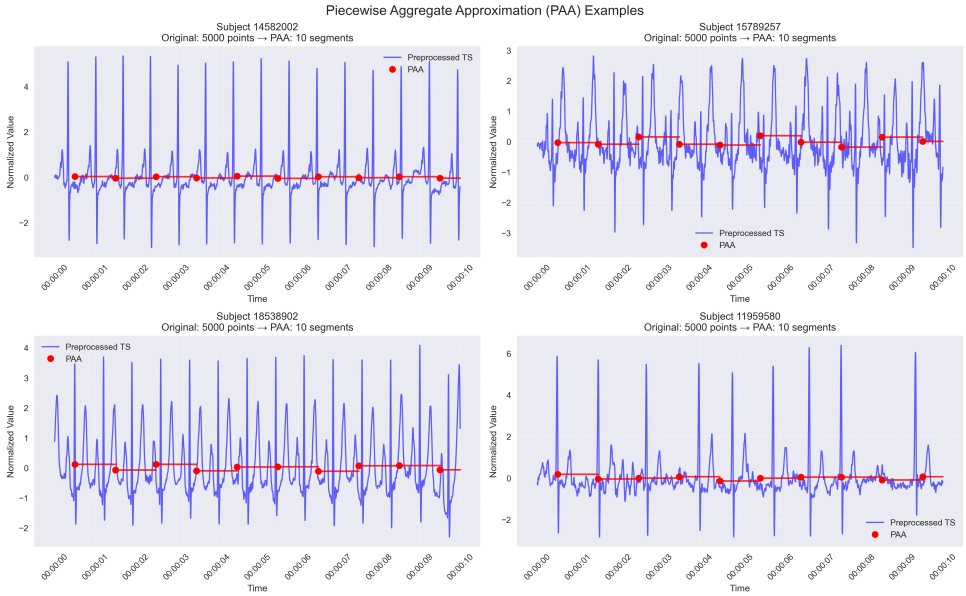


Figure 1.2: Piecewise Aggregate Approximation (PAA) applied to four representative ECG time series. Each signal is compressed from 5,000 samples to 10 segments (500:1 compression ratio) while preserving the overall signal morphology.

The distributions of nine key features across the cohort, shown in Figure 1.3, 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 Chapter 5 and the time series classification tasks in Chapter 6.

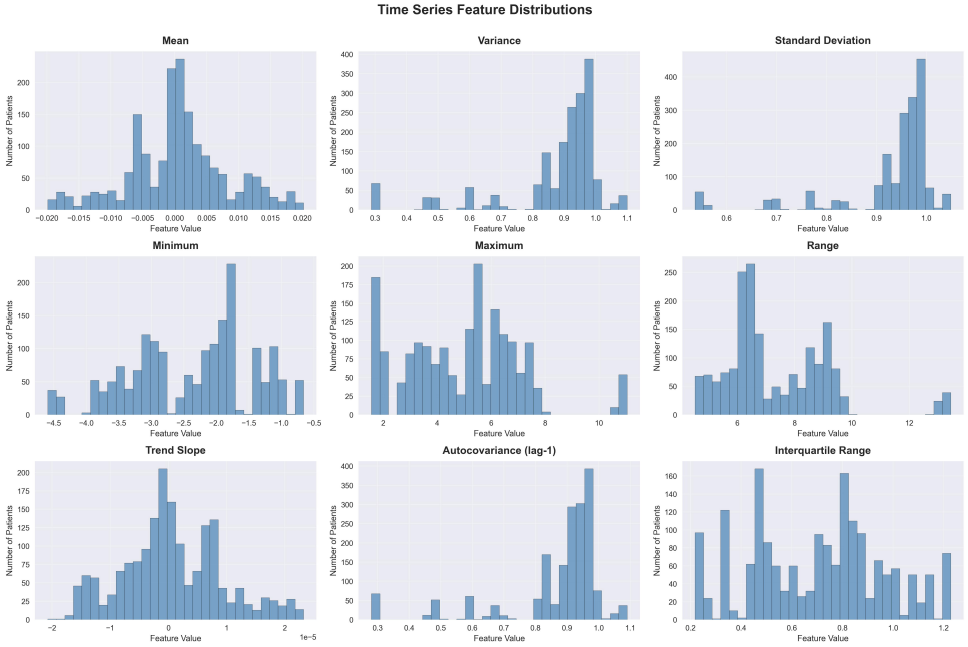


Figure 1.3: Distribution of nine key statistical features extracted from preprocessed ECG time series across 1,786 patients. Each subplot shows the distribution of feature values (e.g., mean, variance, range) computed from individual patient ECG signals.

