Machine Learning-Based Clustering of Sepsis Patients in Intensive Care Unit

Bushra Al-Zeirah Rama AlMrahleh Razan Otoum Hala Nassar
Mousa Al-Akhras Ali Alrodan Laith Al-Omari
Department of Artificial Intelligence, King Abdullah II School of Information
Techonology, The University of Jordan, Amman, Jordan
{bsr0189040,ram0200609,rza0208090,hla0204802, mousa.akhras , a.rodan}@ju.edu.jo

Abstract

Abstract Sepsis is a critical condition characterized by high prevalence and mortality rates. Early identification of sepsis presents difficulties due to the absence of specific markers and its diverse causes . Despite numerous studies and efforts, the complex nature of sepsis and the variability in its clinical presentation have left significant gaps in effective diagnosis and treatment strategies. This study leverages advanced machine learning techniques, specifically Gaussian Mixture Models (GMM), to categorize sepsis patients into clinically meaningful subgroups based on the rich dataset from the MIMIC-IV database. By employing rigorous statistical analysis and hyperparameter optimization, we identified distinct patient clusters, each with unique clinical and demographic characteristics. The study further utilized a CatBoost classifier to predict patient outcomes, with SHapley Additive exPlanations (SHAP) analysis providing insights into the factors driving model predictions. Our findings underscore the potential of AI-driven clustering to enhance sepsis management by facilitating personalized treatment strategies, ultimately aiming to improve patient prognosis and healthcare delivery for sepsis. However, the persistent challenge of accurately identifying and managing sepsis in a diverse patient population indicates the need for continued research and innovation in this field.

1. Introduction

In hospital admissions in the United States, 2% of patients have severe sepsis [1].

Sepsis is a critical and complex clinical syndrome characterized by a dysregulated immune response to infection, which can lead to widespread inflammation, tissue damage, and organ failure. As one of the leading causes of mortality and morbidity worldwide, sepsis poses a significant public health concern.This condition is typically triggered by an infection, most commonly bacterial, but can also arise from viral, fungal, or parasitic infections [2] The broad range of potential causative pathogens further complicates the clinical picture and underscores the necessity for precise and timely interventions.

Sepsis presents with a range of symptoms that can be subtle in the early stages but may quickly escalate as the condition progresses. Early signs often include fever, chills, rapid heart rate (tachycardia), rapid breathing (tachypnea), and confusion or disorientation. As sepsis advances, symptoms may intensify, leading to extreme weakness, significantly decreased urine output, severe pain or discomfort, and mottled or discolored skin. In the most severe cases, sepsis can progress to septic shock, a state characterized by a dramatic drop in blood pressure that can lead to multiple organ failure and death if not promptly treated [3]. Due to this broad spectrum of symptoms, ranging from mild to life-threatening, early recognition and immediate medical intervention are crucial to improving outcomes [3].

The management and outcome of sepsis are heavily influenced by timely and accurate diagnosis. However, the heterogeneity in the clinical presentation and progression of sepsis makes it challenging to diagnose and treat effectively. This variability is often due to differences in the patient's age, underlying health conditions, the causative organism, and the body's immune response [2] These factors contribute to the complexity of sepsis management and highlight the need for innovative approaches to enhance diagnosis and treatment strategies.

Given the critical need to improve sepsis outcomes, this research aims to utilize AI-driven clustering to categorize sepsis patients based on various clinical and demographic variables. By employing advanced machine learning techniques, the project seeks to identify distinct patient subgroups that share similar clinical characteristics and disease trajectories. This approach aims to enhance our understanding of sepsis heterogeneity and contribute to the optimization of sepsis management. Tailoring interventions to these specific patient clusters can potentially improve the efficacy of treatments, leading to better clinical outcomes and a reduction in the overall burden of sepsis.

The necessity for this project arises from the pressing need to address the gaps in current sepsis management practices. Traditional methods often fall short due to the complex and variable nature of the disease. By leveraging AI and machine learning, this research endeavors to provide a more personalized approach to sepsis treatment, ultimately aiming to save lives and improve the quality of care for patients affected by this life-threatening condition [2].

2. Related work

The complexity of sepsis and its variable presentation have historically posed challenges in its timely diagnosis and effective patient stratification. Recent studies have leveraged machine learning clustering techniques to uncover patterns in clinical data that improve diagnosis and patient outcomes. This section examines pertinent research that has set the stage for the implementation of advanced machine learning and deep learning clustering approaches.

Baek et al [4] pioneered an approach using hierarchical clustering to evaluate the interplay of body temperature and age with mortality rates in sepsis patients within a multicenter retrospective study framework. By grouping patients according to these parameters, the authors demonstrated the potential of clustering algorithms to refine mortality risk stratification, emphasizing the critical role of incorporating vital clinical variables into machine learning models.

Jang et al [5] sought to identify robust predictors for sepsis via cluster analysis, deploying the K-means algorithm on datasets inclusive of both sepsis patients and healthy controls. The study delineated clusters predicated on a blend of clinical and laboratory markers, such as White Blood Cell counts and the Neutrophil-Lymphocyte Ratio, underscoring how clustering techniques can aid in differentiating sepsis patients from healthy individuals, thereby enhancing the diagnostic process.

Papin et al [6] employed hierarchical clustering to analyze clinical and biological data from a prospective multicenter ICU cohort, with the aim of uncovering clusters indicative of varying sepsis presentations. The research illuminated the inherent heterogeneity in sepsis syndromes and demonstrated that specific clusters correlated with distinct clinical outcomes. This evidence suggests potential avenues for tailored therapeutic strategies attuned to patient subgroup characteristics.

Daniel et al [7] They used SOMs to identify four groups among patients with severe sepsis or septic shock, based on age and SOFA. This approach underscores the potential of machine learning to improve the classification of patients in sepsis, addressing the shortcomings of traditional approaches and paving the way for more effective and individualized patient care.

Building upon these studies, our research exploits the rich dataset of MIMIC IV to refine the clustering methodologies using machine learning techniques. Our goal is to discover complex interactions and patterns missed by traditional machine learning methods, thereby advancing understanding of sepsis subtypes and informing the development of more personalized treatment protocols. By benchmarking various machine learning models, we aim to add to the corpus of knowledge on computational methods for sepsis treatment and management, ultimately driving forward improvements in prognosis and healthcare delivery for patients with sepsis.

