Journal of Biomedical Informatics 180 (2026) 105059

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Robust end-to-end stratification of amyotrophic lateral sclerosis patients via recurrent variational autoencoder and consensus clustering

Federico De Mori Bajolin [a], Erica Tavazzi [b], Anna M. Bianchi [a,c,*], Martin O. Mendez [a]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy
[b] Department of Information Engineering, University of Padova, Padua, 35131, Italy
[c] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

ARTICLE INFO

Keywords:
Deep learning
Clustering
Unsupervised learning
Amyotrophic Lateral Sclerosis
Stratification

ABSTRACT

Objective: This study aims to develop a data-driven methodology for stratifying Amyotrophic Lateral Sclerosis (ALS) patients based on longitudinal disease progression patterns, using a novel deep learning framework that combines a Recurrent Variational Autoencoder (RVA) with consensus clustering to identify clinically meaningful subgroups.

Methods: The RVA integrates Peephole Long Short-Term Memory networks within the Variational Deep Embedding (VaDE) architecture to simultaneously learn latent representations and cluster assignments from multivariate time-series data. The approach incorporates hyperparameter optimization via prediction strength with two-fold cross-validation, consensus clustering, and internal validation metrics (Silhouette Coefficient, Davies–Bouldin index, Calinski–Harabasz index) for optimal cluster selection. The methodology was validated on simulated data and applied to 3076 ALS patients from the PRO-ACT dataset, using ALSFRS-R total scores, domain subscores, and MiToS staging from the first six months of observation.

Results: Simulation experiments demonstrated that consensus clustering consistently outperformed single-model predictions across all noise levels. Applied to the PRO-ACT real data, the framework identified five distinct patient subgroups. These clusters exhibited distinct progression patterns and statistically significant differences in baseline clinical features, disease onset characteristics, and survival outcomes, with median survival ranging from 12.8 months to 27.5 months.

Conclusion: The proposed deep learning framework effectively captures the heterogeneous nature of ALS progression and identifies clinically relevant patient subgroups using routine clinical assessments. The stratification provides a foundation for personalized prognosis, optimized clinical trial design, and tailored therapeutic strategies, representing a practical tool for improving ALS patient management.

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is one of the most challenging neurodegenerative disorders in clinical neurology. It is characterized by the progressive degeneration of motor neurons, which results in muscular weakness, paralysis, and ultimately respiratory failure [1]. Although significant advances have been achieved in understanding the molecular mechanisms underlying ALS pathophysiology, therapeutic options remain largely palliative, focused primarily on slowing down symptom worsening and extending life expectancy. The median survival after diagnosis ranges between 3 and 5 years [2], although this prognosis is subject to considerable inter-individual variability. Such differences manifest not only in survival outcomes but also in disease progression rates and clinical phenotypes [3]. This variability poses substantial challenges for clinicians in determining the optimal timing for critical interventions (including gastrostomy placement, non-invasive ventilation initiation, and disease-modifying treatments) and complicates the evaluation of treatment efficacy in both clinical practice and research settings.

To monitor disease evolution and functional status, ALS patients are regularly assessed after diagnosis through standardized clinical instruments that evaluate functional disability, respiratory capacity, and disease-related events. The Revised ALS Functional Rating Scale (ALSFRS-R) [4] represents the gold standard tool for this purpose, offering a questionnaire-based approach to quantify and track changes in patient functional status over time. The scale includes both a total score and domain-specific subscores that capture impairment across bulbar, limb, trunk, and respiratory functions. These longitudinal assessments

* Corresponding author at: Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy.
E-mail address: annamaria.bianchi@polimi.it (A.M. Bianchi).

https://doi.org/10.1016/j.jbi.2026.105059
Received 10 March 2026; Received in revised form 15 May 2026; Accepted 27 May 2026
Available online 2 June 2026

Table 1
Statement of significance.