Chapter 2

Time Series Clustering

We apply clustering algorithms to preprocessed ECG Lead II time series data using Piecewise Aggregate Approximation (PAA) features. Unlike tabular clustering on clinical features, this analysis focuses on morphological patterns in ECG signals, revealing continuous variation rather than discrete clinical phenotypes.

2.1 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 2.1 illustrates the transformation.

2.2 KMeans Clustering

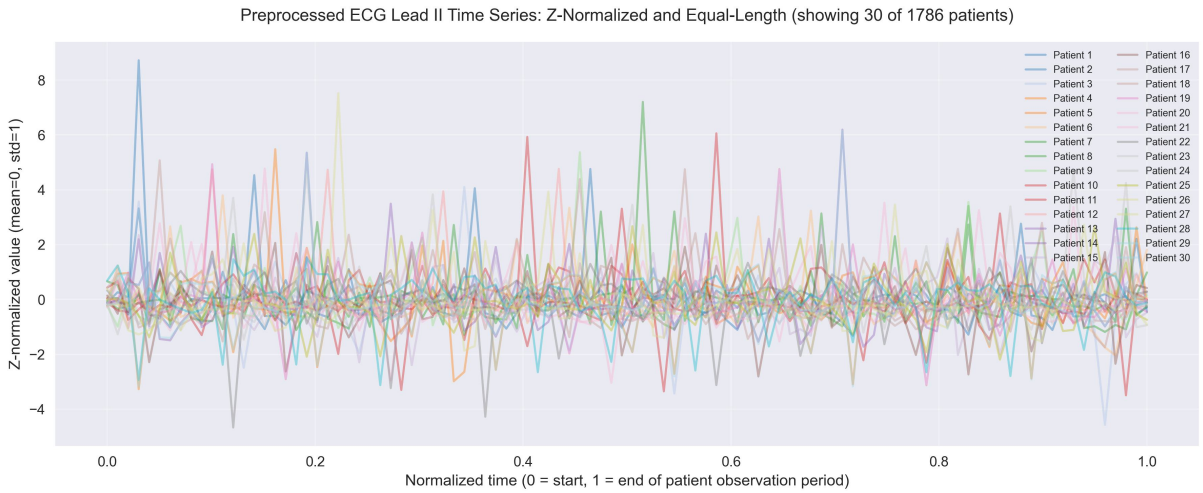
The Elbow Method (Figure 2.2) shows gradual decrease in inertia without pronounced elbow, suggesting continuous variation. We selected $k = 5$ based on diminishing returns. KMeans identified five clusters (68–735 patients). Figure 2.3 shows cluster-average PAA profiles: some exhibit flat profiles (stable baselines), others show oscillations or trends, likely reflecting rhythm variations. Table 2.1 shows Silhouette Score of 0.210, substantially lower than tabular clustering (0.941), indicating continuous processes rather than discrete clinical states.