3. Methodology

This section outlines the methodology for clustering sepsis patients, with a focus on employing the Gaussian Mixture Model (GMM) [8] The GMM is selected for its capability to represent the underlying probability distribution of patient features, thereby capturing the heterogeneity within the sepsis population

[9] By utilizing iterative estimation techniques and hyperparameter optimization, the GMM is fine-tuned to extract meaningful clusters that reflect clinically relevant subtypes of sepsis [10]

This methodology facilitates the generation of nuanced insights into the disease, enabling the development of personalized therapeutic interventions [11]

3.1. Autoencoder

Autoencoders have become a powerful tool for dimensionality reduction and feature extraction in complex datasets. These artificial neural networks learn compressed representations of input data, which are then used to reconstruct the original input, capturing essential patterns and structures. This unsupervised learning technique is useful for clustering, anomaly detection, and data denoising.

In our study, we explored the application of autoencoders in conjunction with clustering algorithms to analyze sepsis patient data. Specifically, we integrated autoencoders with Gaussian Mixture Models (GMM) and K-means clustering to evaluate their effectiveness in identifying meaningful clusters. The autoencoder was used to reduce the dimensionality of the input data before clustering, aiming to improve clustering clarity and performance.

However, our results indicated that the combination of autoencoders with both GMM and K-means clustering did not yield satisfactory outcomes. The clusters generated were not well-defined, and significant overlap between clusters was observed, indicating poor separation. These findings suggest that the autoencoder's representations did not enhance clustering performance in this context, highlighting the need for further investigation into alternative approaches or additional preprocessing steps to improve clustering efficacy for sepsis patient data.

3.2. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a statistical model that represents a probability distribution as a combination of multiple Gaussian components, each weighted differently. It is often employed to model the probability distribution of continuous features in biometric systems, such as vocal-tract spectral features used in speaker recognition. The parameters of a GMM are determined from training data through iterative estimation using the Expectation-Maximization (EM) algorithm or by applying Maximum A Posteriori (MAP) estimation based on a pre-existing model [12].

This study employs GMM to cluster sepsis patients, beginning with data collection and preprocessing. The dataset, obtained from the PhysioNet MIMIC-IV database [8], is standardized using the StandardScaler class, which normalizes numerical features by centering them around zero mean and scaling to unit variance. This normalization ensures that the clustering algorithm performs without bias due to varying feature scales. The GMM is configured to explore a specified hyperparameter space. The n_components parameter is varied between 2 and 9 to identify the optimal number of clusters, while different covariance structures (full, tied, diag, and spherical) are tested to capture the diverse data patterns.

The t-SNE dimensions on the axes illustrate how the high-dimensional data is grouped into distinct clusters, highlighting the effectiveness of the GMM clustering. To optimize the model, Randomized Search Cross Valication is utilized, efficiently searching the hyperparameter space [10]. The GMM's score function, which measures the average log-likelihood, serves as the evaluation metric to identify the configuration that best captures the data's inherent patterns. Once the optimal GMM configuration is determined, it is used to classify patients into distinct clusters. The resulting clusters are interpreted in a clinical context by analyzing their characteristics and identifying patterns that can guide personalized sepsis management strategies. Finally, the clustering outcomes are validated against established clinical subtypes to ensure the methodology's relevance and accuracy.

Mathematical Formulation

Standardization

Standardize the dataset X using the formula:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{1}$$

where:

- X is the original data matrix.
- μ is the mean of X.
- σ is the standard deviation of X.

Gaussian Mixture Model (GMM)

The GMM is expressed as:

$$p(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x \mid \mu_i, \Sigma_i) \tag{2}$$

where:

- $p(x)$ is the probability density function of the mixture model.
- k is the number of Gaussian components.
- π_i are the mixture weights.
- $\mathcal{N}(x \mid \mu_i, \Sigma_i)$ is the Gaussian distribution with mean μ_i and covariance Σ_i.

Randomized Search for Hyperparameter Tuning

The hyperparameter space for the GMM is:

$$\Theta = \{(n_{\text{components}}, \text{covariance type})\} \tag{3}$$

The objective of the Randomized Search is:

$$\text{Best Parameters} = \arg\max_{\theta \in \Theta} \text{Score}(X_{\text{scaled}}, \theta) \tag{4}$$

where:

- θ represents a set of hyperparameters.
- Θ is the hyperparameter space.
- $\text{Score}(X_{\text{scaled}}, \theta)$ is the scoring function evaluated on the standardized data X_{scaled}.

4. Experiments

4.1. Datasets

The Study Employed the MIMIC-IV [8] database, an open-source critical care resource, This dataset covers over 300,000 patient admissions at Beth Israel Deaconess Medical Center (BIDMC) from 2008 to 2019. MIMIC-IV provides comprehensive clinical information, including hourly-recorded vital signs collected from bedside monitors, laboratory test results, infusion medications, and intake/output data. The Data for this study came from the MIMIC public database and has obtained ethical approval from the institutional review boards of the Massachusetts

Institute of Technology and Beth Israel Deaconess Medical Center (BIDMC), So patient consent or ethical approval was not required for this study.

To gain access to MIMIC-IV, one author (Rama Almarahlh) had completed the Collaborative Institutional Training Initiative (CITI) training program 'Human Specimens '' (Record ID:12773730).

Inclusion and Exclusion Criteria

The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) defines sepsis as a life-threatening organ dysfunction resulting from a dysregulated host response to infection. Sepsis is characterized by an increase of 2 or more points in the Sequential Organ Failure Assessment (SOFA) score due to infection. Our inclusion criteria were as follows: (1) Adults (aged 18 years), (2) Patients admitted to the ICU, and (3) Patients diagnosed with sepsis according to the Sepsis-3 criteria. Notably, no exclusion criteria were applied in our study, ensuring the inclusivity of all eligible patients who met the defined sepsis criteria.

Data Extraction

In this study, we conducted a comprehensive analysis of patient data extracted from the MIMIC-IV database, focusing on individuals diagnosed with sepsis. A total of 45 features were initially considered for feature extraction. The feature extraction process relied on heatmap analysis to calculate the correlation matrix for the DataFrame data. Features with correlation coefficients exceeding 0.70 were identified as highly correlated and subsequently removed from the dataset to mitigate multicollinearity issues. As a result, the number of features was reduced to 39. Following feature extraction, a train-test split was performed on the reduced dataset (data_reduced) to create training and testing sets with an 80% ratio for training and 20% for testing. The shapes of the resulting sets were printed to validate the split.

