Chapter 5 Preview: Time Series Clustering

Preview

January 2, 2026

Contents

List of Figures

List of Tables

Chapter 1

Time Series Clustering

In this chapter, we apply clustering algorithms to the preprocessed ECG Lead II time series data described in Chapter 3. Unlike the tabular patient profile clustering analysis, which operated on 23 numerical features derived from clinical records, this analysis focuses exclusively on the morphological patterns captured in the ECG signals themselves. We employ Piecewise Aggregate Approximation (PAA) features extracted from the preprocessed time series to enable efficient distance-based clustering while preserving the essential temporal structure of cardiac electrical activity.

1.1 Feature Representation

The clustering analysis operates on PAA feature vectors derived from the preprocessed ECG signals. Each time series, originally comprising 100 normalized time points (resampled from the 5,000-sample raw recordings), is compressed to 20 segments using PAA. This dimensionality reduction achieves a 5:1 compression ratio while maintaining the overall signal morphology and amplitude characteristics necessary for pattern recognition.

Prior to clustering, the PAA feature matrix is standardized using *sklearn's StandardScaler* to ensure that all segments contribute equally to distance calculations, regardless of their absolute amplitude values. This preprocessing step is essential given that ECG signals exhibit natural amplitude variations across patients due to differences in electrode placement, body composition, and cardiac output.

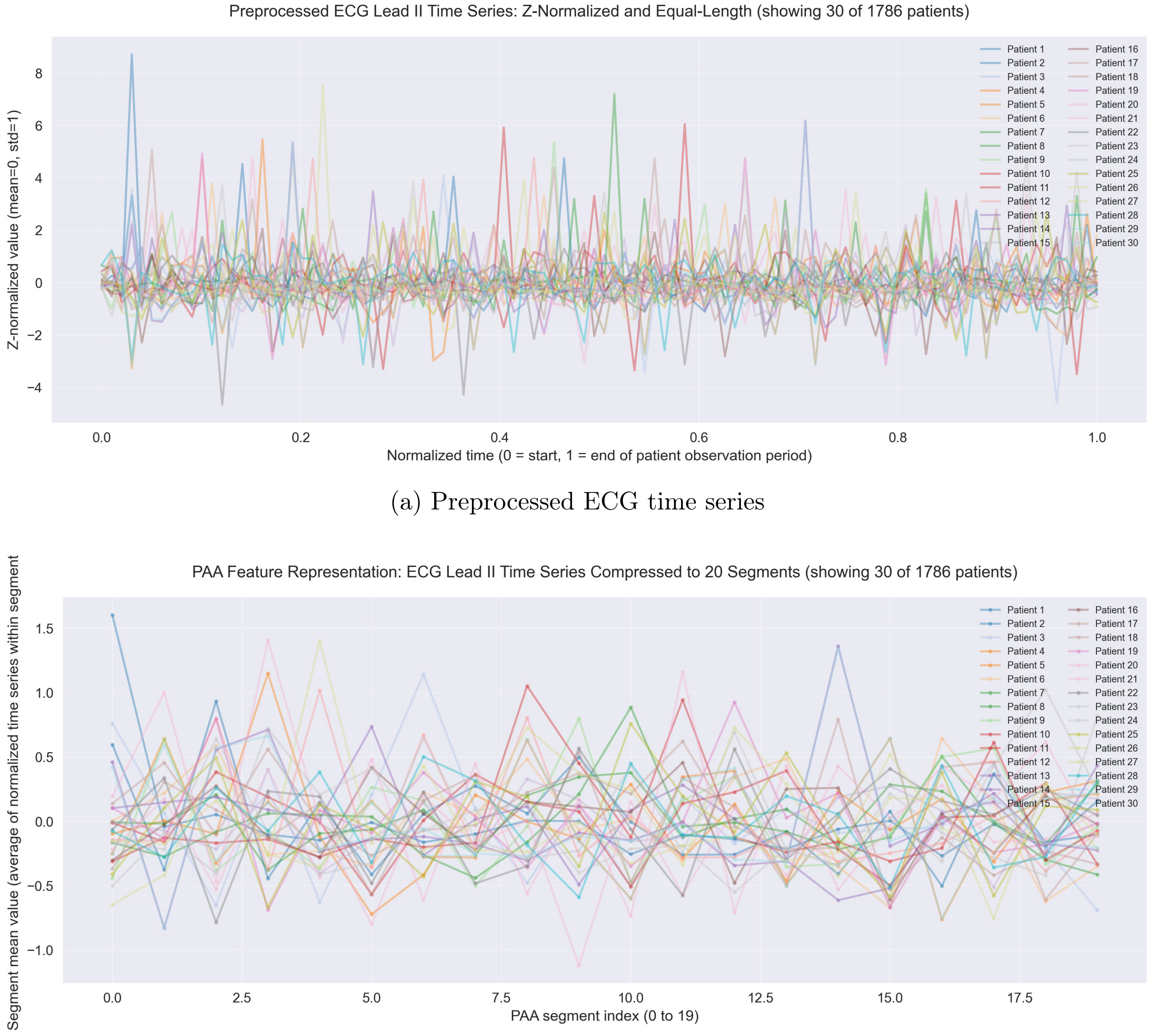
Figure 1.1 illustrates the transformation from preprocessed time series to PAA feature representation. The step-function approximation captures the general morphology and temporal dynamics of the original signals, making it suitable for distance-based clustering algorithms that require fixed-length feature vectors.

