RESEARCH PAPER

Characterizing ARDS Subphenotypes via Autoencoder-Based Clustering

Dr.Nitha V R[1,2], Arun K[3], Neethu Kunjappan[2],Gayathri Ashok[2],Tiny Thampan[2], Alphonsa Sini P J[4]

[1]*Department of Computer Science, Assumption College (Autonomous), Changanassery, India,*

[2]*CVV Institute of Science and Technology, Chinmaya Vishwa Vidyapeeth, Ernakulam, India, Email: nitha.vr@cvv.ac.in*

[3]*Department of Computer Science, University of Kerala, Trivandrum, India, Email: arunk@keralauniversity.ac.in*

[4]*Department of Computer Science, Bharta Mata College, Thrikkakara, India, Email: alphonsa@bharatamatacollege.in*

ABSTRACT

Acute Respiratory Distress Syndrome exhibits substantial clinical heterogeneity, making patient stratification and mortality assessment challenging using conventional analytical approaches. Traditional clustering methods applied to high-dimensional clinical datasets often fail to capture nonlinear relationships among patient variables, limiting their clinical interpretability. This study proposes an autoencoder-based clustering framework that integrates nonlinear representation learning with K-means clustering to identify clinically meaningful subgroups of ARDS patients. The autoencoder learns a compact latent representation that preserves complex interactions among clinical features while reducing redundancy and noise. Clustering in this latent space yields improved subgroup separation compared with linear dimensionality-reduction techniques. The framework was evaluated on an ARDS dataset containing 1,000 patients with 21 clinical features. Experimental results demonstrated superior clustering performance, achieving a Silhouette Score of 0.3892 compared with 0.1006 and 0.0664 for baseline K-means and PCA-based clustering, respectively. The method also achieved improved Davies–Bouldin (0.7837) and Calinski–Harabasz (1361.57) indices, indicating more compact and well-separated clusters. The proposed approach consistently outperformed comparison methods across clustering quality metrics. The model identified five patient subgroups with clear mortality gradients ranging from 13.3% to 36.9%, facilitating intuitive risk stratification. These findings demonstrate that nonlinear representation learning enhances unsupervised clinical phenotyping and supports interpretable risk assessment in critical care settings. The proposed framework highlights the potential of deep learning–based clustering for improving personalized treatment strategies in ARDS management.

Keywords: Machine Learning, Predictive Modelling, Mortality Prediction, Acute Respiratory Distress Syndrome, Classification Algorithms, Clinical Decision Support Systems, Healthcare Data Analytics

How to cite this article: Dr.Nitha V R, Arun K, Neethu Kunjappan, Gayathri Ashok, Tiny Thampan, Alphonsa Sini P J (2026). Characterizing ARDS Subphenotypes via Autoencoder-Based Clustering. International Journal of Drug Delivery Technology; 405-414; DOI: 10.25258/ijddt.16.7s.42

INTRODUCTION

Polycystic Acute Respiratory Distress Syndrome (ARDS) is a very serious illness that can sometimes be fatal. It is marked by quick and widespread inflammation of the lungs. If not treated right away, this inflammation can quickly lead to respiratory failure, block the flow of oxygen, and make the lungs not work properly. This inflammatory reaction harms the alveolar capillary barrier, which causes fluid to build up in the alveoli and makes it harder for gases to move between them. This could cause hypoxaemia and respiratory failure. Patients with this medical condition often need mechanical ventilation and intensive care because of the dysfunction of lungs. The syndrome can develop from several underlying conditions, including sepsis, pneumonia, trauma, aspiration, and other serious illnesses.

This can lead to a variety of clinical signs and symptoms. Even though there are better ways to manage and support critical care, ARDS is still linked to higher rates of death and illness. It is very important to get a diagnosis and treatment right away. To help the patient get better, the main goals of treatment plans should be to make sure the patient gets enough oxygen, make the organs work better, and fix the problem at its source.

Age has a big impact on how likely it is that someone will get sick and how long they will live. Older people are more likely to get sick if they already have health problems like diabetes, heart disease, or lung problems that don't go away. Older people are more likely to die from ARDS than younger people because their lungs are less flexible and their immune systems are weaker. But there is still a potential of having major breathing issues, especially if this

lung condition is caused by an infection, an injury, or a systemic inflammatory illness. Age has a huge effect on how likely a patient is to get better and how fragile they are; therefore, it is vital to think about this thoroughly when making plans for how to treat different groups of patients [1].

Men are more likely than women to get this lung disorder, which may be because their immune systems work differently. Some of the differences in outcomes between men and women may be due to the anti-inflammatory effects of female estrogen. The main sign of ARDS is that breathing gets worse very quickly, usually within hours to days of the event that caused it. Patients frequently exhibit abnormal, accelerated respiratory rates, significant dyspnea, recruitment of accessory respiratory muscles, and symptoms that fail to ameliorate with supplemental oxygen. Vital signs often show low oxygen saturation (SpO2), fast breathing (tachypnea), and, depending on the cause, fever or low blood pressure. Severe hypoxaemia may cause anxiety or change how the brain works.

