SIADS 696: Milestone II- Team project report

Team members: 13-Nicholas Abram- Kasra Afzali- Sawsan Allam

Predicting ED Admission Disposition and Patient Clustering for Resource Optimization

Introduction

Timely care in the Emergency Departments (ED) is paramount for patient outcomes and hospital efficiency. Studies have shown that delays in care increase the risk of complications, disease progression, and in-hospital adverse events, necessitating prolonged hospitalization particularly among older patients who form a significant proportion of ED visits (Morley et al., 2018). As such, increased length of stay (LOS) in the ED department often leads to ED crowding, which strains hospital resources and reduces overall efficiency. For instance, prolonged ED stays translate into delayed transfer of admitted patients to inpatient units, leading to ED boarding and reduced bed turnover rates. Additionally, delays in treatment and inpatient admission increase operational costs since extended stays require more medication, repeated diagnostic tests, and higher resource consumption, imposing a significant financial burden on hospitals. One of the primary drivers of ED crowding and poor patient flow is ineffective patient disposition (Asplin et al., 2003; Lucke et al., 2018)—determining whether a patient should be admitted, discharged, or transferred. Improved admission predictions and deeper insights into ED data distributions, by equipping ED physicians with a reliable decision-support tool, leads to targeted intervention strategies, reduced wait times, lower operational costs, and better patient experiences. By proactively managing high-risk patients, hospitals can ease ED congestion and enhance efficiency.

Our project aims to develop a predictive model to forecast ED admission disposition while leveraging patient clustering techniques to enhance resource allocation. This project was inspired by one team member, an ED physician with over 20 years of experience who has witnessed firsthand the persistent bottlenecks in patient flow, delays in admission decisions, and resource inefficiencies in EDs. Surprisingly, none of the twenty hospital EDs he has worked at implemented such a tool, despite technological advancements. Literature agrees also that no admission prediction tool has been implemented in practice (Brink et al., 2022). As such, a gap exists between technological potential and clinical practice, underscoring the need for a data-driven, AI-assisted approach. A systematic review by Brink et al. analyzed 16 ED admission prediction models conducted in European hospitals. Unlike most studies in the review, our project expands upon these works by integrating both supervised and unsupervised learning, introducing patient clustering to optimize ED resource allocation.

The **supervised model** uses routinely available demographics, triage information, and clinical indicators, to predict a patient's likelihood of admission. This addresses the core challenge of managing patient flow more effectively, reducing ED crowding by identifying which patients are likely to need inpatient care. The main finding from the supervised learning experiments was that simpler, interpretable models, particularly Logistic Regression, significantly outperformed more complex approaches like an XGBoost, a Random Forest, or an Ensemble for predicting patient dispositions. Despite initial promise, complex embedding-based models experienced significant overfitting, especially during final validation. This outcome reinforces the importance of interpretability and simplicity in clinical decision-making contexts. Directly incorporating carefully selected clinical complexity features into a straightforward logistic regression model is effective and practical for predicting patient disposition.

This study also analyses patient data to identify the underlying distribution of data that could explain patterns and groupings of data. The **unsupervised learning** method revealed that age and cardiovascular metrics were the most prominent features explaining underlying data distributions. This study combines multiple preprocessing strategies (robust and aggressive transformations) with complementary clustering techniques (UMAP, GMM) to identify meaningful patient subgroups. Dimensionality reduction techniques (PCA) transform complex clinical data into interpretable visualizations, while Random Forest feature importance analysis characterizes the resulting patient clusters.

This project seeks to enhance patient outcomes and hospital efficiency by improving ED admission predictions and identifying natural patient clusters. Sush data-driven insights help healthcare providers understand resource usage, anticipate potential bottlenecks, and create targeted interventions for different patient groups. As a result, patient flow management becomes smoother, reducing critical care delays, and ultimately offering more personalized, efficient emergency care.

Related Work

Previous studies have investigated various approaches to predicting ED admission, employing statistical and machine learning techniques to enhance decision-making in triage. Kraaijvanger et al. (2018) developed a

logistic regression model using data from all ED patients in a community hospital in the Netherlands, identifying age, triage category, arrival mode, and main symptoms as the strongest predictors of admission. Their model performed well in community hospitals (AUC = 0.87) but had lower predictive accuracy in academic hospitals (AUC = 0.76), indicating potential variability in model generalizability. Similarly, Lucke et al. (2018) examined age as an effect modifier, showing that the strongest predictors of admission differed between younger (<70 years) and older (≥70 years) patients. Also, the AUC decreased from 0.86 to 0.77 for older patients, highlighting age-related differences in admission patterns. More recently, Feretzakis et al. (2024) applied automated machine learning with Gradient Boosting Machines to the MIMIC-IV-ED dataset, achieving an AUC of 0.83. Their study relied on triage data to forecast hospital admissions, focusing solely on supervised learning. The systematic review by Brink et al. showed that most models were based on logistic regression reporting AUC values ranging from 0.66 to 0.88.

Most models studied in the literature relied on logistic regression, limiting the exploration of more complex predictive relationships. Our project builds on this prior work by implementing several model families—Logistic Regression, Random Forest, and XGBoost—and integrating supervised and unsupervised learning. This dual approach aims to predict admission disposition and cluster patients based, offering insights into patient profiles. Furthermore, our study goes beyond previous work by accounting for and categorizing the chief complaint feature, a non-standardized free-text field of the patient's reported reason for presenting to the ED. Since chief complaints are subjective and vary in wording, we applied a complexity score to standardize and extract meaningful predictive patterns. Additionally, while the three models demonstrated strengths in predicting specific dispositions, the study also explored whether an Ensemble exhibits higher predictive power. Our model aligned with previous studies by revealing that Logistic Regression (AUC 0.80) outperformed more complex approaches (Ensemble AUC of 0.74).

Despite extensive past research, no predictive model has been widely implemented in real-world ED workflows. So, our study contributes to the literature by further validating existing findings to speed up the adoption of reliable, typical, standard results across studies. At the same time, identify areas that require further investigation for findings that diverge among studies. For example, multiple studies consistently identify age and mode of arrival as strong predictors of admission. Our unsupervised clustering models showed that the patient clusters may be best described using age and cardiovascular features. Additionally, our model AUC falls within the range of 0.76 and 0.87, as shown in earlier projects. Given that many studies were conducted in Europe, our research—leveraging the MIMIC-IV dataset from the U.S.—also assesses whether similar findings are generalizable across different healthcare systems.

Data Source

Background

Developed by the MIT Laboratory for Computational Physiology, Medical Information Mart for Intensive Care (MIMIC-IV) is a comprehensive ED admissions database at the Beth Israel Deaconess Medical Center from 2011 to 2019. MIMIC-IV is a comprehensive relational database encompassing patient, administrative, clinical, and laboratory data linked across multiple tables. Thanks to its large scale and rich detail, it has become a cornerstone for educational endeavors, appearing frequently in recent academic studies focused on clinical data science, health informatics, and related fields.

Within this broader framework, MIMIC-IV-ED focuses on emergency department admissions and provides detailed patient information—including vital signs, triage assessments, medication reconciliation, medication administration, and discharge diagnoses. In this study, we concentrate on MIMIC-IV-ED data exclusively, except for age, which we derive from related databases.

All data comply with HIPAA Safe Harbor provisions, facilitating diverse education and research studies. While MIMIC-IV-ED requires the completion of privacy training for full access, this training is available to qualified individuals for research purposes. In addition, a publicly accessible demo version of the database allows for preliminary demonstrations and exploration of the data.

Data Source and Access

In this study, all models were trained on the full MIMIC-IV-ED dataset, which the authors accessed through an authorized API hosted by Physionet. A smaller demo dataset is publicly available in our Git repository for those

who want to explore the database structure. While limited in scope, the demo dataset allows for preliminary demonstrations before proceeding to the full dataset under standard permission requirements[1].[1] Because the data collection period has already concluded, no special retrieval procedures were needed beyond authorized access, and any further data preprocessing is detailed in a subsequent section.

Technical Specifications of the Data

Database Structure: The database is organized in a Star Schema, with edstays as a central table. stay_id serving as the primary key.

Size: ~425,000 unique records
Important Keys:

1. Stay_id**:** Primary key
2. Subject_id: Foreign key to multiple tables
3. hadm_id: Foreign key to hospital admissions

Tables:

1. edstays: One row per stay (demographics, admission/discharge timestamps, disposition)
2. diagnosis: ICD-coded ED diagnoses per stay.
3. medrecon: Pre-ED admission medications (reconciliation records).
4. pyxis: Medications dispensed via an automated dispensing system (BD Pyxis).
5. triage: Initial triage vital signs, patient-reported pain, and acuity level.
6. vitalsign: Aperiodic vital signs recorded during the ED stay.

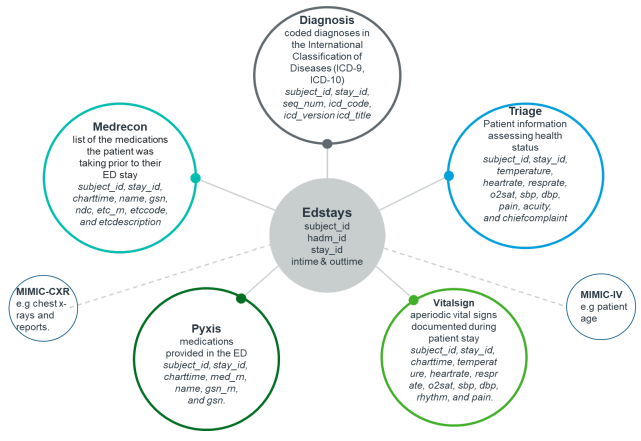


Figure 1. The diagram shows edstays as the central table connecting patient visits to related database tables and external sources. Each table includes key attributes and identifiers, offering a structured overview of the dataset and its connections.