Problem	ALS progression exhibits substantial heterogeneity, which complicates determining intervention timing and evaluating treatment efficacy.
What is Already Known	Traditional stratification methods often assume linear progression or analyze single variables and timepoints, missing complex disease dynamics.
What This Paper Adds	A deep learning framework that learns multivariate, non-linear temporal dependencies across different clinical domains (ALSFRS-R) to robustly identify patient subgroups.
Who would benefit from the knowledge in this paper	Clinicians for personalizing prognosis and care, and researchers for optimizing clinical trial design.

can help identify disease progression patterns and improve clinical understanding of ALS.

Despite this, characterizing patients' progression using these scores is far from straightforward. Clinical evaluations are often performed at different stages of each patient's condition, and although an overall decline is typically observed over time, individuals may still experience intermittent phases of slight improvement or temporary stability [5]. Identifying common patterns of disease evolution and grouping patients according to their progression profiles can play a crucial role in improving patient stratification, supporting the development of more precise inclusion criteria for clinical trials, and enhancing prognostic models [6]. Ultimately, by analyzing these progression patterns, clinicians can make better-informed decisions and design supportive care plans tailored to each patient's specific needs.

In this work, we propose a novel data-driven methodology to stratify ALS patients from longitudinal data by integrating a Recurrent Variational Autoencoder (RVA) with consensus clustering. This methodology consists of four main components: (1) using the RVA to extract latent representations and cluster assignments from multivariate time-series data, (2) selecting optimal hyperparameters for the neural network based on prediction strength and two-fold cross-validation, (3) aggregating multiple clustering solutions to enhance robustness via consensus clustering, and (4) selecting the optimal number of clusters with internal validation metrics. Our findings demonstrate the effectiveness of this pipeline: simulations validated the framework's robustness, while its application to PRO-ACT data revealed five distinct patient subgroups characterized by different progression patterns, and significant differences in both baseline characteristics and median survival.

2. Related work

Patient stratification in ALS is a critical challenge for precision medicine, due to the substantial heterogeneity in clinical manifestations and disease progression. The complexity of this task derives from the need to capture the dynamic nature of the disease, which evolves differently across patients over time. Researchers have developed various approaches to address this challenge, each attempting to group patients based on similar baseline profiles or progression patterns. These approaches can be broadly categorized into five main strategies: clinical scoring-based methods, staging systems, clustering algorithms, non-linear parametric modeling methods, and pattern-mining techniques. While each strategy offers distinct advantages, they often struggle to fully represent the temporal evolution and non-linear progression patterns inherent in ALS.