Arterial blood gas measures, white blood cell counts, platelet counts, and tests of kidney and liver function have a big effect on this disease. Patients with chronic lung illnesses have a restricted ability to tolerate high levels of inflammation. Heart failure or coronary artery disease makes it harder for oxygen to get to tissues, which raises the chance of organ failure. Individuals with weakened immune responses and prolonged recovery are at a higher risk of developing severe infections and sepsis, which are frequent triggers of ARDS [2]. These disorders raise the risk of multi-organ failure by affecting immune function, metabolic homeostasis, and the ability of body to get rid of toxins. Compromised immune systems make people more likely to get very sick and make it harder for them to get better. Chronic inflammation and lung injury make ARDS worse by lowering the breathing ability. Even though better supportive care and ventilatory strategies have led to fewer deaths, ARDS is still a major cause of illness and death around the world [3].

Acute Respiratory Distress Syndrome is considered a heterogeneous syndrome, characterised by highly variable patient outcomes that complicate stratification for prognosis and treatment planning. Conventional stratification techniques often inadequately represent the complex, non-linear interrelations among clinical characteristics that drive disease advancement. Linear methods can hide clinically important structures, so dimensionality reduction techniques are often used to work with clinical data that has a lot of dimensions. Consequently, there exists an urgent requirement for unsupervised learning methodologies capable of deriving low-dimensional representations from complex clinical data while preserving the critical structures relevant to patient outcomes.

The key contributions of this study are as follows:

- An autoencoder-driven clustering framework is presented for ARDS patient stratification, in which nonlinear dimensionality reduction is combined with K-Means clustering to enable the identification of clinically interpretable subgroups based on mortality risk.
- Comprehensive validation is performed using multiple clustering quality metrics, together with clinical interpretability analysis based on mortality rates across the identified subgroups.
- A distance-based risk scoring heuristic is proposed to demonstrate the potential utility of latent representations for mortality prediction.
- Five distinct ARDS patient subgroups are identified, each exhibiting a clear mortality gradient, enabling intuitive risk stratification into low-, intermediate-, and high-risk categories.
- The significance of non-linear representation learning in maintaining clinically pertinent structure within intricate medical data is empirically validated.

Clustering methods show that there is a lot of potential for unsupervised patient stratification in critical care research. Nonetheless, contemporary methodologies frequently fail to adequately handle high-dimensional clinical data and the intricate, non-linear interrelations among physiological attributes. In the subsequent section, pertinent research in clinical phenotyping is examined, the research gaps this study aims to fill are delineated, and the contributions to ARDS patient stratification are articulated.

ASSOCIATED LITERATURE

Machine learning is a very useful way to predict the survival of people with ARDS, a life-threatening condition marked by severe hypoxaemia, diffuse alveolar damage, and a lot of clinical variation. Even though ventilatory strategies, supportive therapies, and intensive care unit management have gotten better, ARDS still has a high death rate [4].

One of the biggest problems with treating ARDS is that patients show up with very different symptoms, causes, inflammatory responses, and organ dysfunction patterns. Because there are so many different types, it is harder to figure out who is at risk, and traditional scoring systems are not as good at predicting outcomes. In the past, prognostic evaluation in critically ill ARDS patients has relied on oxygenation-related indices and scoring systems, such as the Simplified Acute Physiology Score (SAPS) and the Sequential Organ Failure Assessment (SOFA). These tools offer significant clinical insights and are frequently employed for benchmarking and estimating overall

mortality [5]; however, their predictive efficacy at the individual patient level is still constrained. These models typically rely on a limited array of physiological variables gathered at a single time point and employ linear statistical methods, which fail to adequately represent the dynamic characteristics of critical illness, where swift changes in haemodynamic status, ventilatory parameters, and laboratory values can significantly impact prognosis within hours.

Machine learning models can use a wide range of data, such as demographics, lab results, vital signs, comorbidities, and physiological data that changes over time. This is because they can use large electronic health record databases without researchers having to explain how different features work together, which is not possible with traditional regression-based models. Using this method, you can make predictive models that are always changing and that show how patients are doing in the clinic right now. Temporal modelling enables the examination of long-term variations in oxygenation, inflammatory markers, and organ dysfunction. This helps us find phenotypes of ARDS that are more likely to be dangerous and maybe even different biological subtypes.

Several recent systematic reviews and meta-analyses have synthesized the growing body of information that compares machine learning techniques with traditional methods. He et al [6] (2025) evaluated the superiority of artificial intelligence–based models compared to logistic regression in predicting ARDS mortality. Their joint research revealed that AI models exhibited superior sensitivity and specificity, along with enhanced overall prediction accuracy, as evidenced by elevated area under the curve values.

Tan et al. [7] (2025) discovered that machine learning models exhibited robust discriminative capability, achieving C-indexes of 0.84 in training datasets and 0.81 in external validation cohorts, thereby generally surpassing traditional prognostic scoring systems like SOFA and SAPS-II. Nevertheless, both reviews underscored significant heterogeneity among studies and stressed the necessity for stringent external validation before clinical application.