Feature	Mean +/- std
Gender	M: 46% , F: 54%
Age	50 ± 20
Distaloic Blood Pressire (mmHg)	134 ± 22
Systolic Blood Pressure (mmHg)	77 ± 15
Heart Rate (min^{-1})	85 ± 17
Respiratory Rate (min^{-1})	17 ± 2
Oxygen Saturation (%)	98 ± 17
Temperature (°F)	98.0 ± 2
Pain (1-10)	4 ± 4
Total Rows	425,087
Rows Missing Some Values (%)	44,934 (10.47%)

[1] https://physionet.org/content/mimic-iv-ed-demo/2.2/

Data Preprocessing

Summary of Data Processing Strategy

Our study involves two distinct machine learning tasks—supervised learning and clustering—utilizing the same dataset but requiring different data preprocessing workflows. While general quality control steps, such as verifying feature names, data types, relationships across tables, and overall missingness, were applied to both tasks, clustering models necessitated additional constraints.

Unlike supervised learning, which relies on labeled outcomes and can tolerate some noise, clustering algorithms infer structure directly from the data, making them particularly sensitive to outliers, missing values, and measurement errors. We performed a detailed missingness analysis to ensure the integrity of clustering results, revealing non-random and random missing patterns. Clusters exhibiting a high density of missing values were invalidated to prevent biased group formation. Additionally, domain knowledge was used to verify that the values were physiologically plausible.

Missing Analysis

Our systematic analysis also revealed apparent differences in data quality associated with patient encounter types. Specifically, standard clinical pathways such as walk-in or ambulance arrivals, and typical dispositions like discharge to home or hospital admission, exhibited relatively low rates of missing data (2–11%). In contrast, atypical encounters—including helicopter transports or cases where patients expired in the emergency department—demonstrated significantly higher rates of missing values (40–90%). This discrepancy likely reflects differences in clinical priorities and resource allocation during non-routine or critical scenarios. Although these atypical cases represented only a tiny fraction of our dataset, they contain valuable information relevant to predicting patient disposition.

To maintain model accuracy and data integrity, we categorized missing data into random and systematically missing values and values clearly entered incorrectly. Examples of physiologically impossible measurements included blood pressures recorded as 1090 mmHg or temperatures logged at 923°F. Recognizing these apparent errors, we marked data points as missing and employed targeted imputation methods. Given the extensive volume of clinical features and substantial missingness across variables, our strategy aimed to preserve the maximum amount of meaningful information. Imputation approaches prioritized robust statistics, with median values assigned to continuous vital signs, including heart rate (median 84 bpm), respiratory rate (18 breaths/min), oxygen saturation (99%), systolic (133 mmHg) and diastolic (77 mmHg) blood pressures, temperature (98°F), and age (50 years). Due to their discrete, bounded nature, pain scores were imputed using the mean value (4.4).

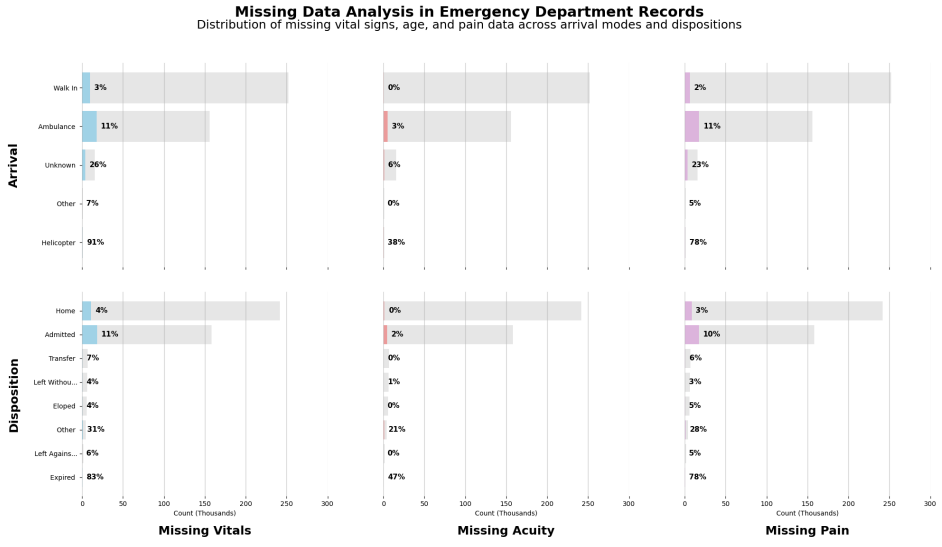


Figure 2. Missing data analysis

Patient disposition categories were standardized to streamline analysis, merging closely related outcomes: categories such as "LEFT AGAINST MEDICAL ADVICE," "LEFT WITHOUT BEING SEEN," and "ELOPED" were consolidated into "HOME." At the same time, cases labeled "TRANSFER" were combined into "ADMITTED." Entries categorized as "OTHER" were removed entirely due to ambiguity. For the unsupervised learning component of the project, the subclasses with high missingness were dropped from the analysis.

Critical fields (anchor_age, disposition, chief complaint) were carefully screened, and records missing essential information were excluded outright to ensure data quality. Furthermore, numeric fields representing vital signs underwent rigorous validation against established clinical norms, resulting in the identification and removal of physiologically implausible measurements, which were subsequently imputed. While introducing selection bias by focusing primarily on standard emergency department encounters, this meticulous approach balanced analytical rigor with practical constraints, ensuring our clustering models were less susceptible to biases stemming from systematic errors or extensive missingness.

For further details regarding missingness analysis refer to appendix II.

Validation Strategies

Our missingness analysis revealed that most missing data was concentrated in non-standard ED cases, particularly among patients arriving by helicopter or those whose workup was interrupted (e.g., patients who passed away in the ED). However, we opted to include these atypical encounters in our dataset for supervised learning analyses, as they retained valuable predictive information and contributed meaningful variation, simplifying our analytical approach without negatively impacting model performance. Conversely, as previously discussed, we excluded these non-standard cases for unsupervised analyses due to their elevated levels of missingness and inherent differences in clinical priorities to preserve the robustness and interpretability of clustering results. Additionally, we adopted physiological validation ranges based on evidence-based medicine and clinical practice experience.

- **Physiological Validity Checks:** Values outside biological plausibility were set to NaN:z

Vital Sign	**Valid Range**
Heart Rate	0–300 bpm
Respiratory Rate	0–99 breaths/min
O_2 Saturation	0–100%
Systolic BP (SBP)	0–400 mmHg
Diastolic BP (DBP)	0–350 mmHg
Temperature	30–115°F

Feature Engineering

General Strategy

We followed a structured and clinically informed strategy to convert raw input data into meaningful features suitable for supervised and unsupervised machine learning methods. Initially, we performed rigorous data validation, filtering physiologically impossible values in vital signs and converting these to missing values. Next, we systematically addressed missingness through robust imputation techniques, employing median values for continuous measures and mean values for discrete pain scores. We also standardized patient disposition categories, merging related outcomes to simplify analysis, and excluded records lacking essential information like age, disposition, or chief complaint.

Feature engineering centered around transforming raw clinical data into actionable, interpretable features aligned with clinical practice. Vital signs including heart rate, respiratory rate, oxygen saturation, systolic and diastolic blood pressures, temperature, pain scores, and acuity ratings were categorized into clinically meaningful groups (e.g., tachycardic, febrile, severe pain, acuity levels). We calculated derived physiological

indicators, such as shock index, mean arterial pressure (MAP), and SIRS criteria, to reflect underlying patient physiology and potential clinical deterioration. Temporal factors were incorporated by indicating whether patient arrivals occurred during daytime shifts (7 AM–7 PM), and patient ages were categorized to represent distinct demographic segments relevant in emergency care.

We quantified the complexity of chief complaints using natural language processing (NLP) techniques to harness predictive insights from unstructured text. Metrics such as text entropy, lexical complexity, part-of-speech complexity, medical entity frequency, and simple counts of words and characters provided additional predictive information. Analysis revealed a nonlinear relationship between text complexity and predictive accuracy, underscoring the importance of balancing information richness and clarity in clinical documentation.

To maximize predictive performance, we included standard and atypical ED encounters for supervised learning models. Conversely, unsupervised analyses excluded highly atypical cases due to their inherent data incompleteness and potential to distort clustering outcomes. This strategic approach balanced comprehensive feature representation with analytical rigor, ensuring robust model performance and clinical applicability.

Some Representative Features:

- **Clinical Metrics:** Derived cardiovascular and metabolic markers:
  - **Pulse Pressure:** SBP - DBP
  - **Shock Index:** Heart rate / SBP (marker of cardiovascular stability)
  - **SIRS Criteria:** Derived from temperature, heart rate, and respiratory rate
- **Vital Sign Categories:** Captured physiological deviations:
  - **Binary flags:** Indicators for abnormal vitals (is_tachycardic, is_hypoxic, is_febrile)
  - **Categorical groupings:** Ranges for heart rate, systolic blood pressure, and pain level
  - **Medically defined thresholds:** (HR > 100 = tachycardic)
- **Demographic Features:**
  - **Age groups:** Child (<18), young adult, adult, middle-aged, senior (>75)
  - **Acuity grouping:** Stratified by triage level
- **Temporal Features:**
  - **Daytime indicator:** Binary flag for 7 AM – 7 PM for dayshift

In supervised learning, features were used as predictors for our binary classification tasks. In unsupervised learning, the same features were used to define cluster characteristics and capture the data's high-dimensional structure in interpretable terms. Using the same feature set for both studies allows for better alignment between classification and clustering methods, and improves the interpretability of results.