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

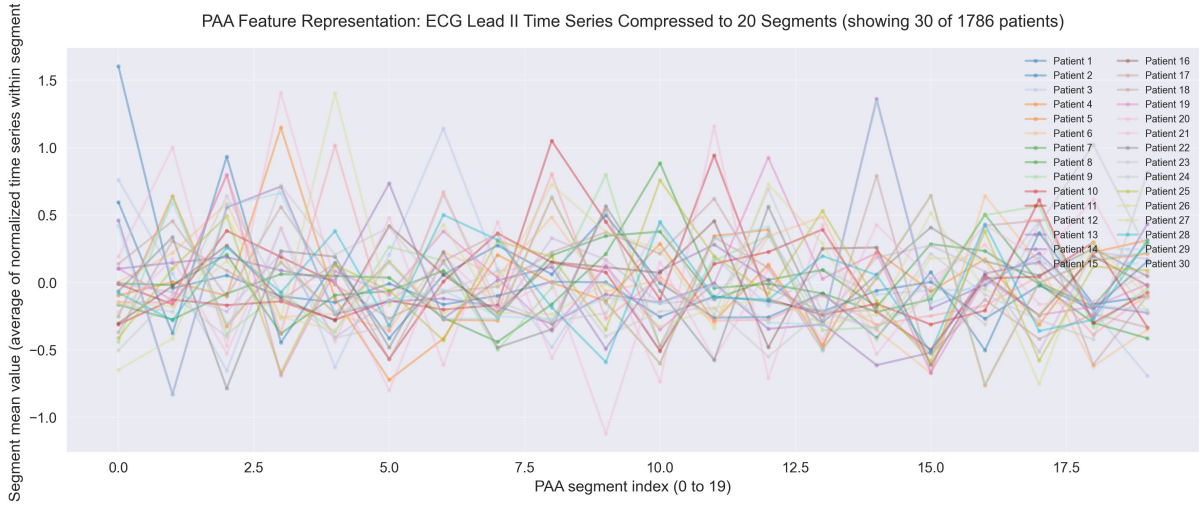
Table 2.1: KMeans clustering evaluation metrics for ECG time series.

2.3 Hierarchical Clustering

Hierarchical clustering with Ward linkage (Figure 2.4) reveals multi-scale structure, contrasting with tabular clustering's binary split. Extracting five clusters (Figure 2.5) 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.



(a) Preprocessed ECG time series



(b) PAA feature representation

Figure 2.1: Feature representation: (a) 30 preprocessed ECG Lead II time series (z-normalized, 100 time points). (b) Corresponding PAA feature vectors (20 segments).

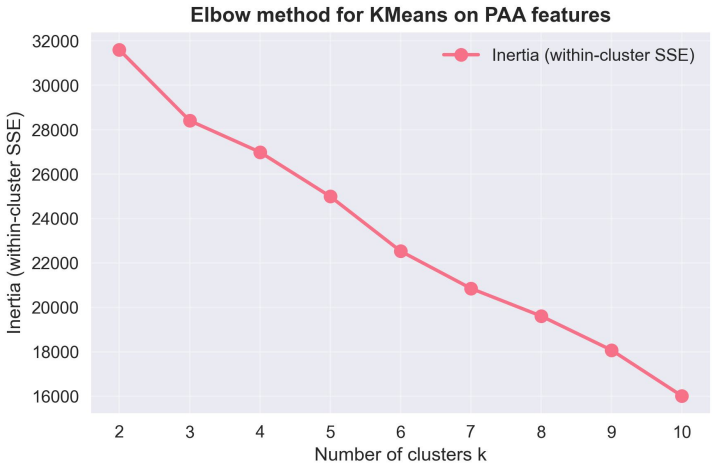


Figure 2.2: Elbow method for KMeans: inertia vs. k. Gradual decrease without pronounced elbow suggests continuous variation in ECG patterns.

Figure 2.3: KMeans cluster-average PAA profiles ($k = 5$), revealing distinct temporal patterns in ECG morphology.