Primary studies have confirmed these results in different groups of people with ARDS. Xu et al. (2025) used MIMICIV data to make several machine learning models that could predict how likely it was that patients with sepsis-related ARDS would die in the hospital. They discovered that random forests were the most effective. The main predictors were the APACHE III score, serum bicarbonate, anion gap, and systolic blood pressure. Lin et al. Lin et al. [8] (2024) utilized machine learning algorithms to forecast both the incidence and mortality of non-pulmonary sepsis-associated ARDS, with XGBoost demonstrating superior performance. Their model exhibited significant external validation, achieving 78.0% accuracy for incidence and 81.4% for mortality across various patient populations.

Villar et al. [9] (2023) conducted an extensive multicenter study with patients suffering from mild to severe ARDS, assessing a range of advanced predictive techniques, such as Random Forest, XGBoost, and logistic regression enhanced by genetic algorithms. The models had AUC values that ranged from 0.87 to 0.91. Some of the most relevant predictors were age, immunosuppression, the PaO_2/FiO_2 ratio, plateau pressure, and the number of organ failures.

Huang et al. [4] showed that Random Forest models trained on MIMIC-III data achieved AUROC values between 0.88 and 0.90, which were higher than the values of the SOFA and SAPS-II scores. Their research demonstrated the significance of ensuring models are comprehensible. SHAP analysis showed that AST, APACHE score, and length of stay in the ICU are all important signs of death.

Researchers have also investigated how to predict the development of ARDS early on. Zhang et al. [10] made machine learning models that can predict ARDS in ICU patients with sepsis early on. The Gaussian Naive Bayes model did the best job, with an AUC of about 0.78. The model found important predictive factors, such as the PaO_2/PAO_2 ratio and C-reactive protein levels. This could give doctors a way to intervene early.

Table I provides an overview of commonly used machine learning approaches for predicting mortality in patients with ARDS. Although supervised learning models often achieve strong predictive performance, they depend on labelled outcomes, which restricts their ability to discover previously unknown or clinically meaningful patient subgroups. Similarly, linear dimensionality reduction techniques such as PCA may fail to capture complex nonlinear relationships among physiological variables. Furthermore, conventional clustering methods applied to either raw features or linearly transformed data frequently produce clusters that are poorly separated and difficult to interpret from a clinical perspective.

Clustering offers a potential solution to these limitations [8], [12]. As an unsupervised approach, it does not require labelled outcomes and can reveal hidden patterns within complex conditions such as ARDS, where distinct patient phenotypes may exist but are not explicitly defined. To address this gap, this study introduces an autoencoder-based clustering framework that integrates nonlinear dimensionality reduction with K-means clustering to identify patient subgroups associated with mortality risk. The proposed approach generates a compact low-dimensional representation while preserving important structural relationships in the data, enabling the identification of five ARDS subgroups exhibiting different mortality profiles. This framework provides a practical

approach for phenotype discovery and risk assessment in critical care research.

TABLE I: Comparison of existing methods

Author	Method	Inferences
He et al. 2025 [6]	Meta-analysis: AI vs. log regression	AI models achieved higher sensitivity, specificity, and AUC; limited by heterogeneity and lack of external validation.
Tan et al. 2025 [7]	Systematic ML review of models	ML outperformed SOFA/SAPS-II (Cindex: 0.84 training, 0.81 validation); emphasized need for validation and interpretability.
Xu et al. 2025 [11]	Multiple ML algorithms on MIMIC-IV	Random forest best; key predictors: APACHE III, bicarbonate, anion gap, systolic BP.
Lin et al. 2025 [8]	XGBoost for Non-pulmonary sepsis-ARDS	strong external validation (78.0% incidence, 81.4% mortality).
Villar et al. 2023 [9]	Random Forest, XGBoost, genetic algorithm	AUC: 0.87–0.91; predictors: age, immunosuppression, PaO_2/FiO_2, plateau pressure, organ failure.
Huang et al. 2021 [4]	Random forest on MIMIC-III/eICU	AUROC: 0.88–0.90; outperformed SOFA/SAPS-II; SHAP identified AST, APACHE, ICU stay.
Zhang et al. 2019 [10]	Multiple ML for early ARDS prediction	Gaussian Naive Bayes best (AUC 0.78); predictors: PaO_2/PAO_2 ratio, CRP.

PROPOSED METHODOLOGY

This study aims to identify clinically meaningful subgroups of patients with ARDS based on specific physiological and clinical characteristics. This is widely recognized as a heterogeneous syndrome, characterized by considerable variability in clinical presentation, underlying pathophysiology, disease severity, and patient outcomes. Such variability results in differences in treatment response, disease progression, and overall prognosis, which complicate clinical decision-making and risk stratification. Identifying patient subgroups may therefore improve understanding of disease heterogeneity and support the development of more personalized treatment strategies.

The study further examines the relationship between the identified subgroups and mortality outcomes, assessing whether particular phenotypes are associated with increased or reduced risk of death. Overall, these findings may contribute to improved risk stratification, support clinical decision-making, and enhance outcome assessment in critically ill ARDS patients (Figure 1).

To achieve this objective, three unsupervised clustering pipelines combining dimensionality reduction techniques with K-means clustering were evaluated. In each pipeline, high-dimensional clinical and physiological data were first transformed into a lower-dimensional representation to retain relevant variability while reducing redundancy and noise. K-means clustering was then applied to group patients according to feature similarity. These approaches were intended to uncover latent patterns within the dataset that could correspond to distinct risk profiles, clinical trajectories, and mortality outcomes among ARDS patients.