For full details on features, see appendix 1.

PART A. Supervised Learning

Methods description

In this project, we employed a structured supervised learning workflow to predict patient dispositions (HOME vs ADMITTED) from emergency department presentations just after patient triage. Three distinct model families were selected to explore underlying mechanisms and enhance predictive robustness: Logistic Regression, XGBoost, and Random Forest.

Logistic Regression was chosen for its interpretability and probabilistic outputs, which are advantageous in clinical decision-making. It provided probability estimates for patient disposition, and L2 regularization was used to mitigate overfitting. We used an extensive list of native features in the dataset and multiple calculated and engineered features.

XGBoost, a gradient-boosting method, was selected for its superior performance with tabular data and its effective handling of complex, non-linear relationships. Given our interest in embedding models, the inherently

high-dimensional nature of embeddings made XGBoost particularly suitable for exploring their predictive value for patient disposition. Additionally, XGBoost provides built-in feature importance metrics, enhancing interpretability and facilitating deeper exploration through techniques such as principal component analysis.

Random Forest was used due to its ensemble approach using decision trees, robust handling of numerical and categorical variables, resilience to outliers, and ability to manage missing values effectively. Although our data preparation mostly mitigated any need for significant missing data robustness. Further, a random forest offers valuable insights into feature importance through multiple measures, allowing exploration of various complexity features and their significance, introducing a new clinical analytical concept in patient triage.

For feature representation, two primary approaches were integrated. Transformer-based embeddings for the free-text chief complaints were generated using the BAAI/bge-m3 model. After extensive experimentation with Principal Component Analysis (PCA), dimensionality was reduced to 80 components, preserving about 90% of the variance and effectively capturing most semantic nuances in the embeddings. Clinical and derived numerical features formed a structured feature set, including vital signs, demographic data, engineered indices such as shock index and SIRS criteria, and NLP-derived complexity scores (entropy, lexical complexity, medical entity counts).

Hyperparameter tuning was rigorously performed using 5-fold cross-validation for each model. Logistic Regression underwent grid search optimization for regularization strength, penalty type (L1, L2), and solver methods. XGBoost hyperparameters like learning rate, max depth, min child weight, number of estimators, subsample, and column sample ratios were optimized. Random Forest optimization targeted the number of estimators, max depth, min samples split, and min samples per leaf.

Results

Supervised evaluation leveraged metrics essential for clinical relevance and model robustness: accuracy for general correctness, weighted F1-score to address class imbalance, ROC-AUC to assess discrimination capability, and precision and recall to focus on the clinical implications of prediction errors. The best results demonstrated that Logistic Regression outperformed other models with an accuracy of 0.7334, F1 score of 0.7910, and ROC-AUC of 0.7970, while Random Forest significantly underperformed despite its moderate complexity.

Feature importance and sensitivity analyses on XGBoost identified embedding components, vital signs (especially heart rate and blood pressure), age, acuity scores, and NLP-derived complexity features as key contributors. Ablation studies demonstrated significant performance degradation when embeddings were removed (~15% reduction), with vital signs contributing around 20% and NLP complexity features adding about 5%. Sensitivity analysis showed a performance plateau after 80 PCA components, indicating an optimal balance between complexity and performance. The model also exhibited robustness across learning rates (0.01-0.1), though lower rates required more iterations for comparable results. However, despite promising performance during training and testing phases, the XGBoost model failed dramatically during final validation, suggesting overfitting and warranting further experimentation. Consequently, we transitioned away from XGBoost and Random Forest models, adopting a refined Logistic Regression model that directly utilized complexity features instead of indirectly through Random Forest.

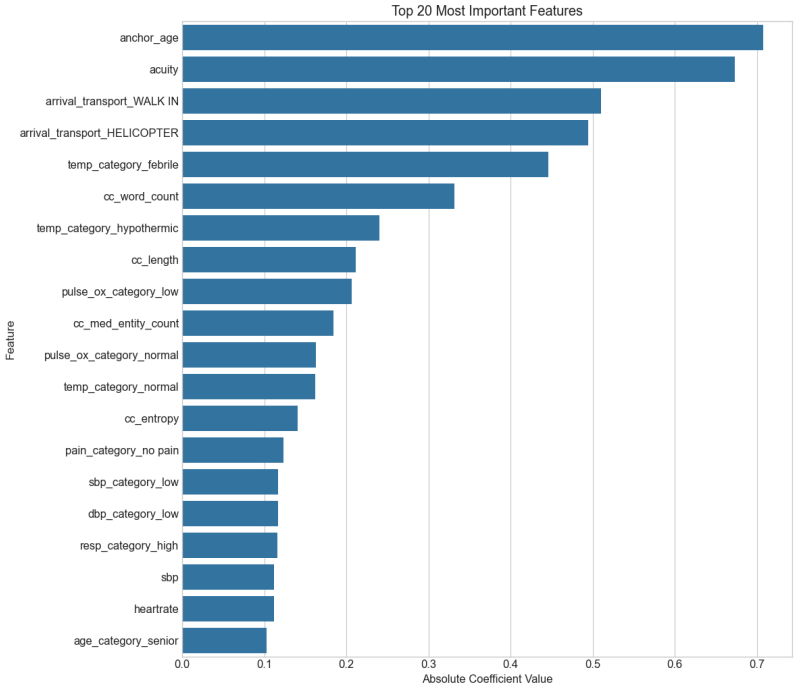


Figure 3. Feature importance

The logistic regression experiments comprehensively evaluated model performance using feature importance analysis, ablation studies, sensitivity analysis, and various sampling strategies to address class imbalance. The feature importance analysis highlighted critical predictors like patient age, fever status, acuity level, and systolic blood pressure, guiding subsequent ablation studies that revealed some features, notably cc_med_entity_count, anchor_age, and acuity, negatively impacted model performance. The sensitivity analysis tested multiple logistic regression parameters (C, penalty type, solver, class weight), demonstrating that the best performance resulted from minimal regularization (C=100) with an l2 penalty and the liblinear solver. Sampling experiments (SMOTE, random over- and under-sampling) showed minimal performance differences, suggesting the model inherently handled mild class imbalance effectively. Overall, these analyses guided refined recommendations for retraining, optimizing the model by removing detrimental features, tuning regularization, and confirming no additional sampling was necessary.

Table 1. Supervised models evaluation

Metric	XGBoost	Random Forest	Logistic Regression	Ensemble
Accuracy	0.6068	0.6578	**0.7318**	0.6788
F1 Score	0.5965	0.6221	**0.7266**	0.6337
Precision	0.5944	0.6531	**0.7283**	0.6976
Recall	0.6068	0.6578	**0.8324**	0.6788
ROC-AUC Score	0.6091	0.6840	**0.7971**	0.7433

Best performing model: Logistic Regression (highest scores across all metrics).

Failure analysis

Failure analysis indicates misclassifications by the model: false positives involve younger patients (ages 20–33) with moderate conditions (e.g., minor trauma, nausea/vomiting), incorrectly predicted as admitted due to disproportionate influence from age and arrival method. Conversely, false negatives now include elderly patients (ages 59–79) presenting with

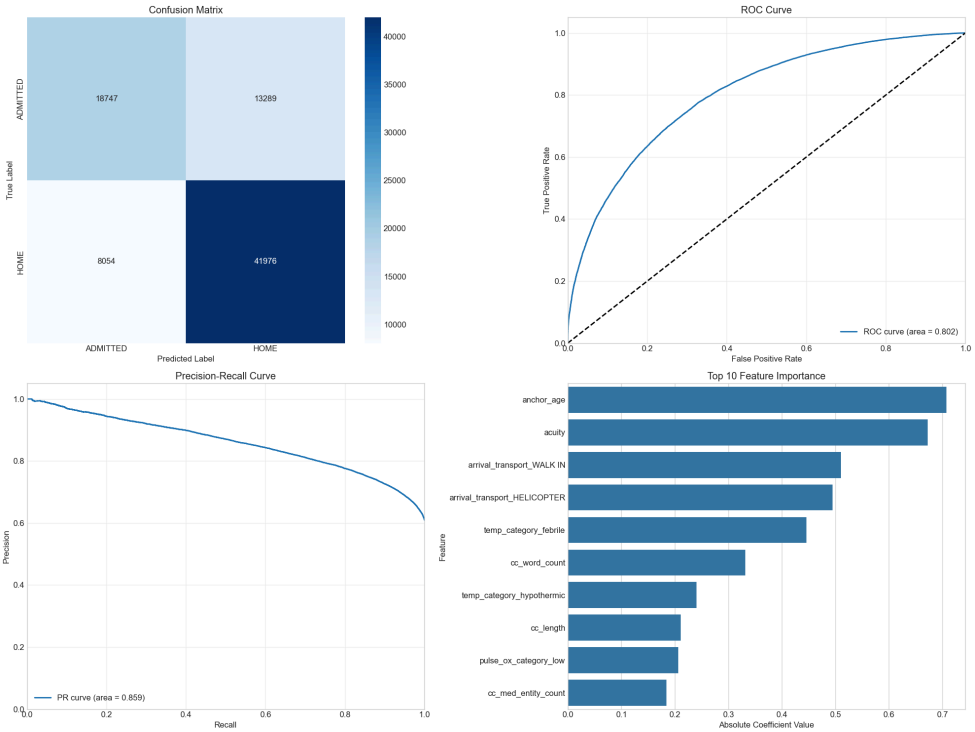


Figure 4. Logistic regression model evaluation

clearly severe respiratory distress or fever, and significantly abnormal vital signs (high heart rate, respiratory rate, low oxygen saturation), incorrectly classified as discharged home. These discrepancies highlight issues in the model's reliance on generalized age categories and vital sign features. Recommendations remain critical, that this is just a decision support for initial resource allocation and not to make judgements about which patients should be discharged home or admitted. Additional clinical workups are still indicated and integrating additional emergency department workups into their disposition decisions is still essential. However, this can still be an important tool in the repertoire of emergency medicine providers.