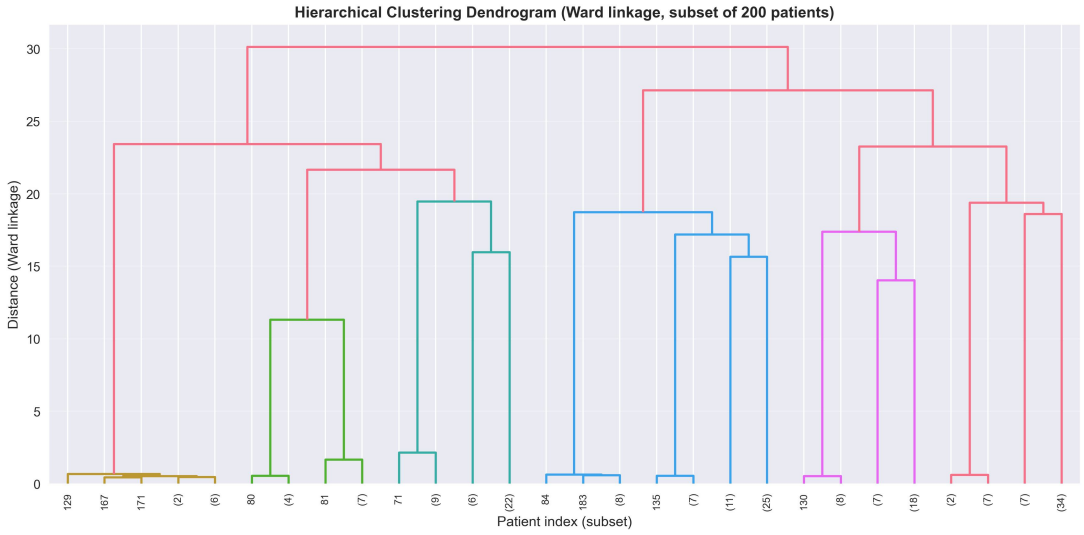


Figure 2.4: Hierarchical clustering dendrogram (Ward linkage, 200 patients). Multiple levels of structure suggest continuous variation rather than discrete categories.

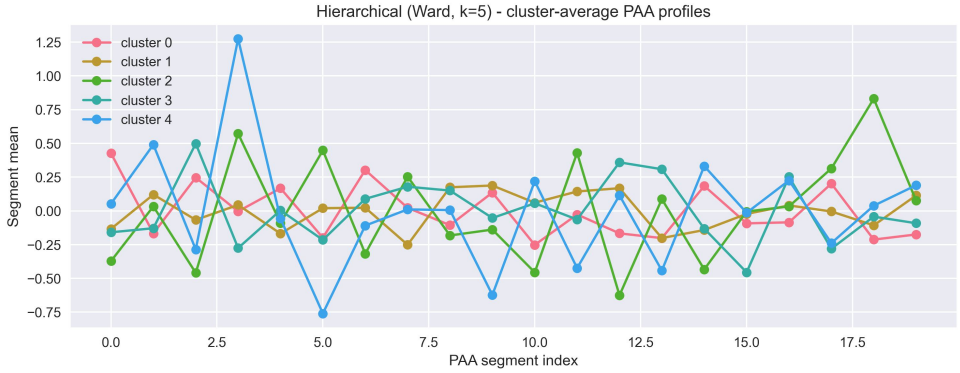


Figure 2.5: Hierarchical clustering PAA profiles ($k = 5$, Ward linkage), showing similar patterns to KMeans with different assignments.

2.4 Density-Based Clustering (DBSCAN)

DBSCAN identifies noise points without requiring a pre-specified number of clusters. Parameter exploration (Table 2.2) led to *eps*=0.8 and *min_samples*=10, producing 27 clusters (39–75 patients each) and 168 noise points (9.4%). Figure 2.6 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.

eps	min_samples	Clusters / Noise
0.5	5	26 / 955
0.5	10	25 / 1067
0.8	5	28 / 147
0.8	10	27 / 168
1.0	10	28 / 68
1.2	10	28 / 35

Table 2.2: DBSCAN parameter exploration results. Lower *eps* values produce many small clusters with high noise rates, while higher values merge clusters but reduce noise detection.