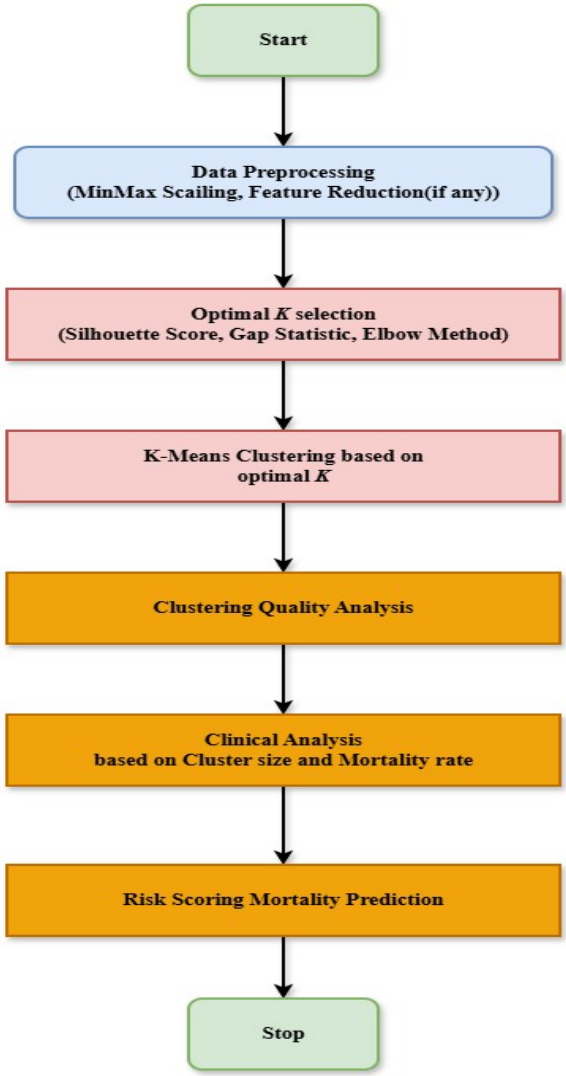


Figure. 1: Workflow of Proposed Methodology

A. Dimensionality Reduction Approaches

Clinical datasets often contain redundant or noisy features that may obscure natural group structures. For this reason, two dimensionality reduction strategies were applied before clustering.

- Autoencoder (AE) is a nonlinear neural network designed to learn a compressed representation of input data by minimising reconstruction error. In this framework, the encoder maps the 21 original clinical features into an eight-dimensional latent

space, capturing informative patterns while reducing redundancy and noise. This lower-dimensional representation preserves essential structure within the data and facilitates downstream analysis. The decoder subsequently reconstructs the original input features from the latent representation, ensuring that important information is retained during compression. The model was trained for 100 epochs using mean squared error as the loss function to measure reconstruction accuracy and guide optimization. After training, the latent representations generated by the encoder were extracted and used as input for K-means clustering to identify patient subgroups.

- Principal Component Analysis (PCA) is a linear dimensionality reduction method that projects high-dimensional data onto orthogonal components capturing maximum variance. In this approach, the feature space is transformed into principal components ranked according to explained variance. Components accounting for 90% of the total variance were retained, resulting in 18 principal components. Whitening was applied to standardise component variance so that each contributed equally during analysis. The transformed data were then clustered using K-means to identify patient groups within the reduced feature space. In addition, K-means clustering was performed directly on the scaled original features without dimensionality reduction, allowing comparison of clustering structure, stability, interpretability, and clinical relevance across representations.

B. Clustering and Choosing the Best k

K-means clustering was applied to each data representation (original features, PCA-transformed features, and autoencoder-derived latent features). The number of clusters (k) was selected automatically using three predefined criteria evaluated on the training data.

- Silhouette Score: The Silhouette Score measures how closely an observation resembles its assigned cluster compared with neighbouring clusters. It reflects both within-cluster cohesion and between-cluster separation, providing an overall indication of clustering quality. Values range from -1 to 1, with higher values indicating stronger structural separation and better cluster assignment [13].
- Gap Statistic: The Gap Statistic compares observed within-cluster dispersion with that expected under a null reference distribution generated from randomly distributed data. For each candidate value of k, the total within-cluster variation is computed and contrasted with the dispersion from the reference dataset. The optimal k corresponds to the value that maximizes the gap statistic, indicating that the clustering structure deviates meaningfully from randomness [14].
- Elbow Method: The elbow method evaluates the within-cluster sum of squares (inertia) as a function of the number of clusters (k). As 'k' increases, inertia decreases because clusters become smaller and more homogeneous. The "elbow" point represents a stage where further increases in k provide diminishing reductions in inertia, suggesting an appropriate and parsimonious cluster number [15].

After selecting the optimal number of clusters (k), the final K-means model was trained on the training set and applied to the test set for evaluation. Cluster quality was assessed using multiple validation metrics. The Davies–Bouldin index measured cluster similarity, where lower values indicate improved separation and compactness. The Calinski–Harabasz index evaluated the ratio of between-cluster to within-cluster dispersion, with higher values indicating better-defined clusters. Together, these metrics provide complementary evaluation of clustering performance and support assessment of subgroup robustness.