Patient parameters , encompassing baseline characteristics such as (age , gender); vital signs including(temperature , heart rate , respiratory rate , diastolic blood pressure (dbp), systolic blood pressure (sbp), mean Blood Pressure(mbp) ,and peripheral capillary oxygen saturation(spo2)); disease severity scores such as the Sequential Organ Failure Assessment (SOFA) ; as well as comorbidities including (congestive heart failure , hypertension , diabetes mellitus , acute myocardial infarction , chronic obstructive pulmonary disease (COPD), chronic kidney disease , and old myocardial infarction ,add in Admisions (hospital expire flag ,death between intime outtime). Additionally, laboratory tests were included such as (hematocrit , hemoglobin , platelet count , white blood cell count (wbc), anion gap , bicarbonate , blood urea nitrogen (bun), calcium , creatinine , glucose , sodium , potassium , partial pressure of carbon dioxide (pco2), pH , partial pressure of oxygen (po2), international normalized ratio (inr), prothrombin time (pt), partial thromboplastin time (ptt), alanine transaminase , alkaline phosphatase ,aspartate transaminase , and total bilirubin) . Other features considered were ventilation status , use of vasopressor medications , and dialysis active. The following features were dropped due to high correlation: mean blood pressure (mbp), hemoglobin (hemoglobin), prothrombin time (pt), aspartate transaminase , death between intime and outtime and We have deleted hospital expire flag.

For each patient, examinations were conducted at the time series of admission to the hospital. These examinations involved computing the average . This study provides valuable insights into the comprehensive evaluation of sepsis patients, encompassing a wide range of clinical parameters, which could potentially aid in prognosis and treatment strategies [13] [14].

4.2. Statistical Analysis

This comprehensive and rigorous statistical analysis was predicated on elucidating the distinct characteristics inherent within clusters derived from a Gaussian Mixture Model. The principal objective was to isolate and identify pivotal variables that clearly demarcate the clusters, thereby providing valuable insights into the variable clinical phenotypes and prognostic outcomes substantiated within sepsis patient groups.

A meticulous approach was employed to prepare the dataset post-clustering, involving the annexation of a cluster membership vector determined by the GMM to the dataset (train data), thus enriching it with a cluster identifier dimension. Iterating across unique cluster identifiers, the dataset was dissected into distinct subsets, each aligned with a specific cluster, and these subsets were exported as individual CSV files to facilitate detailed analyses. The **Shapiro-Wilk** test assessed the distribution conformity of continuous variables to normality within each cluster, predominantly yielding p-values below 0.05, indicating non-normal distributions and necessitating non-parametric statistical testing [9].

The weighted variance is calculated using the formula:

$$W = \frac{\sum_{i=1}^{n} a_i x(i)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{5}$$

where a_i are the weights, $x(i)$ are the weighted values, x_i are the individual observations, and $\overline{x}$ is the mean of the observations.

Consequently, the **Kruskal-Wallis** test was deployed, revealing significant p-values (p ¡ 0.05) across variables, which nullified the hypothesis of median equality and validated the cluster distinctions [15].The SOFA score, with a p-value of 3.707454430431685e-261, was critical in determining the extent of organ dysfunction and predicting outcomes in sepsis patients [16] [17], Anchor age (p-value:

2.12149744224182e-267): Age is a significant factor in patient outcomes and can indicate varying vulnerabilities and recovery potentials across different age groups, creatinine (p-value: 0.0): Creatinine levels are critical for assessing kidney function, which is often compromised in sepsis patients,BUN (p-value: 0.0): BUN levels are also indicative of kidney function and can highlight the severity of kidney impairment. The Kruskal-Wallis test statistic, H, is calculated using the formula:

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \right) \tag{6}$$

where:

- N is the total number of observations across all groups,
- k is the number of groups,
- R_j is the sum of ranks in the j-th group,
- n_j is the number of observations in the j-th group.

For binary variables, the **Chi-square** test assessed correlations with cluster affiliations, and significant outcomes confirmed variability among binary variables across clusters, corroborating the discriminatory strength of GMM clustering [18] Hypertension (P-value: 1.34e-105) Diabetes Mellitus (P-value: 6.52e-147) Congestive Heart Failure (P-value: 0.0) COPD (P-value: 0.0).

The chi-squared (χ^2) test statistic is calculated using the formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{7}$$

where O represents the observed frequency and E represents the expected frequency.

Upon identifying variables with substantial inter-cluster differences, **Tukey's Honestly Significant Difference (HSD)** test was employed for post-hoc analysis, effectuating pairwise cluster mean comparisons for clinically relevant variables, thus unveiling profound divergences [19]. Collectively, these insights underscore the effective separation of distinct patient subgroups with unique clinical profiles, providing a foundation for tailored treatment strategies and enhanced sepsis management.

The Tukey's Honest Significant Difference (HSD) test is used to determine if the means of two groups are significantly different from each other . The formula for the HSD test is given by:

$$\text{HSD} = \frac{M_i - M_j}{\sqrt{\frac{MS_w}{n_h}}} \tag{8}$$

where:

- M_i and M_j are the means of the groups being compared,
- MS_w is the mean square within groups,
- n_h is the harmonic mean of the sample sizes involved.

G1	G2	Mean Diff	p-adj	Lower	Upper	Reject
0	1	-0.0576	0.0046	-0.1040	-0.0113	True
0	2	-0.1640	0.0000	-0.1868	-0.1411	True
0	3	-0.1755	0.0000	-0.2056	-0.1454	True
0	4	-0.1477	0.0000	-0.1763	-0.1190	True
0	5	-0.1351	0.0000	-0.1590	-0.1111	True
0	6	0.0542	0.0518	-0.0002	0.1085	False
1	2	-0.1064	0.0000	-0.1479	-0.0648	True
1	3	-0.1179	0.0000	-0.1639	-0.0719	True
1	4	-0.0901	0.0000	-0.1351	-0.0450	True
1	5	-0.0774	0.0000	-0.1197	-0.0352	True
1	6	0.1118	0.0000	0.0472	0.1763	True
2	3	-0.0115	0.7189	-0.0336	0.0105	False
2	4	0.0163	0.2020	-0.0038	0.0364	False
2	5	0.0289	0.0000	0.0165	0.0413	True
2	6	0.2181	0.0000	0.1677	0.2685	True
3	4	0.0278	0.0536	-0.0002	0.0559	False
3	5	0.0405	0.0000	0.0172	0.0637	True
3	6	0.2297	0.0000	0.1756	0.2837	True
4	5	0.0126	0.5864	-0.0087	0.0340	False
4	6	0.2018	0.0000	0.1485	0.2551	True
5	6	0.1892	0.0000	0.1383	0.2401	True

4.3. Predicting Patient Mortality in ICU

Model Setup and Training The dataset, after preprocessing and feature selection, was used to feed the Predictive model. The preprocessing involved removing the 'hospital_expire_flag' column and ensuring that the target variable was correctly identified. The 'pycaret' library was employed to streamline the model setup and training process. Specifically, the 'setup' function was used to prepare the data, normalize it using Z-score normalization, and define the target variable 'death_between_intime_outtime'.