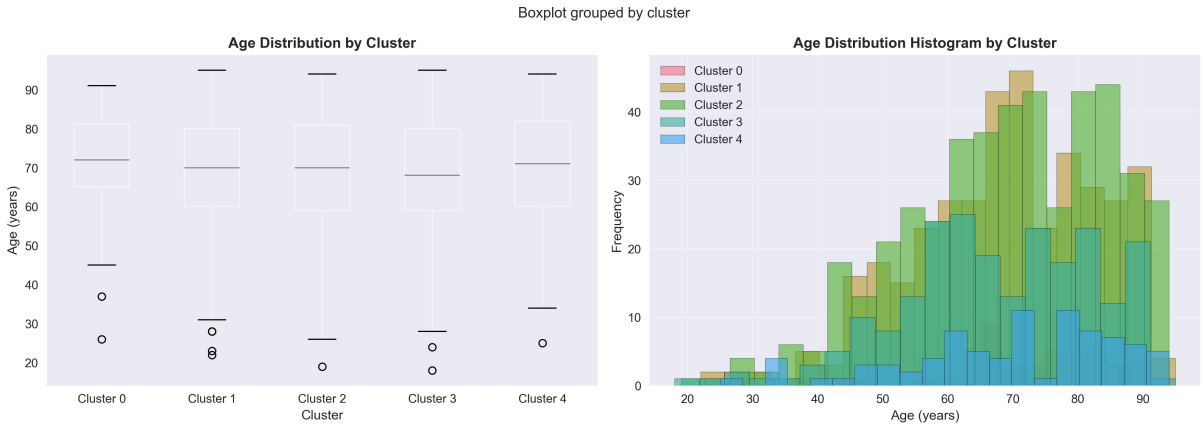
Figure 2.6: DBSCAN cluster-average PAA profiles (eps=0.8, min_samples=10). 27 clusters reflect high diversity of ECG patterns.

2.5 Clinical Interpretation

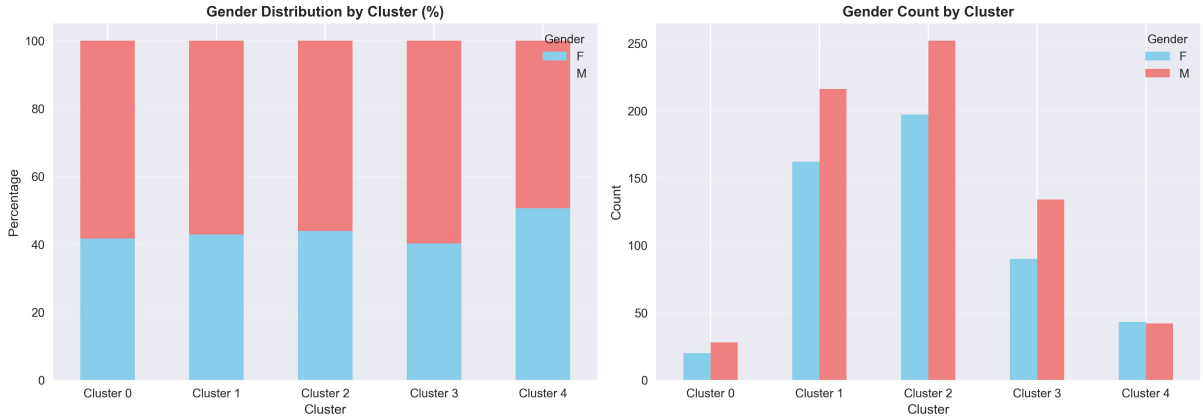
Demographic analysis (Figure 2.7) shows similar age distributions (mean 68.4–70.8 years) and balanced gender proportions. Diagnostic patterns (Figure 2.8, Table 2.3) show AMI, Heart Failure, and Atrial Fibrillation present in all clusters with similar frequencies, consistent with low silhouette scores indicating substantial overlap.

Cluster	Size	Top 1	Top 2	Top 3
0	131	AMI (22.1%)	Heart Failure (17.6%)	Atrial Fibrillation (14.5%)
1	735	AMI (25.4%)	Heart Failure (16.6%)	Atrial Fibrillation (7.1%)
2	292	AMI (19.5%)	Heart Failure (19.2%)	Chronic Ischemic Heart Disease (7.9%)
3	560	AMI (23.8%)	Heart Failure (14.1%)	Chronic Ischemic Heart Disease (7.5%)
4	68	AMI (29.4%)	Heart Failure (8.8%)	Atrial Fibrillation (5.9%)

Table 2.3: Top 3 diagnoses by cluster. AMI is most common; Heart Failure and Atrial Fibrillation show variation in ranks.



(a) Age distribution by cluster



(b) Gender distribution by cluster

Figure 2.7: Demographic characteristics: (a) Similar age distributions across clusters. (b) Balanced gender proportions, indicating ECG patterns not driven by demographics.