C. Heuristic for Scoring Mortality Risk

To examine whether clusters derived through unsupervised learning are associated with mortality outcomes, a simple risk score was calculated for each patient. The score corresponds to the Euclidean distance between the patient representation—derived from the original features, PCA-transformed features, or autoencoder latent space—and the nearest cluster centroid. This distance reflects how typical or atypical a patient is relative to the identified subgroup, with larger distances indicating greater deviation from the cluster center. The continuous score was converted into a binary mortality prediction using a threshold defined as the mean training-set distance. Although this approach does not involve supervised learning, it provides a quantitative indicator for evaluating the clinical relevance of the clustering structure. Linking cluster proximity to outcomes allows assessment of whether identified subgroups correspond to meaningful differences in patient risk profiles.

EXPERIMENTAL ANALYSIS AND FINDINGS

A. Dataset

The ARDS Patient Outcomes Dataset is a clinical tabular dataset that has been compiled to focus on patients diagnosed with ARDS [16]. The dataset includes patient demographics, detailed clinical features, comorbidities, vital signs, laboratory measurements, treatment information, and outcomes such as mortality and length of hospital or ICU stay. It has been structured for use in machine learning and statistical modelling, allowing relationships between patient characteristics and clinical outcomes to be analyzed and predictive algorithms for ARDS prognosis and risk stratification to be developed.

Clinical measurements, demographic information, treatments received, and outcome labels for a cohort of patients are provided, making the dataset suitable for exploring risk factors and building predictive models for mortality or other outcomes in ARDS. In total, the dataset comprises 1,000 patient records, each with 21 features and one target variable (mortality), enabling robust analyses of patient subgroups and outcome associations.

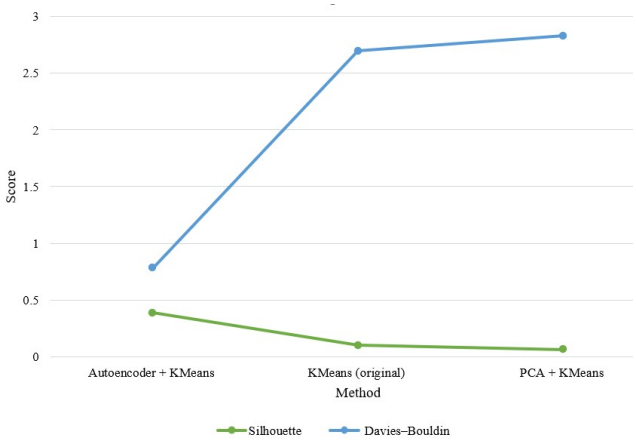


Fig. 2: Clustering quality metrics analysis

B. Plan of Action

The effectiveness of three clustering approaches was examined in this study: autoencoder-based K-means, PCA-based K-means, and conventional K-means applied to the original feature space, to identify clinically meaningful subgroups within the ARDS dataset. To enable a fair comparison, all methods followed the same procedure for determining the optimal number of clusters (k). The selection strategy combined the silhouette score and the gap statistic, which together provide complementary assessments of cluster cohesion and separation. In addition, a consistent distance-based risk scoring heuristic was used to explore potential relationships between cluster membership and patient mortality outcomes.

Cluster quality was assessed using multiple evaluation metrics, including the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index, allowing analysis of cohesion, separation, and overall structural validity. Clinical interpretability was examined by analyzing cluster sizes together with the corresponding mortality rates observed in each subgroup. Predictive usefulness was further explored through ROC-AUC evaluation and classification reports derived from the distance-based risk scores. The autoencoder was trained to learn an eight-dimensional latent representation that captures important data patterns while reducing noise, whereas PCA retained 90% of the total variance and served as a linear dimensionality reduction baseline for comparison.

All experiments were performed on the ARDS dataset after applying Min–Max scaling to standardize feature ranges and support stable model training. The dataset was divided into training and testing subsets, with a 70/30 split used for the autoencoder pipeline and an 80/20 split applied to both PCA-based and baseline K-means models to maintain consistency with prior experimental settings. Random seeds were fixed throughout all experiments to ensure reproducibility of clustering results and analytical outcomes. This standardized preprocessing and partitioning procedure enabled a balanced comparison across the different clustering approaches and their evaluation metrics.

The experiments were conducted on a workstation equipped with 8 GB RAM and a 12th-generation Intel i5 processor operating at 1.30 GHz. Model implementation and analysis were carried out using Python, supported by commonly used scientific libraries including NumPy, pandas, scikit-learn, Matplotlib, and Seaborn.

C. Analysis

1) Clustering Quality: Figure 2 presents the internal validation metrics for all three clustering approaches applied to the ARDS dataset. The autoencoder-based method achieved a Silhouette Score of 0.3892, indicating moderate cohesion within clusters, together with clear separation between them. In comparison, the baseline K-means method using the original features obtained a Silhouette Score of 0.1006, while the PCA-based K-means approach produced a lower score of- 0.0664. These near-zero values indicate that clusters generated using the baseline and PCA representations are weakly defined and exhibit considerable overlap. The observed difference in clustering quality can be linked to the ability of autoencoders to learn nonlinear transformations that preserve the intrinsic structure of the data and capture complex interactions among clinical features. In contrast, PCA relies on linear projections that are less capable of representing relationships present in high-dimensional clinical datasets, leading to overlapping and less informative clusters.