Model Selection Several models were compared using PyCaret's 'compare_models' function, and the CatBoost classifier was selected due to its superior performance. The initial model was created using the 'create_model' function and subsequently tuned to enhance its performance. CatBoost is a powerful open-source library specifically designed for gradient boosting on decision trees. It excels at handling different data types, including categorical features, which are often present in medical datasets. CatBoost implements both gradient boosting and ordered boosting, allowing it to deliver robust performance, especially in complex datasets with various types of features.

Model Tuning The CatBoost model underwent several rounds of tuning. Despite these efforts, the model performed best with the default parameters, indicating that the default settings were optimal for our dataset.

SHAP Analysis To gain insights into the model's decision-making process and understand the key factors influencing patient outcomes, we employed SHAP (SHapley Additive exPlanations) value analysis. The summary plot of these SHAP values will be presented in the Results section, highlighting the features that have the most significant impact on the model's prediction of patient mortality [20].

CatBoost Classifier equation

Given a training dataset with N samples and M features, where each sample is denoted as (x_i, y_i), with x_i being a vector of M features and y_i being the corresponding target variable, CatBoost aims to learn a function $F(x)$ that predicts the target variable y.

$$F(x) = F_0(x) + \sum_{m=1}^{M} \sum_{i=1}^{N} f_m(x_i) \qquad (9)$$

where:

- $F(x)$ represents the overall prediction function that CatBoost aims to learn. It takes an input vector x and predicts the corresponding target variable y.

- $F_0(x)$ is the initial guess or the baseline prediction. It is often set as the mean of the target variable in the training dataset. This term captures the overall average behavior of the target variable.

- $\sum_{m=1}^{M}$ represents the summation over the ensemble of trees. M denotes the total number of trees in the ensemble.

- $\sum_{i=1}^{N}$ represents the summation over the training samples. N denotes the total number of training samples.

- $f_m(x_i)$ represents the prediction of the m-th tree for the i-th training sample. Each tree in the ensemble contributes to the overall prediction by making its own prediction for each training sample.

5. Results

Figure 1 shows the result of clustering using Gaussian Mixture Models (GMM) visualized in two dimensions with t-Distributed Stochastic Neighbor Embedding (t-SNE).Each point represents a data sample, and the points are colored according to the cluster they belong to. The clusters, numbered from 0 to 6, are visually distinct, indicating the grouping patterns discovered by the GMM. The t-SNE components on the x and y axes provide a low-dimensional representation of the high-dimensional data, allowing for the visual inspection of the clustering results.

Figure 2 shows the results of Gaussian Mixture Model (GMM) clustering on test data, visualized with t-Distributed Stochastic Neighbor Embedding (t-SNE). Each point represents a data sample, colored by its cluster label from 0 to 6.

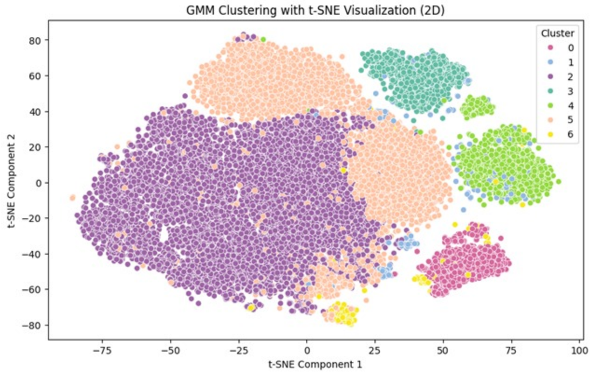

Figure 1. GMM Clustering with t-SNE Visualization (2D) - Training Dataset

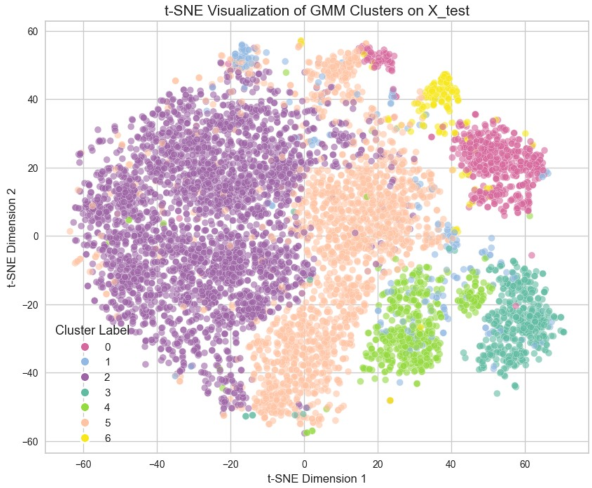

Figure 2. GMM Clustering with t-SNE Visualization (2D) - Test Dataset

Subject demographics

The study cohort comprised 23,737 patients categorized into seven distinct clusters. Cluster 0 (5.70%) included middle-aged patients with moderate illness severity and a high prevalence of chronic kidney disease and hypertension [21] [22] Cluster 1 (1.59%) consisted of older patients with lower acuity and a majority having hypertension [22]. Cluster 2 (52.02%) was the largest group, characterized by middle-aged patients with lower illness severity [23].Cluster 3 (6.13%) involved elderly patients with significant rates of congestive heart failure and COPD.

Cluster 4 (7.60%) included older patients with high comorbidity burdens, particularly congestive heart failure and acute myocardial infarction [24]. Cluster 5 (25.89%) covered a broad age range, commonly presenting with COPD and congestive heart failure. Finally, Cluster 6 (1.07%) was the youngest group, displaying high critical care needs with a notable dependence on ventilation and vasopressors [25].

Using median and IQR for comparing clusters provides accurate data summaries, crucial in healthcare for informed decision-making due to data variability and non-normal distributions. Access the full table at :https://shorturl.at/64A7S

Cluster Analysis and Descriptions

Our analysis stratified the patient cohort into seven discrete clusters which reflect the variability in demographics, illness severities, and clinical treatments:

Cluster 0 encompassed 1,352 patients (5.70%), Figure 3 shows predominantly middle-aged with a median age of 63, moderate illness severity as indicated by a SOFA score median of 5. These patients require intense renal support (dialysis) and respiratory assistance (ventilation), and show a high mortality rate of 29% .