Part B. Unsupervised Learning

General Strategy

This study investigated patient clustering in emergency department data through a dual preprocessing approach: robust and aggressive transformations (including Box-Cox and strict outlier removal). We also used two feature sets for each model, one with all valid numerical features, and one focused feature set with only cardiovascular features. Initially, UMAP dimensionality reduction provided visual insights into natural data groupings. Gaussian Mixture Models (GMM) were then systematically applied to both transformation types, with parameters optimized through grid search to identify optimal cluster configurations. Clusters were subsequently labeled using K-means clustering for preliminary pattern analysis.

To refine the approach, Random Forest feature importance analysis identified the most discriminative variables from the GMM-labeled data. A focused model was then developed using only these top descriptive features, resulting in more interpretable clusters. The final clustered data was projected into PCA space for enhanced visualization, allowing for more precise examination of cluster characteristics and patient groupings. This progressive refinement strategy—from comprehensive analysis to focused modeling—enabled the identification of clinically meaningful patient segments with reduced feature complexity.

Data Transformation

Unsupervised clustering relies entirely on intrinsic data patterns, making transformation choices crucial. Different transformations can significantly alter the discovered structures since clustering algorithms group data points based on relative distances. To address this, we deployed two transformation strategies to capture complementary insights into the data’s geometry and ensure robust clustering outcomes.

We used robust transformations to preserve the natural distribution of features while minimizing outlier influence, ensuring that clusters reflected clinical realities rather than statistical artifacts. Simultaneously, we deployed an aggressive transformation technique. In this study, aggressive transformation refers to tailored transformations applied to different features based on their distributions.

Robust transformation preserves the natural distribution of features while minimizing the influence of outliers, ensuring that relative distances between typical data points remain intact. This approach is particularly practical in maintaining local structures and micro-patterns, allowing clusters to emerge based on clinical realities rather than statistical artifacts. In contrast, aggressive transformation focuses on fully normalizing feature distributions, standardizing scales to create more balanced feature importance. By reducing the impact of skewed distributions on distance calculations, it helps uncover broader structural patterns and identify boundaries that sparse or irregular distributions may otherwise obscure.

Robust Transformation

This approach utilized quantile-based scaling designed to accommodate outliers while reducing their influence on the model.

- **Transformation techniques:** RobustScaler with interquartile range (25th-75th percentile)
- **Coverage:** Applied to 13 numerical clinical features

- **Outlier handling:** Preserved but with reduced statistical influence
- **Special handling:** Random jitter was added to discrete variables to avoid arbitrary cluster formations due to low record precision.

Aggressive Transformation

This approach employed distribution-specific transformations with explicit outlier removal to maximize feature normality.

- **Transformation techniques:**
  - Standard scaling for approximately regular features
  - Box-Cox power transformations for skewed distributions
  - Inversion techniques for bounded metrics (e.g., oxygen saturation)
  - Trigonometric encoding for temporal features
- **Coverage:** Applied to 13 numerical clinical variables plus four temporal features
- **Outlier handling:** Outliers identified and removed using 1.5×IQR Tukey method
- **Special handling:** Random jitter was added to discrete variables to avoid arbitrary cluster formations due to low record precision.
- **Validation:** Shapiro-Wilk normality tests performed on transformed features

For further details regarding transformation techniques please refer to Appendix II.

UMAP Exploration

A comprehensive grid search across UMAP hyperparameters was performed, testing combinations of neighborhood sizes and minimum distances for both transformations. Features were scaled using MinMaxScaler to the range (-1, 1) to ensure equal contribution of all variables to the embedding space. Based on visual assessment of the parameter grid, optimal UMAP parameters were selected to generate the final two-dimensional representation that best preserved the underlying data structure.

While there are some variations in density and subtle patterns, there are no obvious discrete groupings, and none of the parameter combinations show clear, well-separated clusters (Figure 5-6). Even with the focused cardiovascular feature set, no precise, distinct clusters naturally emerge across the different parameter combinations. The data forms a generally continuous, uniform distribution across the embedding space.

This observation is essential and could have several explanations. The clinical data might represent a relatively homogeneous continuous spectrum of patients rather than discrete, separable patient types. The selected features might not capture the distinctions separating patients into clinically meaningful groups, and the relationships in the data might be too complex for UMAP to extract as clear visual clusters in a 2D space. While UMAP preserves non-linear relationships, the complexity of interactions between variables may create a manifold structure that doesn't naturally separate in two dimensions.

The lack of natural clustering suggests that any subsequent k-means clustering imposes artificial boundaries on a continuous distribution. Alternative approaches like Gaussian Mixture models might be more appropriate than clustering for this dataset.

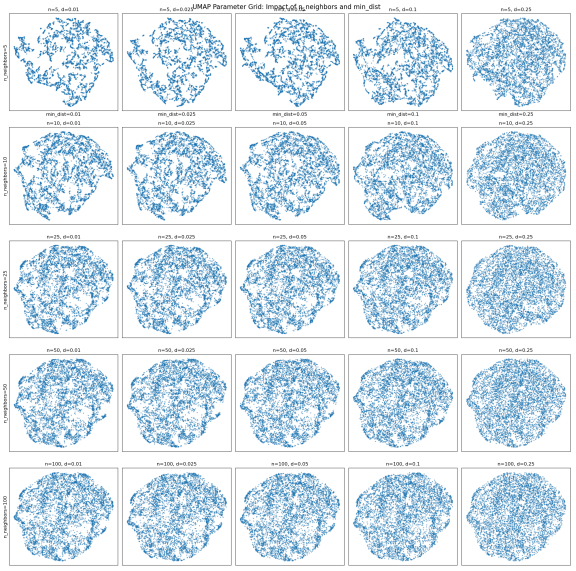


Figure5. UMAP Aggressive Transformations

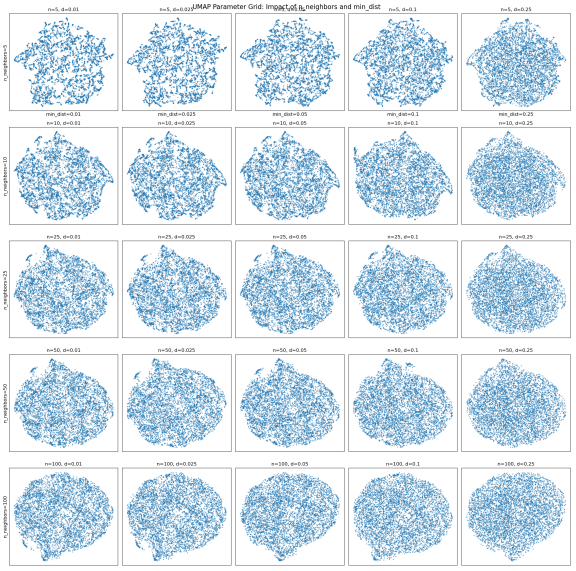


Figure 6. UMAP Robust Transformations

Gaussian Mixture Model

The study implemented a sequential approach to identify natural patient subgroups using Gaussian Mixture Models through structured data transformations and feature sets comparisons. First, the two data preprocessing approaches, aggressive and robust transformation, were evaluated. Both transformations were assessed using GMM models with components ranging from 2-15 clusters, full covariance matrices and consistent hyperparameters. Models were systematically compared using information criteria (BIC/AIC) to balance fit and complexity, and silhouette scores were used to assess cluster cohesion.

The superior approach was identified after comparing model performance metrics across both transformations. Using the selected model and transformation method, two feature sets were compared: a complete feature set including all available clinical parameters, and a focused cardiovascular feature set.

During earlier iterations, we observed artifactual separation of clusters driven predominantly by systolic and diastolic blood pressure measurements. Consequently, we refined our focused cardiovascular feature set by removing SBP and DBP while retaining their derivative metrics. This methodological adjustment resulted in more clinically interpretable clusters that better represented overall cardiovascular status rather than being dominated by raw blood pressure values. This methodological adjustment resulted in more clinically interpretable clusters that better represented overall cardiovascular status rather than being dominated by raw blood pressure values.

Although cluster separation metrics remained suboptimal (Figure 7), this feature engineering approach yielded patient subgroups with more meaningful clinical distinctions, highlighting the importance of selecting physiologically relevant feature combinations over redundant raw measurements. This refinement demonstrates how feature selection in unsupervised learning can significantly impact the clinical relevance of the resulting patient subgroups, even when statistical separation remains challenging.

We comprehensively compared Gaussian Mixture Models by evaluating different transformation methods and feature sets, ranking them based on silhouette scores to assess cluster cohesion. The results indicated low silhouette scores, suggesting poor cluster separation across models. The top two models suggested that a

two-cluster solution was statistically more plausible. However, this simplistic clustering did not adequately capture the diversity of patient types.