1.2 KMeans Clustering

1.2.1 Optimal Cluster Selection (k)

To determine the most appropriate number of clusters for the ECG time series, we employed the Elbow Method, which plots the within-cluster sum of squares (inertia) as a function of k. Figure 1.2 shows the inertia values for $k \in \{2, 3, \ldots, 10\}$.

The elbow plot reveals a gradual decrease in inertia with increasing k, without a pronounced "elbow" that would indicate a clear optimal value. The curve shows diminishing returns after $k = 5$, where the rate of decrease in inertia begins to flatten. This behavior suggests that ECG morphological patterns exhibit more continuous variation compared to the discrete clinical phenotypes identified in the tabular clustering analysis.



(a) Preprocessed ECG time series

(b) PAA feature representation

Figure 1.1: Feature representation for time series clustering. (a) Sample of 30 preprocessed ECG Lead II time series, z-normalized and resampled to 100 time points. (b) Corresponding PAA feature vectors compressed to 20 segments.

Based on the elbow analysis and the need to balance cluster interpretability with granularity, we selected $k = 5$ for the KMeans clustering experiments. This choice allows for the identification of distinct ECG pattern groups while avoiding excessive fragmentation that would reduce clinical utility.

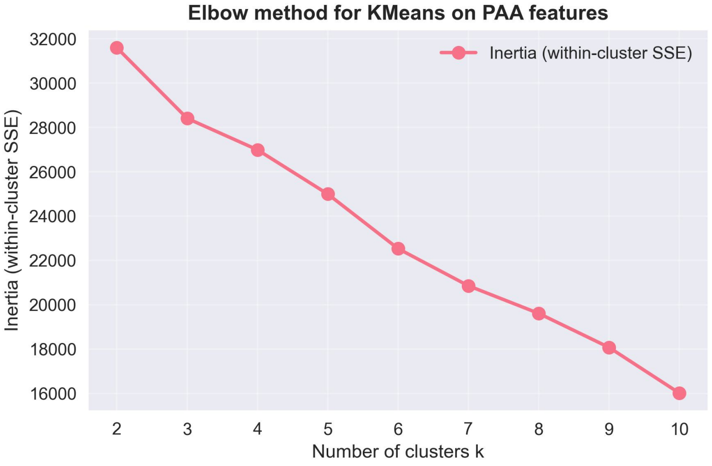


Figure 1.2: Elbow method for KMeans clustering on PAA features. The plot shows within-cluster sum of squares (inertia) as a function of the number of clusters k. The gradual decrease without a pronounced elbow suggests continuous variation in ECG patterns.

1.2.2 Cluster Characterization

The KMeans algorithm identified five clusters with markedly different sizes, reflecting the heterogeneous distribution of ECG patterns in the cardiovascular cohort. Cluster sizes range from 68 patients (Cluster 4, 3.8%) to 735 patients (Cluster 1, 41.1%), indicating that certain ECG morphologies are more prevalent than others.