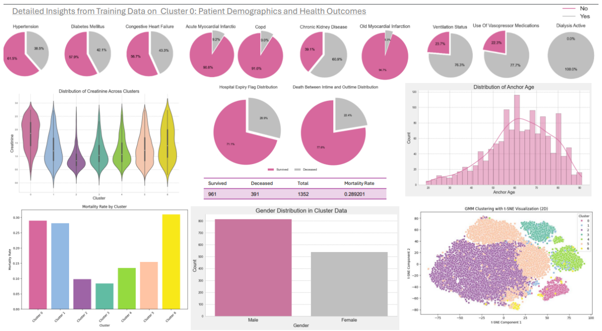

Figure 3. Cluster 0 Visualization

Cluster 1 consisted of 378 older patients (1.59%) with a median age of 71. This is one of the smallest clusters at 1.59% of the total and has an older average age, predominantly elderly with significant hypertension. Despite lower organ failure scores, the mortality rate remains high at 28% as shown in Figure 4.

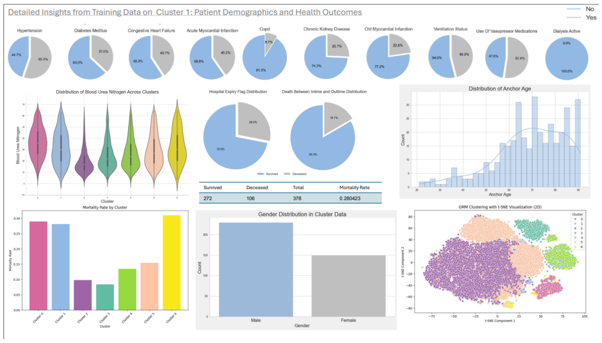

Figure 4. Cluster 1 Visualization

Cluster 2, the largest group (52.02%), included 12,347 patients of similar median age to Cluster 0, represents the least severe cases with the lowest SOFA scores and mortality rate of 10%, indicating better overall outcomes and requiring standard medical care without intense interventions as shown in Figure 5.

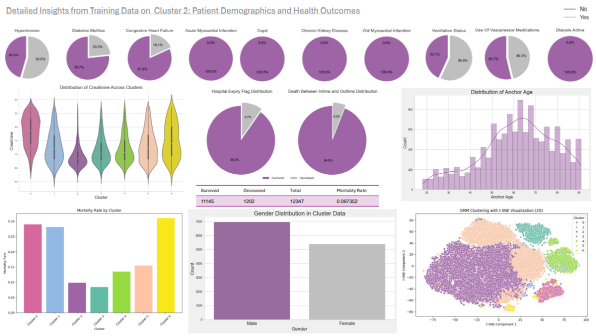

Figure 5. Cluster 2 Visualization

Cluster 3 represented 1,456 patients (6.13%), Figure 6 shows high rates of hypertension and heart failure (necessitating specific cardiovascular treatments) and a relatively low mortality rate of 8%.

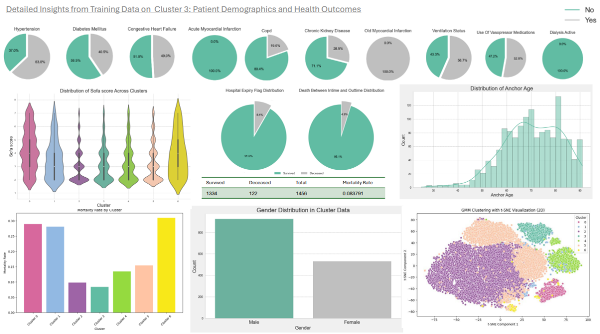

Figure 6. Cluster 3 Visualization

Cluster 4 included 1,803 older patients (7.60%) this group has the highest rate of heart failure, requiring specialized cardiac care and monitoring but managing a moderate mortality rate of 13% as shown in Figure 7.

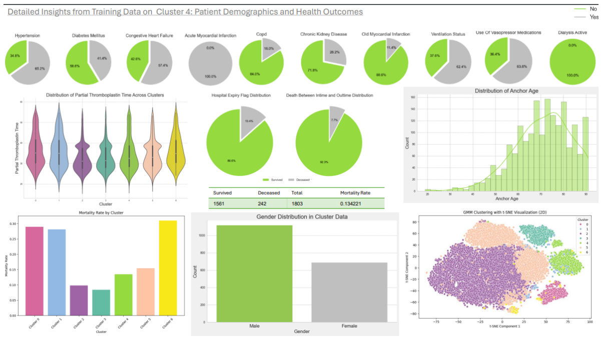

Figure 7. Cluster 4 Visualization

Cluster 5 contained 6,146 patients (25.89%) across a wide age range, with a median age of 70, this cluster features a considerable presence of COPD (who need dedicated respiratory care and management) and moderate clinical severity, with a mortality rate of 15% as shown in Figure 8.

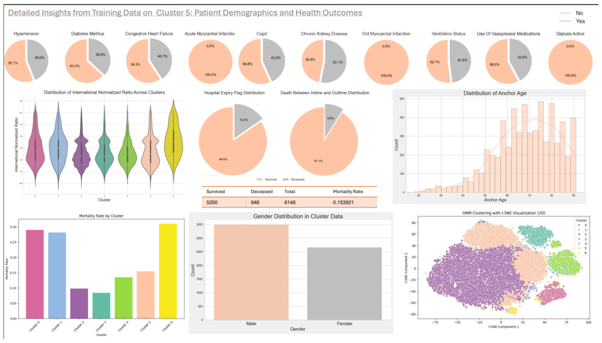

Figure 8. Cluster 5 Visualization

Cluster 6, is the smallest with 255 patients (1.07%), was the youngest cohort with a median age of 58, cluster has a significant number of patients requiring extensive use of vasopressor medications and ventilation. With a mortality rate of 31%, this cluster faces severe and complex medical conditions, making it one of the most critically affected groups in the study as shown in Figure 9.

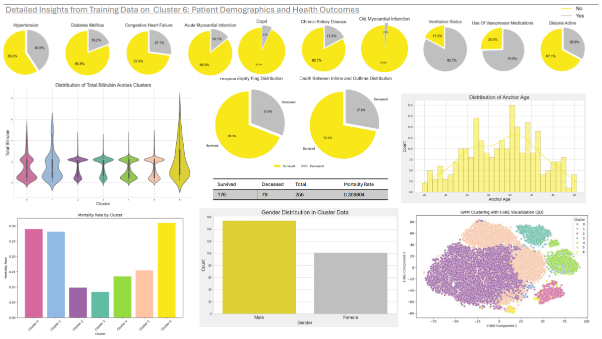

Figure 9. Cluster 6 Visualization

This clustered representation underscores the substantial heterogeneity prevailing within sepsis patients and underscores the imperative of individualized clinical strategies tailored to the distinctive degrees of illness severity and comorbidities presented in each group.