Given the need for more nuanced patient subgrouping, we opted for the third-best performing model, which was trained on robustly transformed data using a focused cardiovascular feature set that excluded SBP and DBP. This choice was driven by the desire to balance statistical plausibility with clinical interpretability, leveraging derived metrics like MAP and Pulse Pressure to provide a more integrated view of cardiovascular health. This approach aimed to create clusters that, while not perfectly separated statistically, offered more meaningful clinical insights.

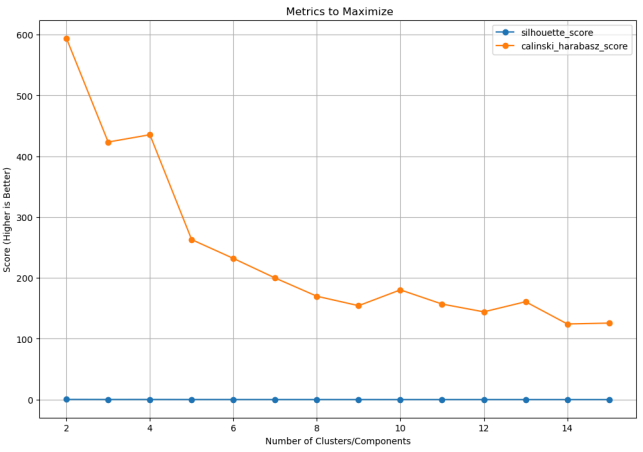


Figure 7.

Feature Importance Analysis

After selecting the optimal GM model, the patient clustering analysis reveals distinct patterns in subgroup distributions and clinical characteristics (Figure 8). Cluster 1 emerges as the largest group, encompassing a significant portion of the population, suggesting a prevalent clinical presentation or standard cardiovascular profile. Clusters 0 and 2 are of moderate size, indicating well-balanced distributions of patients with distinct but frequent physiological traits. In contrast, Cluster 3 is the smallest, potentially representing a unique or less common subgroup with specialized clinical profiles.

Examining feature characteristics, Cluster 0 exhibits moderate cardiovascular metrics, reflecting a balanced health status. Cluster 1 likely includes patients with common cardiovascular patterns, with average MAP and Pulse Pressure values indicative of typical clinical cases. Cluster 2, on the other hand, may be distinguished by specific cardiovascular traits, such as deviations in MAP, signifying distinct physiological conditions. Lastly, Cluster 3 stands out with extreme values in derived metrics, suggesting the presence of rare conditions or outliers within the dataset.

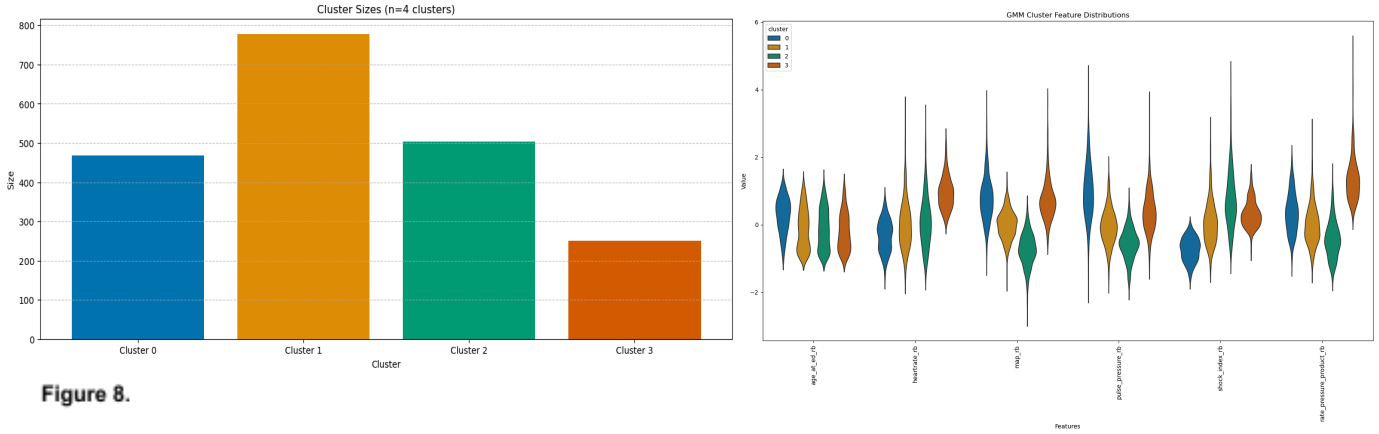


Figure 8.

Validation and Visualization

We projected the validation data into a 2D PCA space to analyze the clustering results to visualize how well the clusters are separated. By plotting the data points and centroids, we assessed the distinctiveness of each cluster and identified any potential overlap.

To further explore cluster characteristics, we utilized multiple visualization techniques. Violin plots were used to illustrate the distribution of feature values across clusters, highlighting variations in density and spread. Bar charts provided a comparative view of mean feature values for each cluster, while heatmaps captured the relative differences in feature means. Additionally, radar plots allowed us to compare multivariate feature profiles across clusters, offering a more holistic perspective. Finally, we generated a statistical summary for each feature within each cluster, providing a comprehensive numerical overview of the dataset's structure.

The PCA (Figure 9) visualization provides insights into the clustering structure and separation achieved by the model. The centroids, marked with red crosses, are positioned centrally within their respective clusters, indicating that they effectively represent the central tendency of each group. While the clusters are distinguishable, there is noticeable overlap, particularly between the purple and yellow clusters. This suggests that while the model has captured underlying patterns, some data points exhibit characteristics that blur the boundaries between clusters.

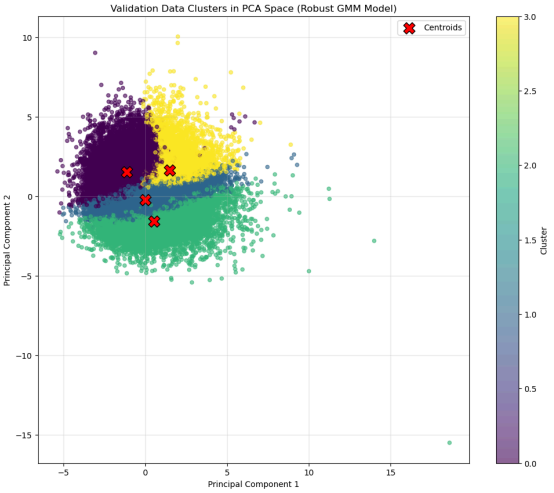


Figure 9.

Since the PCA projection reduces the data to two dimensions, some information from the original high-dimensional space is inevitably lost, potentially affecting how well the clusters are visually separated. We generated two additional graphs to better understand the cluster characteristics (Figure 10).

The violin plot provides a detailed visualization of how feature values are distributed across different clusters. This plot helps us see the spread and density of feature values, capturing the range, skewness, and potential outliers within each cluster. The bar chart presents the mean values of each feature for every cluster, offering a more straightforward way to compare clusters based on their central tendencies. This lets us quickly identify which features have higher or lower averages in different clusters, highlighting their key differences.

Using both graphs allows us to convey a more complete picture of the cluster characteristics. Together, they balance detail and clarity, ensuring a thorough understanding of the clustering structure.

The radar graph compares the feature profiles of different clusters simultaneously. The radar graph was produced by normalizing feature values for each cluster, plotting them on a circular grid with each axis representing a feature, and connecting the points to form a polygon for each cluster.

Each axis represents a feature, and the lines connecting the axes show the relative values of these features for each cluster.

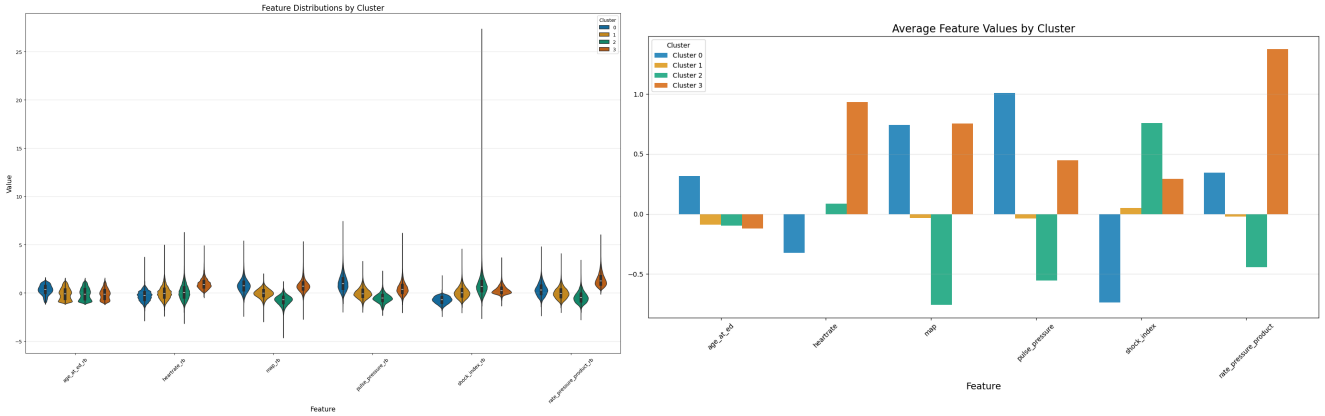


Figure 10.

Results and Discussion

The radar plot (Figure 12) allows for the visualization of multiple clusters on the same graph, making it easy to compare their feature profiles at a glance. By normalizing the feature values, the radar plot enables comparison across features that may initially be on different scales. This helps understand the relative importance or dominance of features within each cluster. The shape and area covered by each cluster's line can reveal patterns or unique characteristics, helping to identify which features are most distinctive for each cluster.