Figure 1.3 presents the cluster-average PAA profiles, which reveal distinct temporal patterns across the five groups. Each profile represents the mean PAA segment values for all patients assigned to that cluster, providing insight into the characteristic ECG morphology associated with each group.

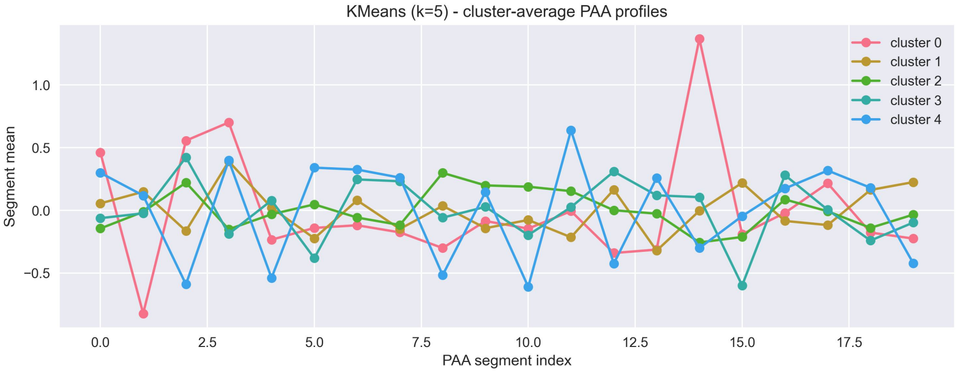


Figure 1.3: KMeans cluster-average PAA profiles for $k = 5$. Each line represents the mean PAA feature vector for patients assigned to that cluster, revealing distinct temporal patterns in ECG morphology.

The cluster profiles exhibit varying degrees of amplitude modulation and temporal dynamics. Some clusters show relatively flat profiles (indicating stable baseline patterns), while others display pronounced oscillations or gradual trends across the observation period. These differences likely reflect variations in cardiac rhythm, conduction abnormalities, or the presence of artifacts in the ECG recordings.

The clustering evaluation, summarized in Table 1.1, reveals a Silhouette Score of 0.210, which is substantially lower than the 0.941 achieved in the tabular patient profile clustering. This discrepancy reflects a fundamental difference between the two analyses: while tabular features capture discrete clinical states (e.g., presence or absence of metabolic stress), ECG time series represent continuous physiological processes with inherent variability.

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

Table 1.1: KMeans clustering evaluation metrics for ECG time series.

1.3 Hierarchical Clustering

To explore the hierarchical structure of ECG patterns and validate the KMeans results, we applied Agglomerative Hierarchical Clustering using Ward linkage. Ward linkage minimizes the within-cluster variance at each merge step, making it well-suited for identifying compact, well-separated groups in continuous feature spaces.

1.3.1 Dendrogram Analysis

Figure 1.4 presents the dendrogram for a subset of 200 patients, illustrating the hierarchical relationships between ECG patterns. The dendrogram reveals multiple levels of structure, with several distinct branches emerging before the distance between merges increases substantially.

The dendrogram structure suggests that ECG patterns form natural groupings at different granularities. At lower levels of the hierarchy, patients with highly similar morphologies are merged, while at higher levels, broader pattern categories emerge. This multi-scale structure contrasts with the binary split observed in the tabular clustering analysis, where patients were clearly separated into high-acuity and baseline groups.

1.3.2 Cluster Profiles

For comparison with the KMeans results, we extracted five clusters from the hierarchical clustering using Ward linkage. The resulting cluster-average PAA profiles, shown in Figure 1.5, exhibit similar temporal patterns to those identified by KMeans, though with different cluster assignments and sizes.

The hierarchical clustering produced clusters of sizes 459, 1001, 126, 68, and 132 patients, with a Silhouette Score of 0.231—slightly higher than KMeans but still indicating substantial overlap between clusters. This modest improvement may reflect Ward linkage's ability to identify more compact groups by minimizing within-cluster variance.

1.4 Density-Based Clustering (DBSCAN)

To account for the non-uniform density distribution and potential outliers in the ECG feature space, we applied DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Unlike partitional methods, DBSCAN does not require a pre-specified number of clusters and can identify noise points that do not belong to any dense region.