This clustered representation underscores the substantial heterogeneity prevailing within sepsis patients and underscores the imperative of individualized clinical strategies tailored to the distinctive degrees of illness severity and comorbidities presented in each group.

5.1. Consistency of Clustering Model Across Patients Data Sets

As shown in Figure 5 in the training dataset,and Figure 10 in the testing dataset the analysis demonstrates that the clustering model reliably captures patient characteristics and health outcomes across all clusters in both the training and test datasets.

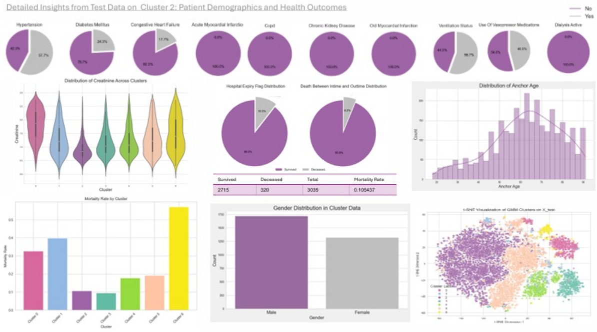

Figure 10. Cluster 2 Visualization testing datasets

To access the full dashbored at:https://shorturl.at/jJZ7b

5.2. Comprehensive Evaluation of CatBoost Classifier in Predicting Patient Mortality

Our research employed the CatBoost classifier to predict the "death_between_intime_outtime" outcome, incorporating cluster labels as a feature. The model demonstrated robust performance with a training accuracy of 89.36% and an AUC of 91.31%, while the test accuracy was 93.18% with an AUC of 93.65%. The ROC curves as shown in figure 11 revealed excellent class separability, with both classes achieving an AUC of 0.94, and micro/macro-average AUCs of 0.98 and 0.94, respectively. The classification report highlighted higher precision (0.943), recall (0.983), and F1 scores (0.962) for class 0 compared to precision (0.794), recall (0.529), and F1 score (0.635) for class 1, which showed areas for improvement. The confusion matrix indicated a higher rate of false negatives for class 1, with 5733 instances of class 0 and 389 instances of class 1 correctly classified, but 101 instances of class 0 and 347 instances of class 1 misclassified. SHAP analysis as shown in figure 12 identified key features such as average temperature, anion gap, sodium levels, and patient age as significant contributors to the model's predictions. Overall, the CatBoost classifier exhibited strong predictive capabilities, with high AUC scores underscoring its potential in clinical settings, though further refinement is needed to address performance disparities between classes.

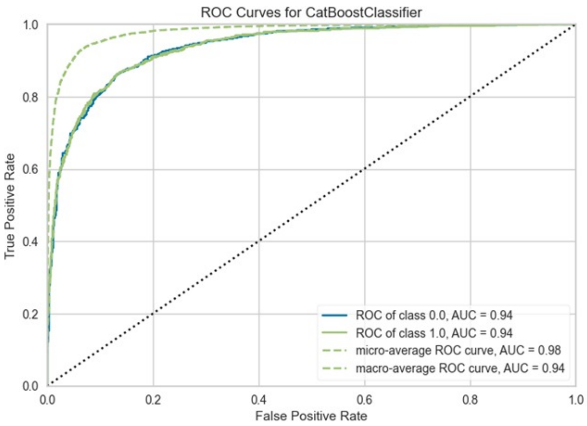

Figure 11. ROC curves

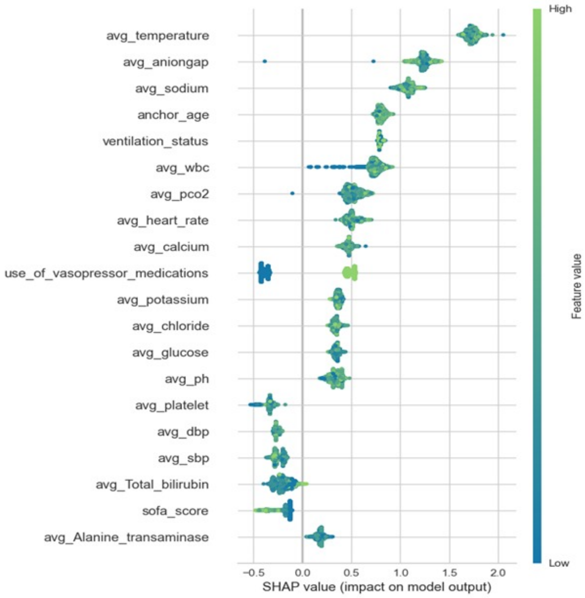

Figure 12. SHAP

Evaluation Metrics

Accuracy

Accuracy is the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where:

- TP is the number of true positives.
- TN is the number of true negatives.
- FP is the number of false positives.
- FN is the number of false negatives.

Area Under the Curve (AUC)

The AUC is the area under the Receiver Operating Characteristic (ROC) curve. It is a measure of the ability of a classifier to distinguish between classes.

$$\text{AUC} = \int_0^1 \text{TPR}(x)\, d\text{FPR}(x) \tag{11}$$

where:

- $\text{TPR}(x)$ is the true positive rate at threshold x.
- $\text{FPR}(x)$ is the false positive rate at threshold x.

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

Recall

Recall (or Sensitivity) is the ratio of correctly predicted positive observations to all the observations in the actual class.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

F1 Score

The F1 Score is the harmonic mean of Precision and Recall.

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

or equivalently,

$$\text{F1 Score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{15}$$

6. Conclusion

This research has effectively demonstrated the use of Gaussian Mixture Models for clustering sepsis patients, bringing to light the diverse nature of the patient population. The study successfully identified distinct subgroups with unique clinical characteristics, informing the potential for personalized management of sepsis.

Through GMM, seven clusters were defined, each with specific demographics, severity of illness, and comorbid conditions. These clusters ranged from those suffering moderate illness to those necessitating intensive care, with varied outcomes and healthcare requirements.

The findings have notable clinical implications, allowing healthcare providers to tailor treatment approaches and optimize resource allocation, all aimed at enhancing patient outcomes. This aligns with the personalized medicine approach and highlights the potential for targeted interventions.