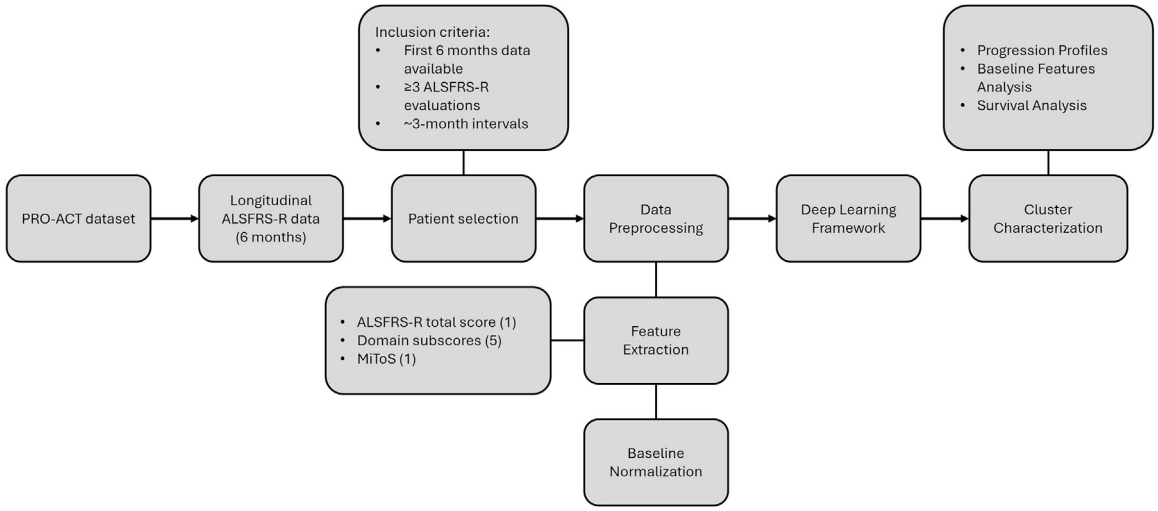
The earliest and most straightforward approaches to patient stratification rely on clinical scoring systems that assess disease severity at specific time points [7]. A fundamental limitation of these approaches is the assumption of linear disease progression in ALS, despite growing evidence of non-linear progression patterns that fluctuate according to disease severity [8]. One widely adopted stratification criterion is the progression rate, which categorizes ALS patients into fast, moderate, and slow progressor groups. This metric is computed as the difference between the maximum possible ALSFRS-R score (48) and the score recorded at the time of assessment, normalized by disease duration measured in months. Patient groupings are then established by applying threshold values to the resulting distribution of progression rates [9,10]. While this approach provides a simple and interpretable classification, it reduces the complexity of disease progression to a single numerical value, potentially overlooking important patterns in how the disease evolves over time.

Staging systems represent an alternative framework that assigns patients to discrete disease phases according to their clinical characteristics. Three principal staging frameworks are employed in ALS research: the Kings' staging system [11], the Milano-Torino Staging (MiToS) system [12], and the FT9 system [13]. These frameworks have been frequently applied in stratification studies [14,15] and offer the advantage of providing clinically meaningful categories that reflect the functional status of patients. However, these systems share a significant limitation with scoring-based methods: they allocate patients to categories based on diagnostic or functional features assessed at a single time point, thereby neglecting longitudinal follow-up information that could reveal important progression dynamics.

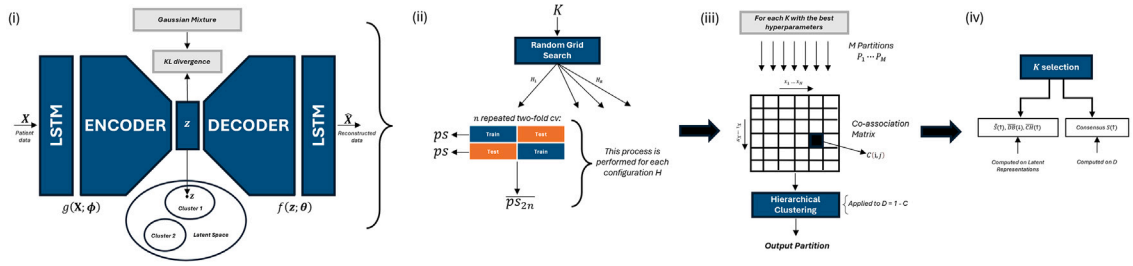
To address the limitations of single-timepoint assessments, researchers have increasingly turned to clustering algorithms that can incorporate temporal data. Pires et al. [6] employed the Expectation–Maximization (EM) algorithm to identify distinct patient profiles within the ALS population using temporal data from multiple clinical visits. This approach represented an important step forward by incorporating longitudinal observations to create patient snapshots and compute evolution classes. However, the clustering procedure itself was applied to snapshot-level data rather than modeling the full temporal trajectory of individual patients' disease progression. Similarly, ensemble-based methodologies [16] utilized data from multiple time points but focused on aggregating temporal information into summary statistics rather than explicitly modeling disease progression dynamics. These methods, while incorporating temporal information, still do not fully capture the continuous nature of disease evolution.