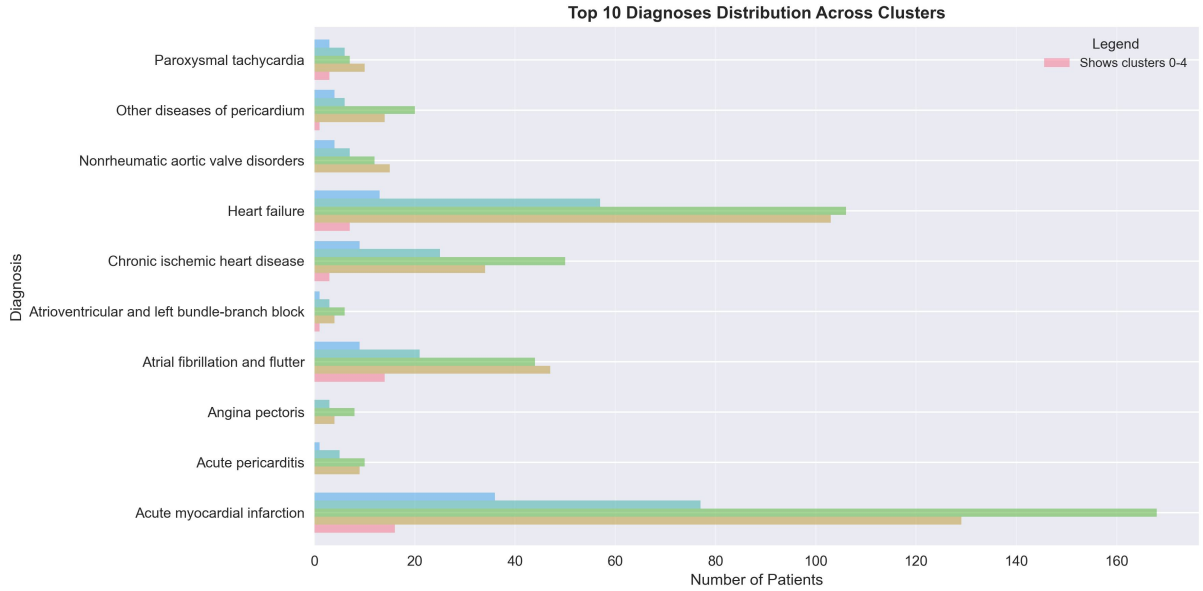


Figure 2.8: Top 10 diagnoses distribution. AMI, Heart Failure, and Atrial Fibrillation appear in all clusters with similar frequencies, suggesting limited discriminative power.

2.6 Evaluation and Comparison

All three methods produce low silhouette scores (0.203–0.231, Table 2.4), 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**	**Key Characteristics**
KMeans	5	0.210	Balanced cluster sizes, interpretable profiles
Hierarchical (Ward)	5	0.231	Compact clusters, hierarchical structure
DBSCAN	27 + 168 noise	0.203	High granularity, noise detection

Table 2.4: Comparison of clustering methods applied to ECG time series PAA features. DBSCAN silhouette score calculated on non-noise points only.

Interpretation: 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.

Clinical Argument: 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. Future work should explore fine-grained morphological feature extraction.

Chapter 3

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 Chapter 3.

3.1 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*.

3.2 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.

3.3 Results and Evaluation

Table 3.1 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

Table 3.1: Classification performance metrics. The Shapelet classifier achieves the highest F1-score and recall, indicating superior sensitivity for detecting ischemic patients.