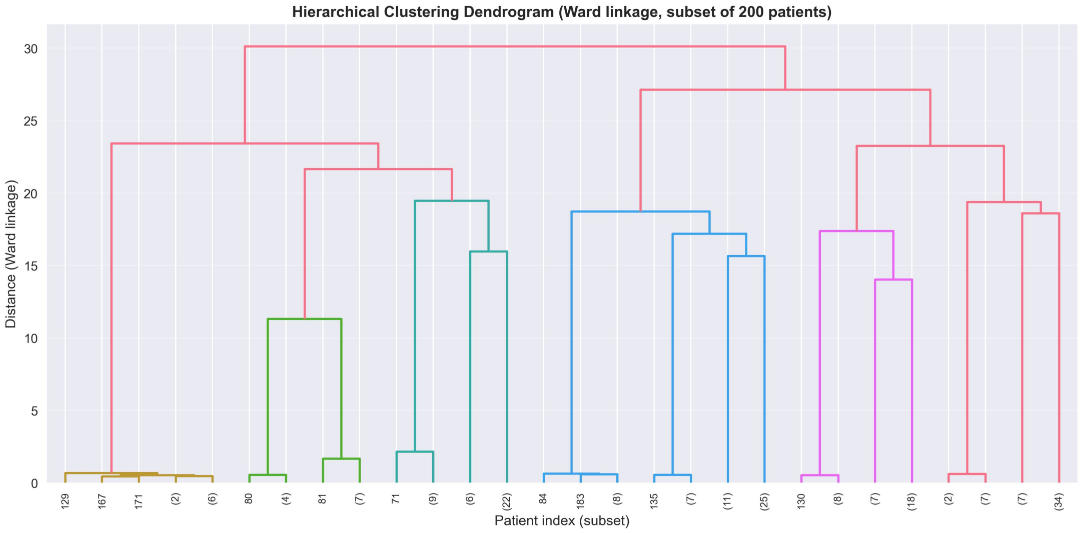


Figure 1.4: Hierarchical clustering dendrogram using Ward linkage (subset of 200 patients). The vertical axis represents the distance (dissimilarity) between merged clusters. Multiple levels of structure are evident, suggesting continuous variation in ECG patterns rather than discrete categories.

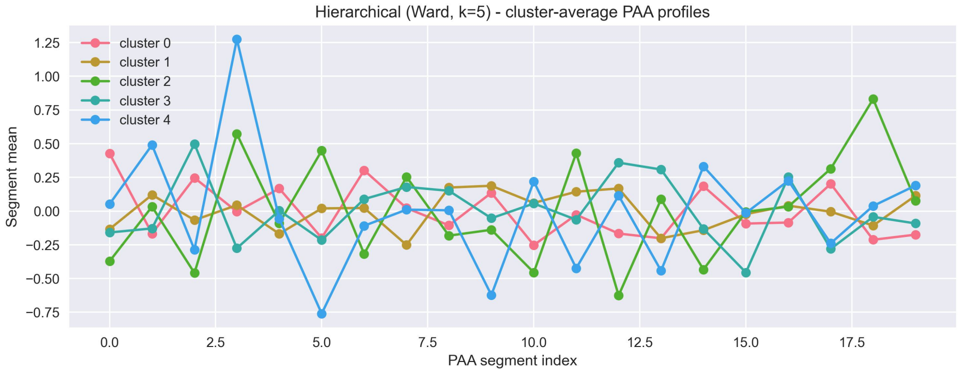


Figure 1.5: Hierarchical clustering cluster-average PAA profiles for $k = 5$ using Ward linkage. The profiles show similar temporal patterns to KMeans, though with different cluster assignments reflecting the hierarchical merging strategy.

1.4.1 Parameter Selection

We explored multiple combinations of *eps* (neighborhood radius) and *min_ samples* (minimum points required to form a dense region) to identify parameters that balance cluster coherence with noise detection. Table 1.2 summarizes the results of this parameter search.

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

Based on this analysis, we selected *eps*=0.8 and *min_ samples*=10, which produced 27 clusters and identified 168 noise points (9.4% of the dataset). This configuration provides a reasonable balance between cluster granularity and noise detection, allowing us to identify both dense pattern groups and idiosyncratic ECG morphologies.