Cluster 0: Hypertensive Older Patients

Patients in Cluster 0 exhibit high mean arterial pressure (MAP) and pulse pressure (PP), relatively older age and low heart rate and shock index. Clinically, this suggests a group of older individuals with chronic or uncontrolled hypertension, which is often associated with conditions such as arterial stiffness, cardiovascular disease, and long-term hypertension-related complications. The combination of high MAP and PP indicates a significant cardiovascular burden, which requires careful management to prevent adverse outcomes.

Figure 11.

Cluster 1 & Cluster 2: Patients with High Shock Index and Elevated Heart Rate

Both Cluster 1 and Cluster 2 contain patients with high shock index and elevated heart rate, indicating a state of low blood perfusion—a key clinical concern in cases of hemorrhage, dehydration, or systemic infections.

- Cluster 1: Younger Patients with Normal Blood Pressure
  Patients in this group have a relatively normal MAP and PP, suggesting that despite high heart rate and shock index, their physiological compensatory mechanisms effectively maintain blood pressure. This could indicate an early-stage or well-compensated shock state, where the body can still sustain adequate circulation.
- Cluster 2: Older Patients with Low MAP and PP
  Unlike Cluster 1, Cluster 2 consists of older patients with low MAP and PP, suggesting that their physiological response is insufficient to sustain blood pressure. This could indicate a more advanced or severe shock stage where compensatory mechanisms fail, making these patients more vulnerable to hemodynamic instability.

Cluster 3: Younger Patients with Physiological Distress or Stimulant Influence

Patients in Cluster 3 are younger, with very high heart rates but relatively normal or high MAP and PP. This pattern suggests that these individuals might be experiencing physiological distress—such as pain, anxiety, or acute stress—or external influences like stimulant substances (e.g., caffeine, drugs) that elevate heart rate and blood pressure. Unlike Clusters 1 and 2, where shock-related conditions are suspected, Cluster 3's physiological profile does not indicate compromised perfusion but rather an increased sympathetic response.

Ethical Considerations

Using predictive machine learning models and patient clustering techniques in ED settings introduces several ethical considerations. While nuances between the supervised and unsupervised learning methods exist, they share the following ethical concerns: .

- ***Privacy concerns:*** arise with retrospective patient data since it limits patients' autonomy as they do not explicitly consent to using their records.

  Such concern is addressed partly by the MIMIC-IV, which consists of de-identified patient records and is HIPAA-compliant. Furthermore, our team ensured ethical data handling by completing the required training on human subjects research and adhering to the data use agreement (DUA). Additionally, the study maintained secure storage and access control, ensuring that only authorized team members worked with the data in compliance with the dataset's usage policies. Also, the study focuses on aggregated trends reinforcing privacy protection. Concurrently, the study provides a publicly available demo dataset in the Git repository, enabling others to explore the database structure to enhance learning experience.

- ***Bias and algorithmic unfairness***

  - **Disparities in collected data:** the data collected at ED departments may inherently include disparities since it is primarily sourced from patients with access to healthcare facilities. Consequently, it may inadvertently exclude disadvantaged groups, such as individuals with financial hardships or those with different healthcare-seeking behaviors. This selection bias could be accentuated in the unsupervised learning method as the study excluded atypical encounters and outliers due to the sensitivity of clustering to these cases.

  - **Geographic bias:** the choice of the MIMIC-IV dataset, sourced from a single hospital in Boston, U.S. as the primary focus for the project introduces geographic bias. This selection means that other countries and populations with diverse cultures and ethnicities are excluded from the study, restricting representativeness.

  Since Boston is a cosmopolitan city, it enhances fairness by gathering data from patients of different age groups, gender, and race. Furthermore, the study included the entire MIMIC-IV-ED dataset (>400,000 records) , This is evident from a broad age range (18–89, with older patients capped), a balanced gender split (46% male, 54% female), and a race distribution where 58% are White, 21% Black, 8% Other, 8% Hispanic, and 4% Asian. On the other hand, excluding atypical cases in clustering stemmed from prioritizing robustness and interpretability of results to prevent potential harm from inaccurate findings. We further enhanced fairness, by retaining such cases in supervised learning, maintaining their valuable predictive information. Furthermore, we supplemented our conclusions by referring to previous research from other countries to validate common findings.

- ***Dependence on AI in Clinical Decision-Making:*** While our model serves as a decision-support tool, there is a risk that clinicians may become over-reliant on AI predictions, potentially overriding clinical judgment.

  Our model is designed to complement, not replace, physician decision-making. As such, the model sought to incorporate the "chief complaint" feature via NLP techniques, ensuring that medical professionals' assessments remain central to the model's decision-making process.

While the study has taken measures to address ethical concerns, limitations may remain. To promote ethical accountability, all project work, including methodologies and findings,s are shared to inform learners about how datasets are employed and contribute to transparency.

Statement of Work

- **Sawsan Allam:** Health background insights, ethical considerations, qualitative dimensions, report writing and review.
- **Kasra Afzali:** GitHub setup, unsupervised learning code implementation, feature engineering, model development, report writing, and review.
- **Nicholas Abram:** Data evaluation, supervised learning code implementation, cleaning, feature engineering, model development, report writing, and review.

References

Asplin BR, Magid DJ, Rhodes KV, et al. A conceptual model of emergency department crowding. Ann Emerg Med 2003;42:173–80.

Brink, A., Alsma, J., van Attekum, L. A., Bramer, W. M., Zietse, R., Lingsma, H., & Schuit, S. C. (2022). Predicting inhospital admission at the emergency department: a systematic review. Emergency Medicine Journal, 39(3), 191-198.

Feretzakis, G., Sakagianni, A., Anastasiou, A., Kapogianni, I., Tsoni, R., Koufopoulou, C., ... & Verykios, V. S. (2024). Machine learning in medical triage: A predictive model for emergency department disposition. *Applied Sciences*, *14*(15), 6623.

Graham, B., Bond, R., Quinn, M., & Mulvenna, M. (2018). Using data mining to predict hospital admissions from the emergency department. *Ieee Access*, *6*, 10458-10469.

Kraaijvanger, N., Rijpsma, D., Roovers, L., van Leeuwen, H., Kaasjager, K., van den Brand, L., ... & Edwards, M. (2018). Development and validation of an admission prediction tool for emergency departments in the Netherlands. *Emergency Medicine Journal*, *35*(8), 464-470.

LaMantia, M. A., Platts-Mills, T. F., Biese, K., Khandelwal, C., Forbach, C., Cairns, C. B., ... & Kizer, J. S. (2010). Predicting hospital admission and returns to the emergency department for elderly patients. *Academic emergency medicine*, *17*(3), 252-259.

Lucke, J. A., de Gelder, J., Clarijs, F., Heringhaus, C., de Craen, A. J., Fogteloo, A. J., ... & Mooijaart, S. P. (2018). Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. *Emergency Medicine Journal*, *35*(1), 18-27.

Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., & Kinsman, L. (2018). Emergency department crowding: a systematic review of causes, consequences and solutions. PloS one, 13(8), e0203316.

Acknowledgements

Windsurf IDE - Agentic IDE providing intelligent code assistance and pair programming capabilities

Claude 3.5 Sonnet - Advanced AI model used for code generation, debugging, and technical documentation integrated into Windsurf IDE.

Grammarly - This writing assistant is a plugin for Google Docs and Microsoft Word for documentation proofreading, grammar checking, clarity, and readability.

Appendix 1. Features List

Data Description

Features Overview

The dataset is processed using a binary classification model with the following feature categories:

Target Variable

- `disposition`: Binary target variable for classification

Feature Types

1. **Numeric Features**
    - Processed using StandardScaler for normalization
    - Missing values are imputed with median values
    - Features are scaled to have zero mean and unit variance
2. **Categorical Features**
    - Processed using OneHotEncoder
    - Missing values are imputed with mode (most frequent value)
    - Handles unknown categories during prediction
    - Sparse encoding is disabled for better interpretability

Data Processing

- **Class Balance**: Dataset undergoes undersampling to achieve 1:1 class distribution
- **Cross-validation**: Uses k-fold cross-validation (default 5 folds) for model evaluation
- **Feature Engineering**:
    - Automatic feature preprocessing pipeline
    - Maintains feature order consistency between training and prediction

Model Configuration

- **Hyperparameter Optimization**:
    - Solver options: liblinear (l1/l2) and lbfgs (l2)
    - Regularization strength (C): [0.001, 0.01, 0.1, 1, 10]
    - Maximum iterations: [100, 500, 1000]

Model Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC
- Confusion Matrix

Feature Details

Numeric Features

Summary Statistics

Feature	Mean ± Std	Median [IQR]	Range
temperature	98.02 ± 2.11	98.00 [97.60-98.50]	33.70-111.40
heartrate	85.03 ± 17.33	84.00 [73.00-95.00]	1.00-256.00
resprate	17.57 ± 2.31	18.00 [16.00-18.00]	0.00-98.00
o2sat	98.40 ± 2.41	99.00 [98.00-100.00]	0.00-100.00
sbp	134.85 ± 22.02	133.00 [120.00-147.00]	1.00-312.00
dbp	77.45 ± 14.55	77.00 [68.00-86.00]	0.00-345.00
pain	4.36 ± 3.73	4.40 [0.00-8.00]	0.00-10.00
shock_index	0.65 ± 0.38	0.63 [0.53-0.74]	0.01-102.00
sirs	0.34 ± 0.48	0.00 [0.00-1.00]	0.00-1.00
anchor_age	50.31 ± 20.09	50.00 [32.00-66.00]	18.00-91.00
acuity	2.58 ± 0.82	3.00 [2.00-3.00]	-1.00-5.00