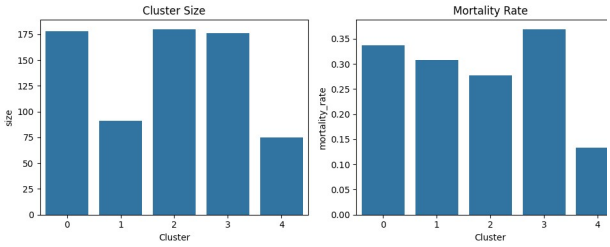
The Davies–Bouldin Index further reflects these performance differences. The autoencoder-based method achieved a value of 0.7837, indicating low within-cluster dispersion relative to inter-cluster separation and therefore more coherent groupings. By comparison, the baseline K-means approach yielded a higher score of 2.6930, and PCA-based clustering produced 2.8285, suggesting diffuse and poorly separated clusters. The Calinski–Harabasz Index, which measures the ratio of between-cluster variance to within-cluster variance, provides additional support for the effectiveness of the autoencoder representation. The autoencoder achieved a value of 1361.57, substantially higher than the baseline (53.76) and PCA (41.98) results, indicating dense and well-separated clusters within the learned latent space. Taken together, these metrics indicate that nonlinear feature extraction through the autoencoder captures structural patterns in the ARDS dataset more

effectively than either linear PCA or clustering performed directly on the original feature space.

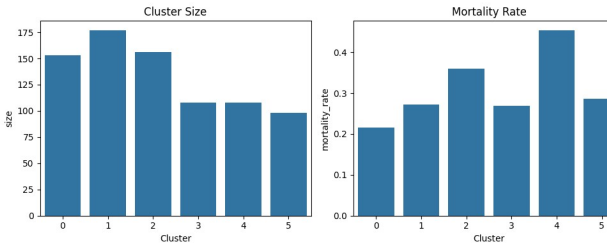
2) Clinical Interpretability of Clusters: Beyond statistical evaluation, the clinical usefulness of clustering depends on whether identified subgroups correspond to meaningful differences in patient outcomes. Figure 3 shows the cluster sizes and associated mortality rates for each method.

Autoencoder with K-Means: The autoencoder-based approach produced five distinct patient subgroups after applying K-means clustering to the latent representation (Figure 3a). These clusters display a monotonic progression in mortality risk when ordered by increasing mortality rate, ranging from 13.3% (Cluster 4) to 36.9% (Cluster 3). This pattern enables clinical stratification into three categories: low-risk (Cluster 4, 13.3% mortality), intermediate-risk (Clusters 0, 1, and 2, with mortality rates of 33.7%, 30.8%, and 27.8%), and high-risk (Cluster 3, 36.9% mortality). The distribution of patients across clusters provides further insight into the disease spectrum. The two largest clusters (Clusters 2 and 0) together include 358 patients and show intermediate mortality rates, suggesting that many ARDS patients fall within a moderate-risk category. In contrast, the smallest cluster (Cluster 4, n=75) exhibits the lowest mortality risk and may represent a clinically distinct phenotype associated with less severe disease presentation. The high-risk cluster (Cluster 3, n=176) represents a substantial subgroup with elevated mortality, indicating the need for closer clinical monitoring and potentially more intensive intervention strategies

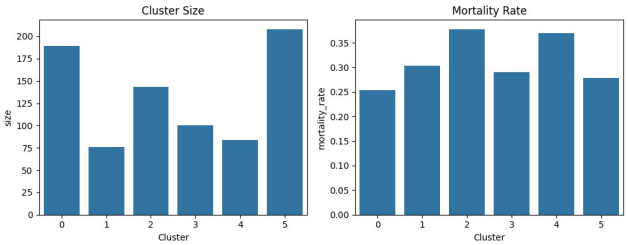
Baseline K-means: The standard K-means algorithm applied directly to the scaled feature space produced



(a) Autoencoder with K-means



(b) Baseline K-means



(c) PCA with K-means

Fig. 3: Cluster summaries for each method showing cluster size (left bars) and mortality rate (right bars).

Six clusters according to the selected optimal (k) criteria (Figure 3b). Although mortality rates across clusters spanned a broader range (21.6% to 45.4%) compared with the autoencoder-derived groups, several observations raise concerns regarding their clinical validity. Cluster 4 shows the highest mortality rate (45.4%) and includes 108 patients, which could be clinically meaningful if consistently validated. However, the absence of a monotonic mortality progression across clusters suggests suboptimal grouping. Mortality rates varied irregularly (21.6%, 27.1%, 35.9%, 26.9%, 45.4%, 28.6%) rather than following a gradual trend that would indicate a continuum of disease severity.

- PCA with K-Means: The PCA-based approach generated six clusters after dimensionality reduction, retaining 18 components that explained 90% of the variance (Figure 3c). Cluster sizes varied considerably, with Cluster 5 containing the largest subgroup (n=208), followed by Cluster 0 (n=189), Cluster 2 (n=143), Cluster 3 (n=100), Cluster 4 (n=84), and Cluster 1 (n=76). These distributions show that the PCA-transformed representation partitions patients into groups of uneven size.