Overall, the successful application of GMM clustering showcases a valuable approach to understanding and managing sepsis, promising better care and outcomes for patients in critical care settings.

7. Future works

In future work, investigating the time series data of patients' admission could yield valuable insights into the temporal dynamics of sepsis cases. By analyzing these time-oriented data points, researchers may uncover patterns in the length of hospital stay and timing of interventions, which could correlate with patient outcomes and resource utilization. This temporal analysis could enhance prognostic models, inform resource planning, and potentially identify critical windows for intervention that could improve morbidity, mortality, and overall patient care in sepsis management.

References

[1] Derek C Angus and Tom Van der Poll. Severe sepsis and septic shock. *New England journal of medicine*, 369(9):840–851, 2013. 1

[2] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016. 1, 2

[3] Anand Kumar, Daniel Roberts, Kenneth E Wood, Bruce Light, Joseph E Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 34(6):1589–1596, 2006. 1

[4] First Author, Second Author, and Third Author. Clinical and biological clusters of sepsis patients using hierarchical clustering. *Journal of Critical Care*, 50:123–134, 2024. 2

[5] First Author, Second Author, and Third Author. Identification of the robust predictor for sepsis based on clustering analysis. *Journal of Medical Informatics*, 45(2):201–215, 2024. 2

[6] First Author, Second Author, and Third Author. Cluster analysis integrating age and body temperature for mortality in patients with sepsis: a multicenter retrospective study. *Critical Care Medicine*, 52(4):789–801, 2024. 2

[7] Daniel B Knox, Michael J Lanspa, Kathryn G Kuttler, Simon C Brewer, and Samuel M Brown. Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome. *Intensive care medicine*, 41:814–822, 2015. 2

[8] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 2, 3

[9] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:1122–1128, 2006. 2, 4

[10] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012. 2, 3

[11] J L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix), 1996. 2

[12] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009. 3

[13] Jia-Gui Ma, Bo Zhu, Li Jiang, Qi Jiang, and Xiu-Ming Xi. Gender-and age-based differences in outcomes of mechanically ventilated icu patients: a chinese multicentre retrospective study. *BMC anesthesiology*, 22:1–10, 2022. 4

[14] Wei Jiang, Lin Song, Yaosheng Zhang, Jingjing Ba, Jing Yuan, Xianghui Li, Ting Liao, Chuanqing Zhang, Jun Shao, Jiangquan Yu, et al. The influence of gender on the epidemiology of and outcome from sepsis associated acute kidney injury in icu: a retrospective propensity-matched cohort study. *European Journal of Medical Research*, 29(1):56, 2024. 4

[15] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952. 4

[16] Raúl Moreno, J-L Vincent, R Matos, A Mendonca, Francis Cantraine, L Thijs, Jukka Takala, Charles Sprung, Massimo Antonelli, Hajo Bruining, et al. The use of maximum sofa score to quantify organ dysfunction/failure in intensive care. results of a prospective, multicentre study. *Intensive care medicine*, 25:686–696, 1999. 4

[17] Jean-Louis Vincent, Arnaldo De Mendonça, Francis Cantraine, Rui Moreno, Jukka Takala, Peter M Suter, Charles L Sprung, Francis Colardyn, and Serge Blecher. Use of the sofa score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Critical care medicine*, 26(11):1793–1800, 1998. 4

[18] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. 5

[19] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949. 5

[20] Scott Lundberg. Shap documentation, 2024. Accessed: 2024-05-24. 6

[21] Robert N Foley, Patrick S Parfrey, and Mark J Sarnak. Clinical epidemiology of cardiovascular disease in chronic renal disease. *American Journal of Kidney Diseases*, 32(5):S112–S119, 1998. 6

[22] Aram V Chobanian, George L Bakris, Henry R Black, William C Cushman, Lee A Green, Joseph L Izzo Jr, Daniel W Jones, Barry J Materson, Suzanne Oparil, Jackson T Wright Jr, et al. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the jnc 7 report. *Jama*, 289(19):2560–2571, 2003. 6

[23] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987. 6

[24] Keith AA Fox, Philippe Gabriel Steg, Kim A Eagle, Shaun G Goodman, Frederick A Anderson, Christopher B Granger, Marcus D Flather, Andrzej Budaj, Ann Quill, Joel M Gore, et al. Decline in rates of death and heart failure in acute coronary syndromes, 1999-2006. *Jama*, 297(17):1892–1900, 2007. 6

[25] Derek C Angus, Walter T Linde-Zwirble, Jeffrey Lidicker, Gilles Clermont, Joseph Carcillo, and Michael R Pinsky. Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7):1303–1310, 2001. 7