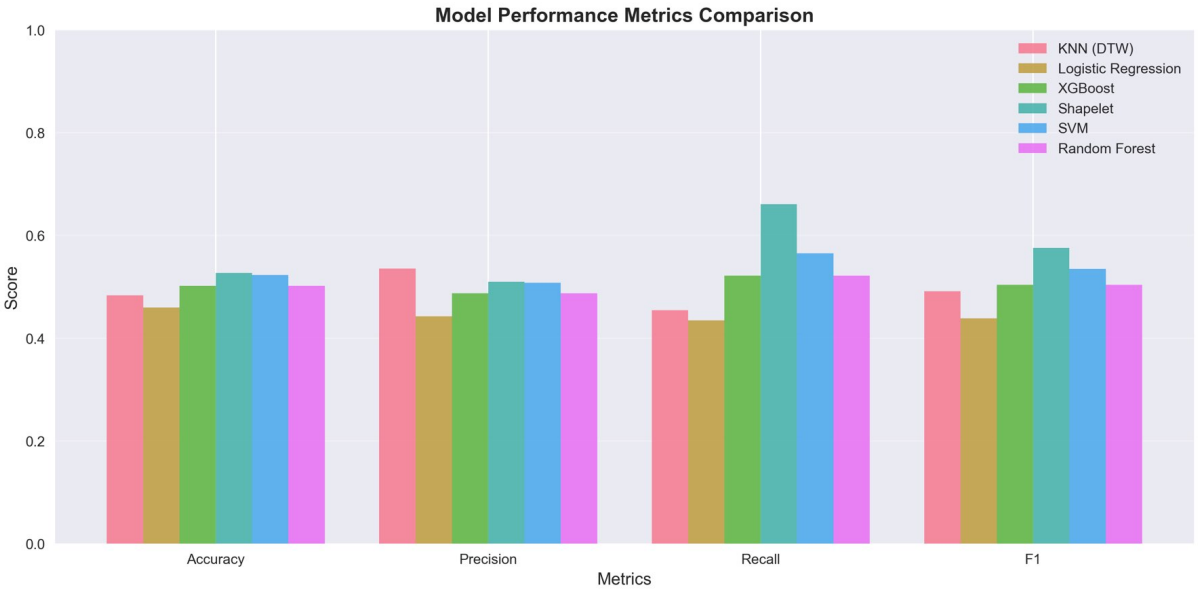


Figure 3.1: Model performance comparison across all metrics. Shapelet, SVM, and ensemble methods outperform the linear baseline and KNN with DTW.

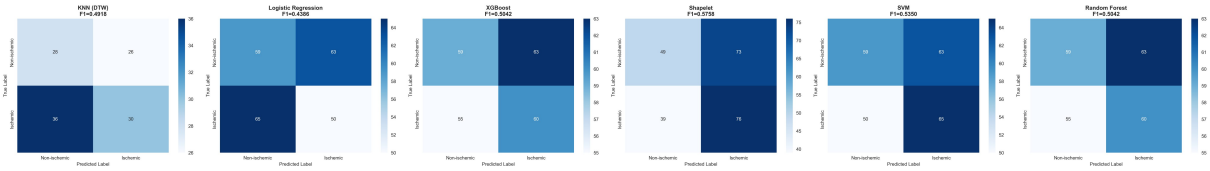


Figure 3.2: Confusion matrices for all models. The Shapelet classifier shows the highest true positive rate (76) but also the highest false positive rate (73), consistent with its high recall and moderate precision.

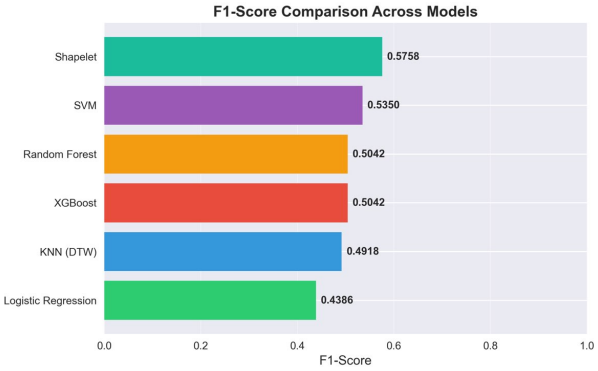
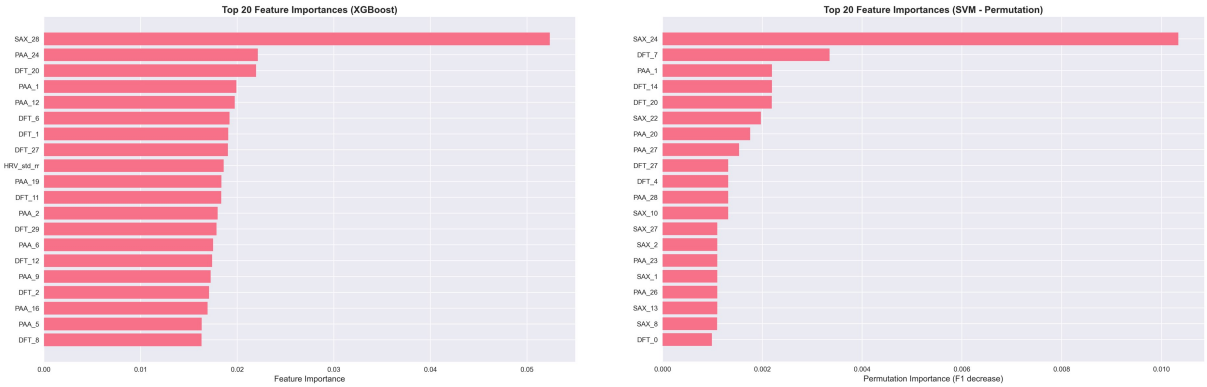