1.4.2 Cluster Characterization

The application of DBSCAN resulted in 27 distinct clusters, ranging in size from 39 to 75 patients, plus 168 noise points. This high number of clusters—substantially more than the five identified by KMeans and Hierarchical methods—reflects DBSCAN's sensitivity to local density variations in the feature space.

Figure 1.6 presents the cluster-average PAA profiles for the non-noise clusters. The profiles exhibit considerable diversity, with some clusters showing distinct temporal patterns while others display more subtle variations. The high cluster count suggests that ECG morphological patterns form a continuum rather than discrete categories, with DBSCAN identifying local density peaks within this continuous distribution.

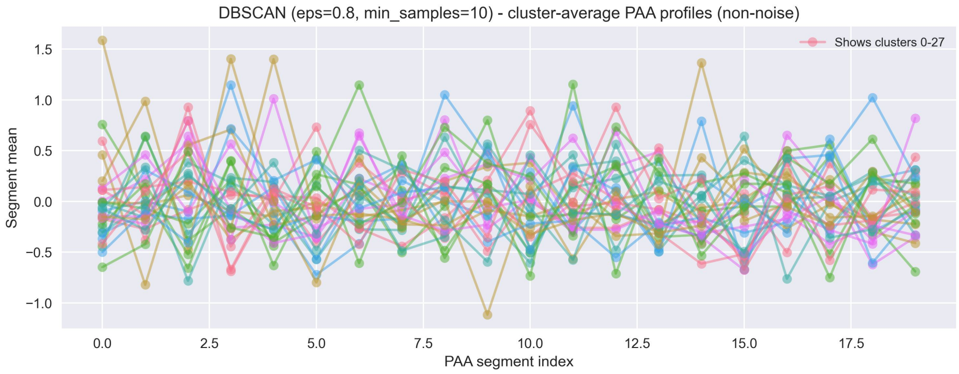


Figure 1.6: DBSCAN cluster-average PAA profiles for non-noise clusters (eps=0.8, min_samples=10). The algorithm identified 27 clusters, reflecting the high diversity of ECG patterns in the dataset.

The identification of 168 noise points (9.4% of patients) is particularly noteworthy. These patients exhibit ECG patterns that do not conform to the dense regions identified by the algorithm, potentially representing rare arrhythmias, measurement artifacts, or patients with unique clinical presentations that deviate from common patterns.