Feature	Mean ± Std	Median [IQR]	Range
cc_entropy	1.09 ± 0.73	1.00 [1.00-1.58]	0.00-3.42
cc_lexical_complexity	1.00 ± 0.03	1.00 [1.00-1.00]	0.25-1.00
cc_pos_complexity	1.21 ± 1.10	1.00 [0.00-2.00]	0.00-11.00
cc_med_entity_count	0.58 ± 0.65	0.00 [0.00-1.00]	0.00-5.00
cc_length	14.56 ± 8.06	13.00 [8.00-19.00]	1.00-136.00
cc_word_count	2.43 ± 1.24	2.00 [2.00-3.00]	1.00-16.00

Categorical Features

Vital Signs Categories

Feature	Categories	Description
hr_category	• normal • tachycardic • bradycardic	Heart rate classification
resp_category	• normal • high • low	Respiratory rate classification
pulse_ox_category	• normal	Oxygen saturation classification

Feature	Categories	Description
	• low	
`sbp_category`	• normal • high • low	Systolic blood pressure classification
`temp_category`	• normal • febrile • hypothermic	Temperature classification
`dbp_category`	• normal • high • low	Diastolic blood pressure classification
`pain_category`	• no pain • mild • moderate • severe	Pain level classification

Patient Demographics

Feature	Categories	Description
`age_category`	• young_adult • adult • middle_aged • senior	Age group classification

Feature	Categories	Description
`gender`	• F • M	Patient gender

Hospital Operations

Feature	Categories	Description
`day_shift`	• True • False	Indicates day vs night shift
`arrival_transport`	• AMBULANCE • WALK IN • HELICOPTER • OTHER • UNKNOWN	Mode of arrival transport

Data Quality Measures

- Comprehensive logging of missing value statistics
- Tracking of rows dropped due to missing critical data
- Validation of vital signs against physiological ranges
- Special handling for pain scores with unclear documentation

Key Strengths

1. Robust error handling and logging
2. Physiologically appropriate value ranges
3. Differentiated handling of critical vs. non-critical missing data
4. Transparent imputation strategies
5. Comprehensive documentation of data transformations

Detailed Feature Lists

Numeric Features

1. **Vital Signs**

- `temperature`: Patient's body temperature
- `heartrate`: Heart rate measurement
- `resprate`: Respiratory rate
- `o2sat`: Oxygen saturation level
- `sbp`: Systolic blood pressure
- `dbp`: Diastolic blood pressure
- `pain`: Pain level assessment

2. **Clinical Scores**

    - `shock_index`: Clinical measure of shock
    - `sirs`: Systemic Inflammatory Response Syndrome score
    - `acuity`: Patient acuity score

3. **Patient Demographics**

    - `anchor_age`: Patient age at time of visit

4. **Text Analysis Features**

    - `cc_entropy`: Entropy measure of chief complaint text
    - `cc_lexical_complexity`: Lexical complexity score of chief complaint
    - `cc_pos_complexity`: Part-of-speech complexity measure
    - `cc_med_entity_count`: Count of medical entities in chief complaint
    - `cc_length`: Character length of chief complaint
    - `cc_word_count`: Word count in chief complaint

Categorical Features

1. **Vital Sign Categories**

    - `hr_category`: Heart rate category
        - normal
        - tachycardic
        - bradycardic
    - `resp_category`: Respiratory rate category
        - normal
        - high
        - low
    - `pulse_ox_category`: Pulse oximetry category
        - normal
        - low
    - `sbp_category`: Systolic blood pressure category
        - low
        - normal
        - high
    - `temp_category`: Temperature category
        - normal
        - febrile
        - hypothermic
    - `dbp_category`: Diastolic blood pressure category
        - low
        - normal
        - high

- `pain_category`: Pain level category
    - severe
    - no pain
    - moderate
    - mild

2. **Temporal Features**

    - `day_shift`: Indicator for day shift vs night shift
        - True
        - False

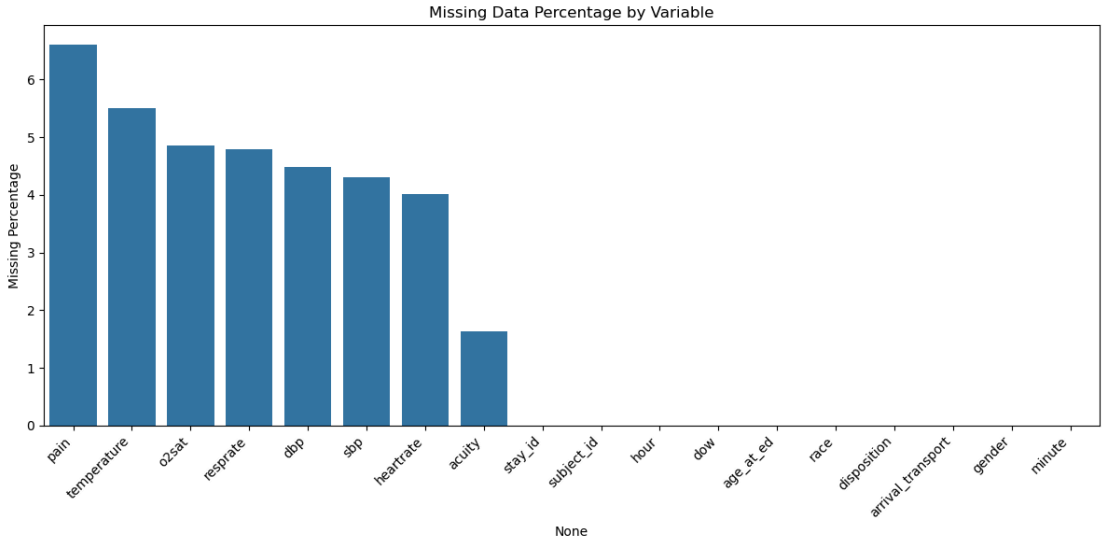
3. **Patient Demographics**

    - `age_category`: Age group category
        - adult
        - middle_aged
        - young_adult
        - senior
    - `gender`: Patient gender
        - F
        - M

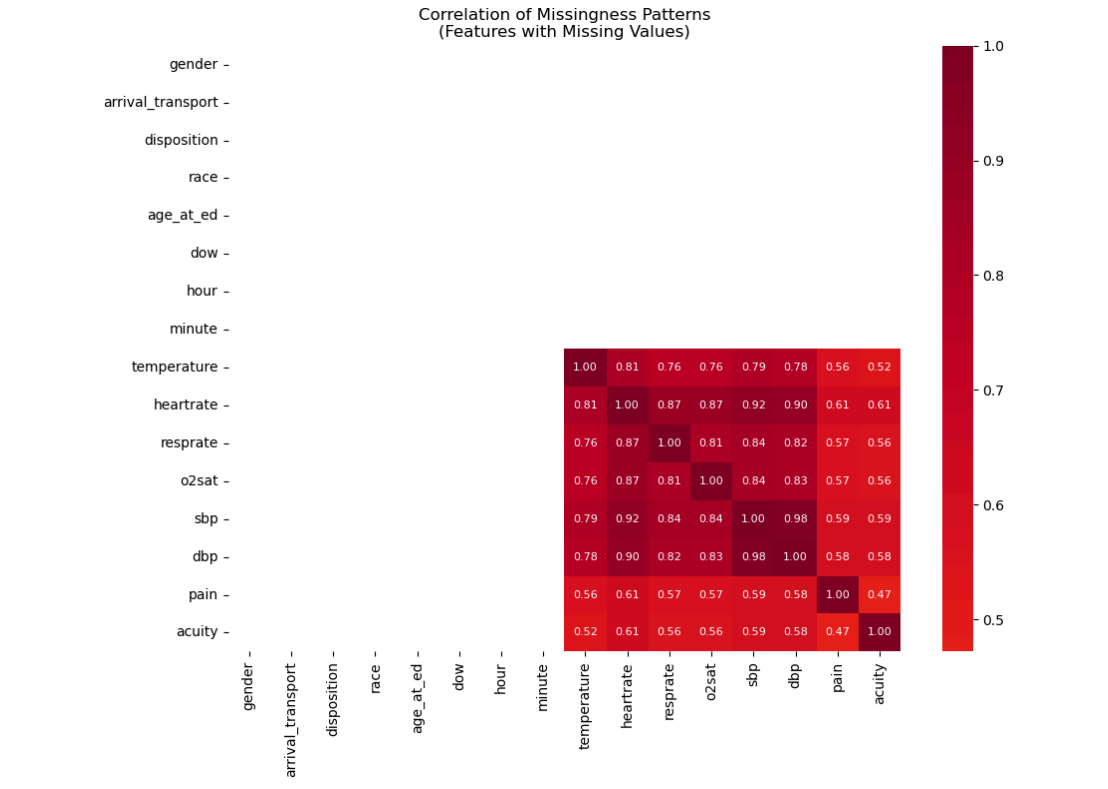
4. **Hospital Operations**

    - `arrival_transport`: Mode of arrival transport
        - AMBULANCE
        - WALK IN
        - UNKNOWN
        - OTHER
        - HELICOPTER

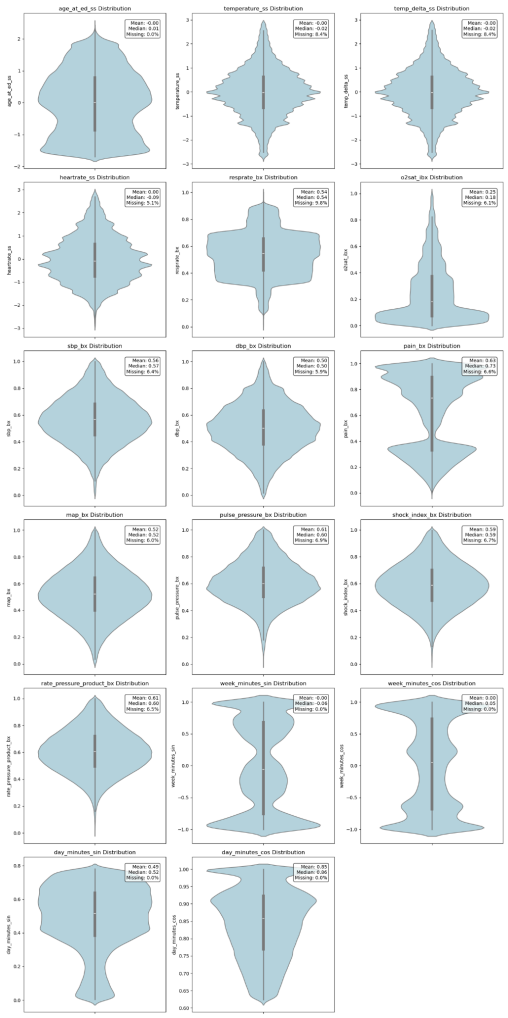
APPENDIX II.