The evidence that ALS progression follows complex, non-linear patterns has led to more sophisticated modeling approaches. While non-linear parametric methods [17–20] can capture curved trajectory shapes and have improved our understanding of disease dynamics, their reliance on predefined functional forms may limit their ability to identify unexpected progression patterns. To overcome this limitation, Ramamoorthy et al. [8] developed a methodology for characterizing ALS progression based on mixture of Gaussian processes (MoGP). Their framework integrates Gaussian process regression to enable identification of non-linear trajectory patterns, employs a Dirichlet process mixture model for data clustering, and incorporates a Monotonic Inductive Bias to promote the learning of declining trajectories. This non-parametric approach successfully identified patient clusters exhibiting comparable disease progression patterns and revealed the coexistence of both linear and non-linear trajectories across patients, thereby providing justification for flexible modeling approaches capable of accommodating diverse trajectory types. However, this approach remains univariate, focusing exclusively on ALSFRS-R score without accounting for relationships between different functional domains.

More recently, pattern-mining methodologies have emerged as complementary approaches for investigating ALS disease progression, offering the ability to discover complex relationships within temporal data by considering multiple features simultaneously. Amaral et al. [21] developed ClusTric, which integrates triclustering with hierarchical clustering to achieve patient stratification based on disease progression patterns. Their three-step approach discovers coherent progression



(a) General workflow.



(b) Deep Learning Framework.

Fig. 1. Overview of the proposed approach. **(a)** Schematic representation of the proposed deep learning-based stratification workflow applied on ALS real data. **(b)** Structure of the developed deep learning framework, which includes: (i) a recurrent variational autoencoder; (ii) hyperparameter optimization: $\overline{ps}_{2n}$ (average prediction strength over $2n$ train–test splits from n repetitions); (iii) consensus clustering: D (distance matrix); (iv) internal validation and optimal number of clusters: $\overline{S}$, $\overline{DB}$, $\overline{CH}$ (average internal metrics computed across M runs).

patterns across patients, features, and time using TCtriCluster [22], computes representative 3D virtual patterns, and groups patients via hierarchical clustering based on pattern similarity.

While ClusTric excels at discovering interpretable progression patterns, it presents several methodological limitations that motivated, in this work, the adoption of a deep learning framework (RVA) (see Table 1). First, it may struggle to capture complex non-linear dynamics inherent in disease trajectories. Second, this approach requires extensive preprocessing, increasing analytical complexity. Third, its sequential pipeline optimizes triclustering and hierarchical clustering independently, potentially leading to suboptimal patient stratification. Finally, ClusTric's deterministic clustering method lacks mechanisms to quantify uncertainty in patient assignments, making it difficult to distinguish high-confidence classifications from boundary cases. Conversely, RVA intrinsically captures non-linear dynamics, enables end-to-end optimization, and quantifies uncertainty. Notably, similar deep learning approaches have already successfully stratified patients in other neurodegenerative diseases [23].

3. Methods

The overall stratification workflow is illustrated in Fig. 1(a), which provides a high-level overview of the sequential steps from raw clinical data to the identification of patient subgroups. Specifically, the main steps include (1) patient selection based on defined inclusion criteria, (2) data preprocessing via feature extraction and baseline normalization, (3) the application of the deep learning framework, and (4) the final cluster characterization. The practical implementation of the initial data curation and preprocessing steps is detailed in the Experimental Setup (Section 4). The deep learning framework underlying this workflow comprises four principal modules, as depicted in Fig. 1(b): (1) a Recurrent Variational Autoencoder (RVA), (2) hyperparameter optimization, (3) consensus clustering, and (4) internal validation for selecting the optimal number of clusters. The following sections provide a detailed description of each component.

3.1. Recurrent variational autoencoder

The RVA is the core component of the proposed framework, designed to integrate Peephole Long Short-Term Memory (LSTM) networks [24] within the Variational Deep Embedding (VaDE) architecture [25]. This specific configuration allows for the simultaneous learning of meaningful latent representations and cluster assignments from multivariate time-series data, addressing two fundamental challenges in clinical trajectory analysis: effectively capturing multivariate temporal dynamics and avoiding suboptimal two-step clustering.