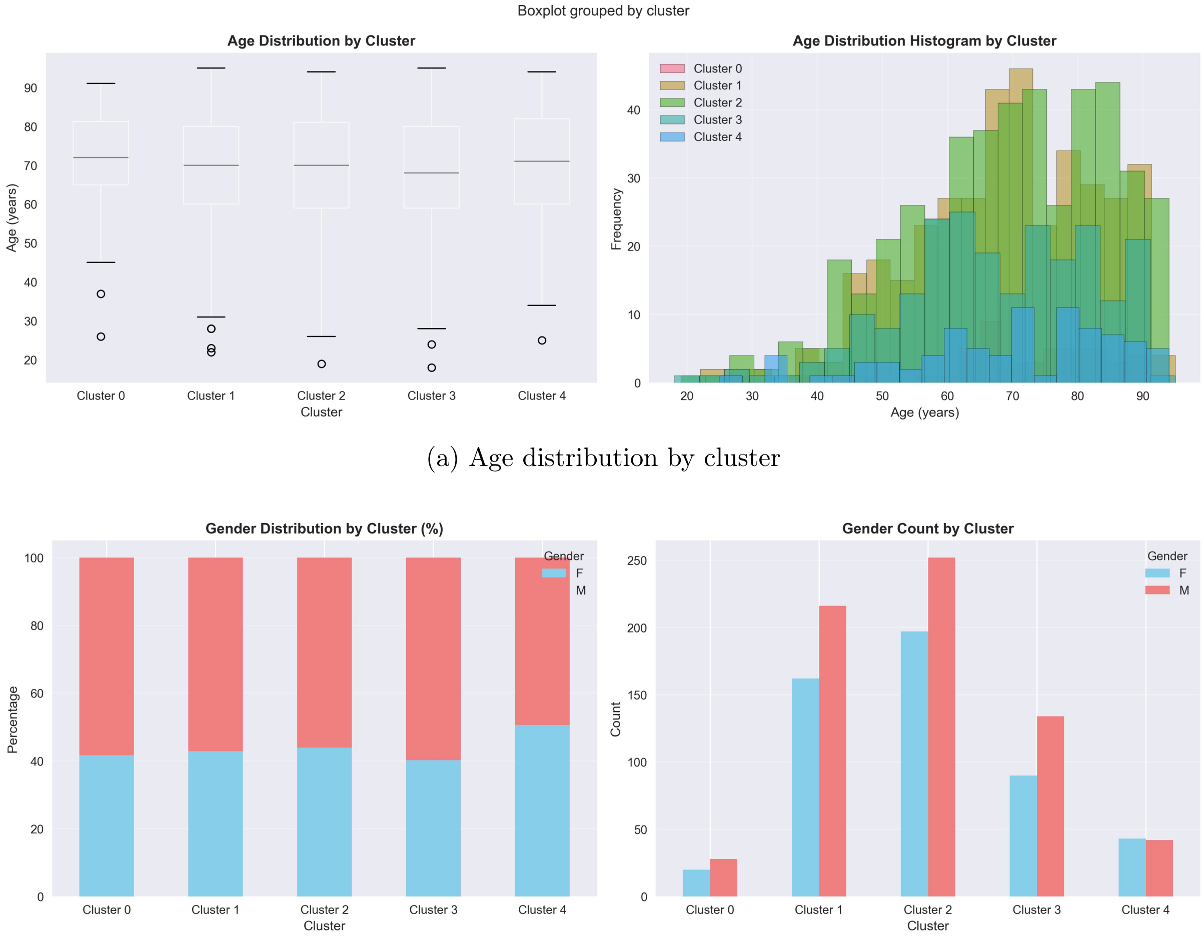
1.5 Clinical Interpretation

To assess the clinical relevance of the time series clusters, we examined the distribution of patient demographics and diagnoses across the KMeans clusters. This analysis provides insight into whether ECG morphological patterns correlate with clinical characteristics.

1.5.1 Demographic Analysis

Figure 1.7 presents the age and gender distributions across the five KMeans clusters. The age distributions are remarkably similar across clusters, with mean ages ranging from 68.4 to 70.8 years and overlapping interquartile ranges. This uniformity suggests that ECG morphological patterns are not strongly associated with patient age in this cohort.

Similarly, the gender distribution shows relatively balanced proportions across clusters, with no cluster exhibiting a pronounced gender bias. This finding indicates that the ECG patterns identified by clustering are not primarily driven by demographic factors but rather reflect underlying cardiac electrical activity patterns.



(a) Age distribution by cluster

(b) Gender distribution by cluster

Figure 1.7: Demographic characteristics across KMeans clusters. (a) Age distributions show similar means and spreads across clusters. (b) Gender distributions are relatively balanced, indicating ECG patterns are not primarily driven by demographics.

1.5.2 Diagnostic Patterns

The distribution of primary cardiac diagnoses across clusters, shown in Figure 1.8, reveals that the most common conditions—Acute Myocardial Infarction (AMI), Heart Failure, and Atrial Fibrillation—are present in all clusters with similar relative frequencies. This finding suggests that ECG morphological patterns, as captured by PAA features, do not strongly discriminate between major cardiac diagnoses at the level of resolution provided by this analysis.

Table 1.3 summarizes the top diagnoses for each cluster. While there are minor variations in the relative frequencies of conditions (e.g., Cluster 4 shows a higher proportion of AMI at 29.4%), the overall pattern is one of similarity rather than distinctiveness. This observation aligns with the low silhouette scores observed in the clustering evaluation, indicating substantial overlap between clusters.

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 1.3: Top 3 diagnoses by KMeans cluster. While Acute Myocardial Infarction (AMI) is the most common diagnosis across all clusters, the second and third most frequent diagnoses show variation, with Heart Failure and Atrial Fibrillation appearing at different ranks.