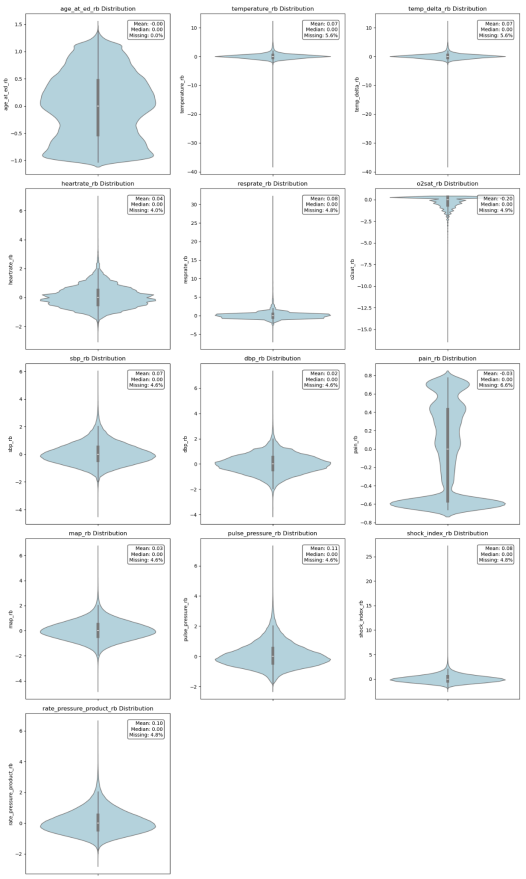


Missingness By Feature. The bar chart displays missing data percentages across variables in our emergency department dataset. Pain scores show the highest missingness at approximately 6.5%, followed by temperature at 5.5%



This heatmap illustrates the correlation of missingness patterns among features in our emergency department dataset. The visualization reveals strong relationships in how data is missing across vital sign measurements.



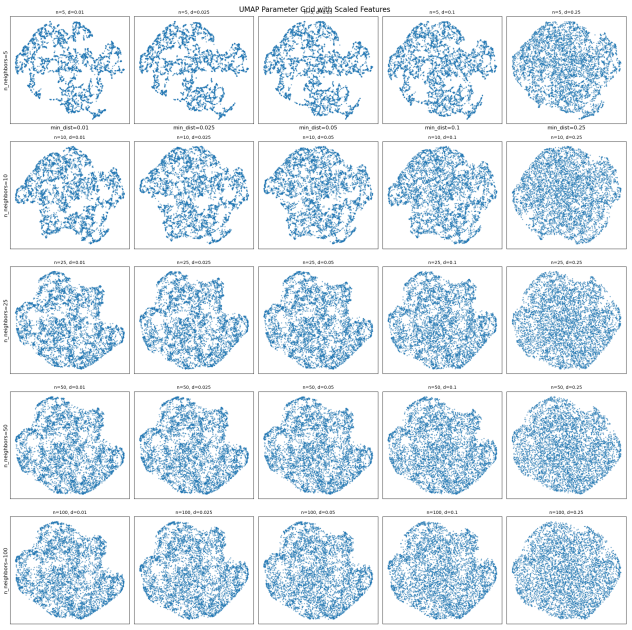


Value Distribution by Feature with Aggressive Transformation.

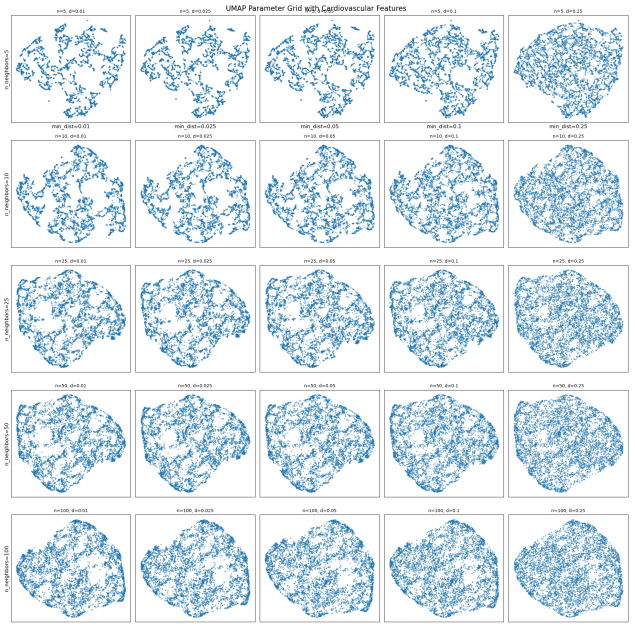
This visualization shows the results of aggressive data transformation techniques applied to our emergency department dataset for unsupervised learning. The violin plots display how each feature has been normalized, with most distributions now showing more standardized patterns.

Value Distribution by Feature with Robust Transformation

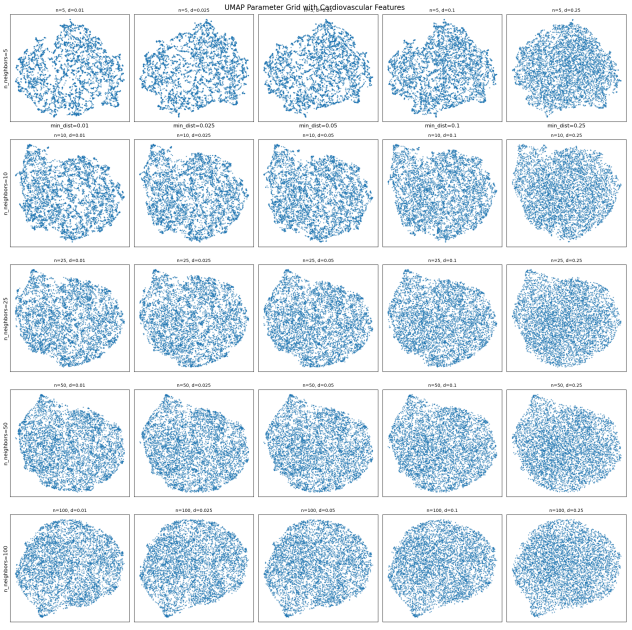
This visualization shows the results of aggressive data transformation techniques applied to our emergency department dataset for unsupervised learning. The violin plots display how each feature has been normalized, with most distributions now showing more standardized patterns. Compared to aggressive transformation, this transformation maintains more of the original data distribution. The robust transformation strikes a balance between normalization and preserving the underlying data structure, making it potentially more suitable for unsupervised learning models where maintaining natural clusters and relationships in the data is important.



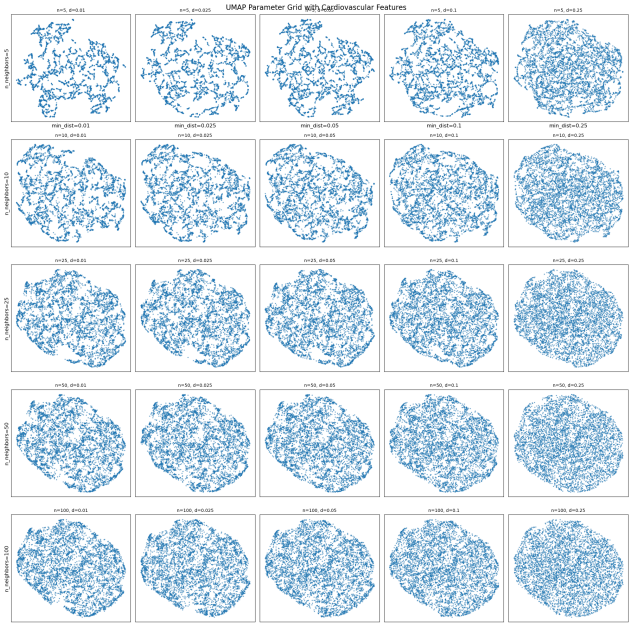
UMAP representation of aggressive transformation with complete feature set



UMAP representation of aggressive transformation with focused cardiovascular feature set



UMAP representation of robust transformation with complete feature set



UMAP representation of robust transformation with focused cardiovascular feature set

UMAP representation of the data using a combination of aggressive or robust transformations, with full feature or focused cardiovascular feature. The visual inspection of the UMAP graphs does not suggest any definitive clusters.