  The corresponding mortality rates display an inconsistent pattern that limits clinical interpretability. Mortality does not follow a monotonic progression when clusters are ordered by risk (25.4% for Cluster 0, 30.3% for Cluster 1, 37.8% for Cluster 2, 29.0% for Cluster 3, 36.9% for Cluster 4, and 27.9% for Cluster 5). The alternating pattern, in which higher-mortality clusters appear among lower-risk groups, suggests that the clustering solution does not reflect a meaningful severity continuum. This outcome indicates that linear transformations introduced by PCA may obscure relationships associated with disease progression, producing numerically distinct clusters that do not align clearly with clinically relevant outcome differences.

3) Mortality Prediction via Distance Scores: As a secondary evaluation, the distance from each patient to the nearest cluster centroid was used as a heuristic risk score, thresholded at the mean training distance to predict mortality. Table II reports the ROC-AUC and accuracy on the test set.

TABLE II: Performance of the distance-based risk score for mortality prediction.

Method	Test ROC-AUC	Test Accuracy
Autoencoder with K-Means	0.5244	0.53
K-Means (original)	0.4365	0.48
PCA with K-Means	0.4876	0.52

The autoencoder risk score achieved the highest ROC-AUC (0.524), although only slightly above random performance. The baseline and PCA scores remained close to 0.5 or below. This modest predictive ability is expected because the score represents a heuristic measure rather than a trained predictive classifier. Nevertheless, the autoencoder-based representation shows a slightly stronger association with mortality, indicating that the latent space retains information related to patient outcomes.

Among the three evaluated methods, the autoencoder-driven clustering approach demonstrates overall advantages. It produces statistically consistent clusters with clearer clinical interpretation, reveals a meaningful risk stratification pattern, and highlights the importance of nonlinear dimensionality reduction for analysing complex clinical datasets. These results support the use of autoencoder-based clustering as a useful approach for phenotype discovery in critical care research.

DISCUSSION

Autoencoder-driven clustering improves unsupervised ARDS classification of patients by learning non-linear representations that capture complex clinical relationships.

Traditional clustering methods struggle with noisy, high-dimensional medical data, but the autoencoder uses a neural network to compress high-dimensional features into a meaningful latent space. The autoencoder-based method outperformed standard K-Means and PCA-based K-Means in all internal validation metrics on the ARDS dataset. The autoencoder method outperformed baseline and PCA-based clustering in Davies-Bouldin Index values for cluster compactness and separation. The Calinski-Harabasz Index confirmed these findings, showing that autoencoders outperformed other methods.

For clinical interpretability, the autoencoder identified five patient subgroups with clear mortality rates, enabling risk stratification. While distance-based risk score had modest predictive power, it outperformed baseline K-Means and PCA-based clustering. Autoencoder clusters showed smooth mortality progression, while comparative methods showed less interpretable patterns, demonstrating similar advantages in clinically coherent subgroup discovery. Autoencoder-driven clustering is a promising method for finding meaningful patient subgroups in critical care settings, which could improve ARDS risk stratification and personalized treatment strategies.

LIMITATIONS

Several limitations should be considered when interpreting the findings of this study. First, the analysis was conducted using a single dataset, and therefore, the generalizability of the identified subgroups to other patient populations remains uncertain. Differences in clinical practices, demographic distributions, and data collection procedures across institutions may influence clustering outcomes. Second, the autoencoder architecture employed in this work was intentionally simple to maintain interpretability and computational efficiency; deeper or alternative architectures may uncover additional latent patterns. Third, the distance-based mortality risk score represents a heuristic rather than a supervised predictive model, which explains its modest predictive performance. Finally, although internal validation metrics demonstrate improved clustering structure, external clinical validation is required before practical deployment in real-world healthcare settings.

FUTURE WORK

Future research should focus on validating the proposed framework using independent and multi-institutional datasets to assess generalizability across diverse patient populations and clinical environments. External validation is essential to determine whether the identified ARDS subgroups remain stable under varying healthcare practices and demographic conditions.

Another promising direction involves extending the model to incorporate multi-modal clinical data, such as longitudinal physiological measurements, medical imaging, genomic information, and biomarker profiles. Integrating these heterogeneous data sources through multi-modal autoencoder architectures may enable more refined patient phenotyping and improved risk characterization.

Improving interpretability of the latent representations also remains an important objective. Techniques such as feature attribution methods and latent space visualization could help clinicians understand how specific clinical variables influence subgroup formation. Additionally, prospective clinical studies are needed to evaluate whether treatment strategies guided by subgroup identification can improve patient outcomes.

Finally, developing a real-time clinical decision support system based on the proposed framework could facilitate

practical deployment in intensive care settings, enabling earlier intervention and optimised resource allocation.

CONCLUSION

This study presented an autoencoder-based clustering framework for identifying clinically meaningful subgroups among patients with ARDS. By learning nonlinear representations of high-dimensional clinical data, the proposed approach addressed limitations associated with traditional clustering methods that rely on linear assumptions. Experimental evaluation demonstrated that the autoencoder model significantly outperformed baseline K-means and PCA-based clustering across all validation metrics.