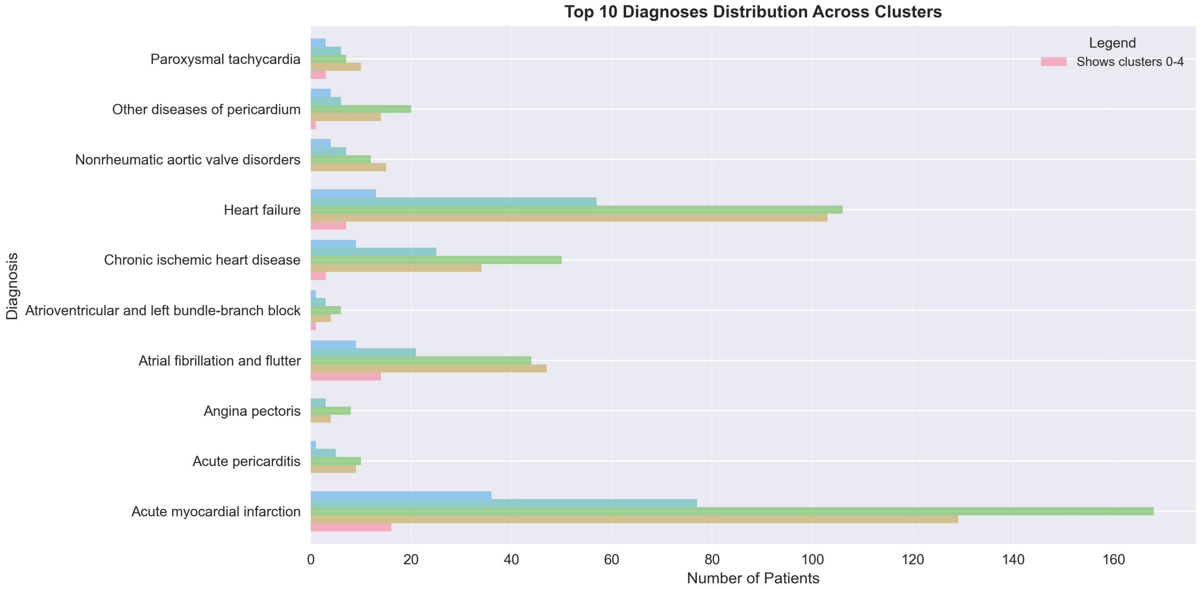


Figure 1.8: Top 10 diagnoses distribution across KMeans clusters. The most common cardiac conditions (AMI, Heart Failure, Atrial Fibrillation) appear in all clusters with similar frequencies, suggesting ECG patterns do not strongly discriminate between diagnoses at this level of analysis.

1.6 Evaluation and Comparison

Table 1.4 provides a comprehensive comparison of the three clustering approaches applied to the ECG time series data. All three methods produce relatively low silhouette scores (0.203–0.231), substantially lower than the 0.941 achieved in the tabular patient profile clustering analysis.

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

Interpretation: The low silhouette scores observed in time series clustering reflect a fundamental difference between ECG morphological patterns and tabular clinical features. While tabular features capture discrete clinical states (e.g., presence/absence of metabolic stress, history depth), ECG time series represent continuous physiological processes with inherent temporal variability. The PAA representation, while effective for dimensionality reduction, may not capture the fine-grained temporal patterns necessary to distinguish between subtle ECG morphologies.

The high number of clusters identified by DBSCAN (27 clusters) further supports the hypothesis that ECG patterns form a continuum rather than discrete categories. The algorithm's ability to identify local density peaks suggests that while certain patterns are more common (forming dense regions), the overall distribution is continuous with gradual transitions between pattern types.

Clinical Argument: The finding that ECG morphological patterns do not strongly correlate with demographic characteristics or primary diagnoses suggests that Lead II ECG signals, when analyzed at the level of overall morphology captured by PAA features, may be more indicative of general cardiac electrical activity patterns than specific disease states. This observation aligns with clinical knowledge that ECG interpretation requires expert analysis of specific waveform components (P waves, QRS complexes, ST segments) rather than overall signal shape. Future work might explore feature extraction methods that capture these fine-grained morphological characteristics, potentially improving the discriminative power of time series clustering for clinical applications.