Figure 3.3: F1-score rankings. The Shapelet classifier achieves the highest F1-score (0.5758).

3.3.1 Feature Importance Analysis



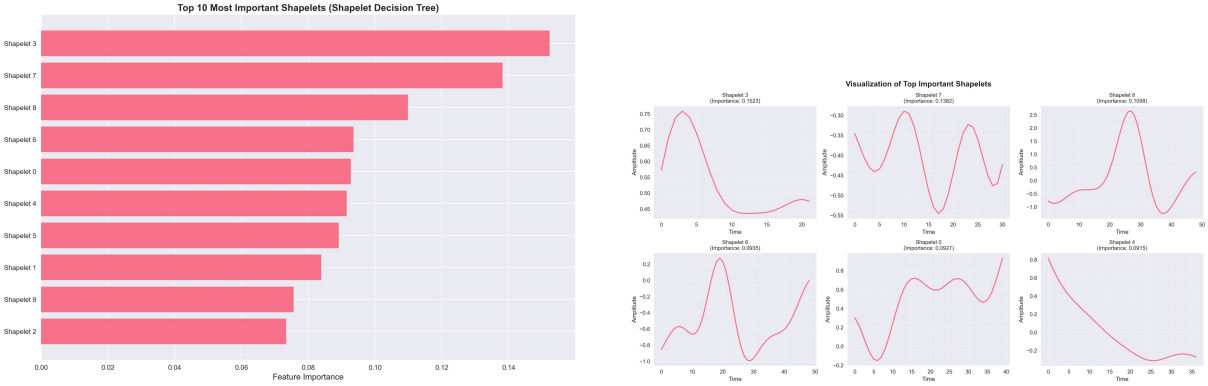
(a) XGBoost feature importance

(b) SVM permutation importance

Figure 3.4: Feature importance analysis. SAX features, particularly `SAX_28`, dominate XGBoost rankings, while SVM permutation importance shows contributions from multiple feature types.

SAX features, particularly `SAX_28`, dominate XGBoost importance rankings, followed by PAA and DFT coefficients. HRV features (`HRV_std_rr`) also appear among top contributors. SVM permutation importance shows a more distributed pattern across all feature types.

3.3.2 Shapelet Analysis



(a) Shapelet feature importance

(b) Top discriminative shapelets

Figure 3.5: Shapelet analysis: (a) Feature importance scores showing relatively uniform contributions, with Shapelet 3 (length 22) and Shapelet 7 (length 31) having highest importance. (b) Visualizations of discriminative patterns, likely corresponding to ECG waveform components (QRS complexes, ST segments) altered in ischemic conditions.

The superior performance of the Shapelet classifier suggests that local pattern matching captures discriminative temporal structures more effectively than global approximation-based features.

3.4 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. However, several limitations likely contribute to the limited performance: (1) preprocessing may remove discriminative high-frequency components or normalize away clinically significant amplitude differences, (2) feature extraction operates at coarse temporal resolutions (30 segments for 5,000-sample signals), potentially missing fine-grained patterns, (3) binary classification aggregates diverse ischemic conditions (acute MI, chronic ischemic heart disease) into a single class, and (4) Lead II signals alone may not capture the full spatial information needed for comprehensive ischemic detection.