The proposed method achieved superior clustering quality, including a Silhouette Score of 0.3892, a Davies–Bouldin Index of 0.7837, and a Calinski–Harabasz Index of 1361.57. Furthermore, the identified clusters revealed a clear mortality gradient ranging from 13.3% to 36.9%, enabling clinically interpretable risk stratification. Although the distance-based risk scoring approach achieved only modest predictive accuracy, it demonstrated that the autoencoder latent space captures outcome-relevant information. These findings emphasize the importance of nonlinear dimensionality reduction for uncovering meaningful patterns within complex healthcare datasets.

Future validation using external cohorts remains necessary before clinical deployment; however, the proposed framework shows strong potential for supporting clinical phenotyping, improving risk assessment, and advancing personalized treatment strategies for ARDS patients.

Beyond ARDS research, the proposed framework illustrates how representation learning can support broader applications in clinical data analysis. As healthcare datasets continue to grow in complexity and scale, methods capable of uncovering latent patient structures may play an increasingly important role in precision medicine. Integrating such approaches with clinical decision support systems could enable continuous risk assessment and adaptive treatment planning, ultimately contributing to improved patient outcomes and more efficient utilisation of intensive care resources.

REFERENCES

[1] K. Aronson and K. Rajwani, “The acute respiratory distress syndrome: a clinical review,” Journal of Emergency and Critical Care Medicine, vol. 1, no. 9, 2017.

[2] L. Al-Husinat, S. Azzam, S. Al Sharie, M. Araydah, D. Battaglini, S. Abushehab, G. A. Cortes-Puentes, M. J. Schultz, and P. R. Rocco, “A narrative review on the future of ards: evolving definitions, pathophysiology, and tailored management,” Critical Care, vol. 29, no. 1, p. 88, 2025.

[3] E. Fan, D. Brodie, and A. S. Slutsky, “Acute respiratory distress syndrome: advances in diagnosis and treatment,” Jama, vol. 319, no. 7, pp. 698–710, 2018.

[4] B. Huang, D. Liang, R. Zou, X. Yu, G. Dan, H. Huang, H. Liu, and

Y. Liu, “Mortality prediction for patients with acute respiratory distress syndrome based on machine learning: a population-based study,” Annals of translational medicine, vol. 9, no. 9, p. 794, 2021.

[5] N. Ding, T. Nath, M. Damarla, L. Gao, and P. M. Hassoun, “Early predictive values of clinical assessments for ards mortality: a machinelearning approach,” Scientific reports, vol. 14, no. 1, p. 17853, 2024.

[6] Y. He, N. Liu, J. Yang, Y. Hong, H. Ni, and Z. Zhang, “Comparison of artificial intelligence and logistic regression models for mortality prediction in acute respiratory distress syndrome: a systematic review and meta-analysis,” Intensive Care Medicine Experimental, vol. 13, no. 1, p. 23, 2025.

[7] R. Tan, C. Ge, Z. Li, Y. Yan, H. Guo, W. Song, Q. Zhu, and

Q. Du, “Early prediction of mortality risk in acute respiratory distress syndrome: systematic review and meta-analysis,” Journal of Medical Internet Research, vol. 27, p. e70537, 2025.

[8] J. Lin, C. Gu, Z. Sun, S. Zhang, and S. Nie, “Machine learning-based model for predicting the occurrence and mortality of nonpulmonary sepsis-associated ards,” Scientific Reports, vol. 14, no. 1, p. 28240, 2024.

[9] J. Villar, J. M. Gonzalez-Mart´ ´ın, J. Hernandez-Gonz´ alez, M. A. Ar-´ mengol, C. Fernandez, C. Mart´ ´ın-Rodr´ıguez, F. Mosteiro, D. Mart´ınez, J. Sanchez-Ballesteros, C. Ferrando´ et al., “Predicting icu mortality in acute respiratory distress syndrome patients using machine learning: the predicting outcome and stratification of severity in ards (postcards) study,” Critical care medicine, vol. 51, no. 12, pp. 1638–1649, 2023.

[10] Z. Zhang, “Prediction model for patients with acute respiratory distress syndrome: use of a genetic

algorithm to develop a neural network model," PeerJ, vol. 7, p. e7719, 2019.

[11] Z. Xu, K. Zhang, D. Liu, and X. Fang, "Predicting mortality and risk factors of sepsis related ards using machine learning models," Scientific Reports, vol. 15, no. 1, p. 13509, 2025.

[12] P. Sinha, A. Spicer, K. L. Delucchi, D. F. McAuley, C. S. Calfee, and M. M. Churpek, "Comparison of machine learning clustering algorithms for detecting heterogeneity of treatment effect in acute respiratory distress syndrome: a secondary analysis of three randomised controlled trials," EBioMedicine, vol. 74, 2021.

[13] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 747–748.

[14] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 63, no. 2, pp. 411–423, 01 2002.

[15] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," Journal of the royal statistical society. series c (applied statistics), vol. 28, no. 1, pp. 100–108, 1979.

[16] Ziya, "Ards patient outcomes dataset," Kaggle, 2025, accessed: 2025-04-02. [Online]. Available:

https://www.kaggle.com/datasets/ziya07/ards-patient-outcomes-dataset