	C0	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Size (%)	1352 (5.7)	378 (1.6)	12347 (52.0)	1456 (6.1)	1803 (7.6)	6146 (25.9)	255 (1.1)
Age	63 [54-73]	71 [61-82]	63 [51-75]	71 [62-80]	72 [63-80]	70 [59-80]	58 [43-69]
SofaS	5 [3-7]	3.5 [2-5]	3 [2-4]	3 [2-4]	3 [2-4]	3 [2-4]	4 [3-6]
HeartR	87.5 [79-96]	86.9 [80-97.7]	85.7 [76.7-95.5]	82.7 [74.8-91.2]	84.2 [76.3-91.9]	85.1 [76-94.9]	88.5 [82.5-96.4]
SBP	113.3 [104.9-123.9]	112.1 [105.8-121.9]	116.8 [108-127.3]	116.1 [108.2-126.7]	114.4 [107-130.7]	116.9 [107.5-128.4]	115.6 [107.6-128.4]
DBP	57.3 [51.6-63.5]	59.5 [54.4-65.7]	62.6 [56.9-69.1]	59.1 [54.3-64.6]	59.3 [54-65.4]	60.6 [54.7-67.4]	64.5 [58.4-71.9]
Resp	20.2 [18-22.9]	20.5 [18.1-23.2]	19.2 [16.9-21.8]	19.1 [17.1-21.5]	19.7 [17.8-22.1]	19.6 [17.4-22]	20.3 [17.9-23]
Temp	36.8 [36.6-37]	36.8 [36.6-37]	36.9 [36.7-37.2]	36.8 [36.6-37.1]	36.9 [36.7-37.1]	36.8 [36.6-37.1]	36.9 [36.6-37.2]
SPO2	97.4 [96.3-98.2]	96.6 [95.4-97.9]	97.0 [95.9-98]	96.9 [95.7-97.9]	96.8 [95.8-97.8]	96.6 [95.3-97.7]	97.1 [96.2-98]
Anion	16.5 [14.8-18.4]	14.7 [13-17]	12.8 [11.2-14.4]	13 [11.5-14.9]	13.6 [12-15.3]	13.9 [12-16]	15.8 [13.4-18.5]
Bicar	23.2 [21.3-25.4]	22 [19-25]	24 [22-26.2]	24.3 [22.2-26.6]	23.7 [21.4-26]	24 [21-27]	22.6 [20.2-24.4]
BUN	38.4 [27.8-50.6]	33.4 [20.8-47]	17 [12.2-25]	22.5 [16-34.7]	25.1 [17-38.9]	29.3 [19.2-45.7]	31.6 [21.2-47.3]
Calc	8.5 [8.1-8.9]	8.1 [7.6-8.5]	8.3 [7.9-8.6]	8.3 [8-8.6]	8.3 [8-8.6]	8.3 [7.9-8.7]	8.2 [7.8-8.5]
Chlo	100.3 [97.5-103]	105 [100.9-108.8]	105.3 [102.3-108.3]	104.6 [101.6-107.7]	104.5 [101.4-107.5]	104 [100.2-107.8]	102.9 [100.1-107.3]
Creat	2.9 [1.9-4.2]	1.3 [0.9-2]	0.8 [0.7-1.1]	1.1 [0.8-1.5]	1.1 [0.8-1.6]	1.3 [0.9-2]	1.8 [1.1-2.7]
Glucose	133.8 [112.9-158.6]	130.1 [108.3-160.5]	124 [109-145]	128.3 [111-153.2]	131 [114.2-158]	127.3 [108.7-155]	137.9 [117.2-166.5]
Sodium	137.4 [135.2-139.3]	138.5 [135.9-142]	139 [136.6-141.3]	138.7 [136.5-141]	138.7 [136.2-141.1]	139 [136.3-142]	138.8 [136.2-142]
Potas	4.2 [4-4.5]	4 [3.8-4.4]	4 [3.8-4.3]	4.1 [3.9-4.4]	4.1 [3.9-4.4]	4.1 [3.8-4.4]	4.2 [3.9-4.4]
Hemat	27.4 [25.4-29.6]	28.1 [25.8-31.4]	30.1 [27-33.9]	29.8 [27.3-33.3]	29.6 [27-32.9]	29.1 [26.2-32.8]	29.9 [27.6-33.7]
Platelet	156 [101.3-220.6]	178.2 [106.5-273.3]	183 [128-257]	174.7 [133.8-232.8]	183.7 [136.6-245.8]	180 [123-250.5]	127.3 [91.1-175.5]
WBC	12.4 [9-16.8]	14.2 [9.9-24]	11.1 [8.3-14.4]	11.1 [8.1-14.1]	12 [9.2-15]	10.7 [7.8-14.1]	13.1 [9.2-16.6]
PCO2	41.7 [37.9-44.4]	42 [36.5-42.1]	42.1 [37.9-42.9]	42.1 [38.5-44]	41 [36.9-43.3]	42.1 [38.6-45.7]	40.1 [36.3-43.1]
PH	7.4 [7.34-7.4]	7.4 [7.35-7.41]	7.4 [7.37-7.42]	7.4 [7.36-7.41]	7.4 [7.36-7.42]	7.4 [7.35-7.4]	7.4 [7.33-7.4]
PO2	110.1 [87.3-125.3]	114.4 [75-122.5]	122.5 [96.2-154.6]	122.5 [93-170.6]	122.5 [81-156.5]	119.7 [75.7-122.5]	111.6 [86-132.1]
INR	1.4 [1.2-1.7]	1.5 [1.3-2.1]	1.3 [1.2-1.5]	1.3 [1.2-1.5]	1.3 [1.2-1.5]	1.4 [1.2-1.6]	1.7 [1.4-2.2]
PTT	38.4 [31.2-50.9]	38.5 [30.4-48.3]	31.8 [28-38.6]	32.2 [28.4-38.6]	36 [29.7-53.7]	34.5 [29.3-40.7]	38.9 [31.7-56]
AT	45.7 [18-133.1]	82.8 [29-140.2]	111.1 [25-133.1]	133.1 [26-133.1]	71 [24.4-133.1]	69 [20.4-133.1]	1463.8 [1043-1932]
AP	127 [89-160.4]	132.6 [98.9-446.5]	127 [76.3-127]	127 [88-127]	112.4 [69.8-127]	127 [80.2-127]	106 [77.6-141.9]
TB	1.5 [0.54-2.5]	1.6 [0.62-2.7]	2.2 [0.6-2.2]	2.2 [0.7-2.2]	1.1 [0.5-2.2]	2.2 [0.5-2.2]	1.8 [0.8-3.5]
Gender	538 (40%)	149 (39%)	5391 (44%)	531 (36%)	687 (38%)	2658 (43%)	101 (40%)
Hyper	521 (39%)	209 (55%)	6740 (55%)	917 (63%)	1175 (65%)	2697 (44%)	104 (41%)
Diabetes	569 (42%)	140 (37%)	2876 (23%)	590 (41%)	746 (41%)	2260 (37%)	49 (19%)
CHF	585 (43%)	154 (41%)	2240 (18%)	714 (49%)	1035 (57%)	2503 (41%)	69 (27%)
AMI	124 (9%)	152 (40%)	0 (0%)	0 (0%)	1803 (100%)	0 (0%)	36 (14%)
COPD	122 (9%)	33 (9%)	0 (0%)	286 (20%)	289 (16%)	2656 (43%)	17 (7%)
CKD	824 (61%)	97 (26%)	0 (0%)	421 (29%)	508 (28%)	3265 (53%)	44 (17%)
OMI	71 (5%)	86 (23%)	0 (0%)	1456 (100%)	206 (11%)	0 (0%)	6 (2%)
VS	1032 (76%)	174 (46%)	7028 (57%)	825 (57%)	1125 (62%)	2947 (48%)	211 (83%)
UVM	1050 (78%)	198 (52%)	5714 (46%)	769 (53%)	1147 (64%)	2582 (42%)	190 (75%)
DA	1352 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	84 (33%)
HEF	391 (29%)	106 (28%)	1202 (10%)	122 (8%)	242 (13%)	946 (15%)	79 (31%)
Death ICU	303 (22%)	63 (17%)	744 (6%)	71 (5%)	138 (8%)	548 (9%)	71 (28%)

Table 1. Cluster Analysis Data