First, to model the temporal evolution of the disease progression, the integrated Peephole LSTMs are used similarly to the approach of Roque et al. [26]. These architectures have demonstrated superior performance compared to classical methods for multivariate time-series modeling [27,28]. Such networks are particularly suited for capturing precise temporal patterns, such as critical phases of deterioration within a clinical trajectory. By incorporating peephole connections, the LSTM allows its gating mechanisms to directly access the internal memory cell state. In a multivariate context, this structural feature enables the network to leverage its internal memory as a dynamic, step-by-step

global summary of interacting clinical domains. Consequently, it can explicitly track both autocorrelations (the temporal evolution within a single functional score) and complex cross-correlations (the interaction between different domains) without losing critical state transitions.

Second, to avoid the aforementioned suboptimal two-step clustering, the actual partitioning is performed through the VaDE architecture, which integrates a Gaussian Mixture (GM) prior [25]. Conventional sequential clustering methods often rely on a disconnected two-step process (e.g., autoencoder-based dimensionality reduction followed by post-hoc clustering). This fragmented approach can fail to preserve the underlying data structure and lead to suboptimal grouping. In contrast, integrating a GM prior within a classical variational autoencoder (VAE) ensures that latent representation and cluster assignment are jointly optimized. As demonstrated in the original VaDE study [25], this joint optimization significantly outperforms traditional methods.

This approach facilitates the discovery of patient subgroups with heterogeneous disease trajectories. The detailed mathematical formulation of these modules is presented in Section 3.1.1.

3.1.1. Model architecture and mathematical formulation

The presented architecture builds on an encoder–decoder framework designed to effectively handle multivariate longitudinal data. In this context, each patient is characterized by a time-ordered sequence of clinical assessments $\mathbf{X} = \{\mathbf{y}_t, \mathbf{y}_{t+1}, \dots\}$, where $\mathbf{y}_t \in \mathbb{R}^M$ represents the M-dimensional vector of clinical variables at visit t, and the full trajectory $\mathbf{X} \in \mathbb{R}^{T\times M}$ spans T time points.

The encoder network $g(\mathbf{X};\boldsymbol{\phi})$ consists of two sequential modules, where $\boldsymbol{\phi}$ denotes the learnable parameters of the encoder network. In the first module, the input trajectory $\mathbf{X}$ is processed by a Peephole LSTM, an extension of the standard LSTM architecture. Unlike a classical LSTM, in which gate activations depend exclusively on the hidden state and the current input, the Peephole LSTM extends this mechanism by incorporating direct connections from the previous cell state to the gating units [24]. In the second module, the representation obtained from the Peephole LSTM is fed into a fully connected neural network that parametrizes a probabilistic mapping to the latent space. Specifically, it models $q(\mathbf{z}|\mathbf{X}) = \mathcal{N}(\mathbf{z};\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\sigma}}^2\mathbf{I})$ in the variational posterior $q(\mathbf{z}, c \mid \mathbf{X}) = q(\mathbf{z} \mid \mathbf{X})\, q(c \mid \mathbf{X})$, where $\mathbf{z} \in \mathbb{R}^J$ is the latent representation vector of dimension J, $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^J$ is the mean vector, $\tilde{\boldsymbol{\sigma}}^2 \in \mathbb{R}^J$ is the variance vector, and $c \in \{1, \dots, K\}$ represents the cluster assignment variable among K predefined clusters. Through this formulation, the encoder learns to map each input trajectory to a multivariate Gaussian distribution in the latent space, from which a latent embedding $\mathbf{z}$ is sampled. The mean $\tilde{\boldsymbol{\mu}}$ encodes the location in the latent space, while the variance $\tilde{\boldsymbol{\sigma}}^2$ captures the uncertainty in this mapping.

While classical VAEs regularize this encoded distribution towards a standard normal prior $\mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$, where $\mathbf{0} \in \mathbb{R}^J$ is the zero vector and $\mathbf{I} \in \mathbb{R}^{J\times J}$ is the identity matrix [29], the latent space in RVA is approximated by a GM [25]. The model assumes the existence of K clusters, each characterized by a prior probability $p(c) = \mathrm{Cat}(c|\boldsymbol{\pi})$, where $\mathrm{Cat}(\cdot|\boldsymbol{\pi})$ denotes the categorical distribution parametrized by the mixing coefficients $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ with $\sum_{k=1}^K \pi_k = 1$, and a cluster-conditional Gaussian component distribution $p(\mathbf{z}|c) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2\mathbf{I})$, where $\boldsymbol{\mu}_c \in \mathbb{R}^J$ and $\boldsymbol{\sigma}_c^2 \in \mathbb{R}^J$ are respectively the mean and variance of cluster c in the latent space. Based on this mixture prior, cluster assignments are inferred by evaluating which Gaussian component better explains the encoded representation. Specifically, for a given patient, the posterior probability of belonging to cluster c is computed as:

$$q(c|\mathbf{X}) \equiv p(c|\mathbf{z}) = \frac{p(c)p(\mathbf{z}|c)}{\sum_{c'=1}^K p(c')p(\mathbf{z}|c')} \tag{1}$$

Consequently, the mixture prior shapes the latent space into distinct regions corresponding to different Gaussian components. Patients whose trajectories are encoded into the same region of this space are therefore more likely to be associated with the same cluster, since their latent representations are better explained by the same Gaussian component.

The decoder network $f(\mathbf{z};\boldsymbol{\theta})$, with learnable parameters $\boldsymbol{\theta}$, reconstructs the original patient trajectory from the latent representation $\mathbf{z}$, modeling the conditional distribution $p(\mathbf{X}|\mathbf{z})$. A feed-forward network transforms the latent embedding into the initial input vector of the Peephole LSTM, which then generates the sequence auto-regressively by predicting each clinical state $\hat{\mathbf{y}}_t \in \mathbb{R}^M$ based on the previous internal state and the past outputs $\hat{\mathbf{y}}_{<t}$.

3.1.2. Training and optimization

The RVA model is trained by maximizing the Evidence Lower Bound (ELBO), which jointly optimizes data reconstruction and latent-space regularization [25]:

$$\mathcal{L}_{\mathrm{ELBO}}(\mathbf{X}) = \mathbb{E}_{q(\mathbf{z},c|\mathbf{X})}[\log p(\mathbf{X}|\mathbf{z})] - D_{\mathrm{KL}}\big(q(\mathbf{z},c|\mathbf{X}) \parallel p(\mathbf{z},c)\big) \tag{2}$$

The first term, known as reconstruction loss, corresponds to the likelihood of observing $\mathbf{X}$ given a latent representation $\mathbf{z}$. This term forces the algorithm to learn good reconstructions of its input data. The second term is the Kullback–Leibler divergence from the GM prior $p(\mathbf{z}, c)$ to the variational posterior $q(\mathbf{z}, c|\mathbf{X})$, which regularizes the latent embedding $\mathbf{z}$ to lie on a GM manifold. Balancing these two objectives allows the model to learn interpretable and well-separated representations of disease progression: the reconstruction term ensures that the latent representation $\mathbf{z}$ encodes sufficient information to accurately reconstruct each patient's individual clinical trajectory $\mathbf{X}$, while the regularization term organizes these representations into distinct disease subgroups corresponding to the K cluster components.

To maximize $\mathcal{L}_{\mathrm{ELBO}}$, all model parameters are jointly optimized, including the mixing coefficients $\boldsymbol{\pi}$, cluster-specific means and variances $\{\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c\}$ where $c \in \{1, \dots, K\}$, and the encoder and decoder parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$. Gradient-based optimization is enabled through the reparameterization trick, expressing latent samples as:

$$\mathbf{z} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\sigma}} \odot \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

This formulation allows gradients to propagate through the stochastic sampling operation during backpropagation [29].

3.2. Hyperparameter optimization

To identify optimal hyperparameters for each candidate number of clusters K, a hyperparameter optimization procedure is employed, using prediction strength as the evaluation metric [30]. The hyperparameters considered include the number of hidden layers ($n_{\text{hidden layers}}$), number of neurons per layer (n_{neurons}), learning rate (η), and batch size.

For each K, hyperparameter configurations are sampled via random grid search and evaluated using n repeats of two-fold cross-validation. The choice of two-fold cross-validation is driven by the specific mechanics of the prediction strength metric, which requires fitting a fully independent clustering model directly on the test set. Higher-order splits would drastically reduce the test set size, introducing two critical issues: first, due to the inherent phenotypic imbalance within the ALS cohort, minority subgroups would not be reliably represented; second, the RVA would lack the data volume required to learn robust representations. Consequently, the independent test set clustering would fail. A two-fold split prevents this by ensuring the test set remains sufficiently large to guarantee adequate representation of all latent clusters and stable model convergence, an approach consistent with the original methodology by Tibshirani et al. [30]. To mitigate potential partitioning bias, the two-fold process is repeated n times using distinct random seeds. Crucially, for each single repetition, all hyperparameter configurations are evaluated on the same data partition, which ensures a fair comparison between different hyperparameter settings. This procedure leads to the selection of robust hyperparameters that generalize effectively across diverse data splits. The evaluation process works as follows: the dataset is randomly split into two equal parts, called fold A and fold B. In the first phase, fold A serves as the training set and fold B as the test set. RVA is fitted on fold A to obtain a clustering model,

which is then used to assign cluster labels to the samples in fold B. Next, RVA is fitted separately on fold B, producing a second independent clustering model that assigns a set of cluster labels to the samples in fold B. Prediction strength is then used to quantify the agreement between these two cluster assignments for fold B. In the second phase, the roles are reversed: fold B becomes the training set and fold A becomes the test set, and the same procedure is repeated. This means that for each repetition, both folds serve once as training data and once as test data, yielding two prediction strength measurements per repetition.

Formally, for a candidate number of clusters K, let $A_{K1}, A_{K2}, \ldots, A_{KK}$ be the indices of the test observations in test clusters $1, 2, \ldots, K$, where these clusters are obtained by fitting RVA directly on the test set. Let $m_{K1}, m_{K2}, \ldots, m_{KK}$ be the number of observations in these clusters. The prediction strength (ps) of the clustering $S(\cdot, K)$ is defined by:

$$\mathrm{ps}(K) = \min_{1 \le j \le K} \frac{1}{m_{Kj}(m_{Kj}-1)} \sum_{i \neq i' \in A_{Kj}} E[S(\mathbf{X}_{\mathrm{tr}}, K), \mathbf{X}_{\mathrm{te}}]_{ii'} \tag{4}$$

where $E[S(\mathbf{X}_{\mathrm{tr}}, K), \mathbf{X}_{\mathrm{te}}]_{ii'} = 1$ if test observations i and i' are assigned to the same cluster by the training-based clustering, and 0 otherwise. The prediction strength is the minimum of this quantity over the K test clusters. In other words, for each cluster of the test data model, compute the fraction of pairs of samples in that cluster that are also assigned to the same cluster by the training data model. Prediction strength is defined as the minimum proportion across all clusters of the test data model.

Based on this metric, for each K, the hyperparameter configuration with the highest average prediction strength computed across $2n$ independent train–test splits is selected. This approach ensures that the chosen hyperparameters produce clustering models that generalize well to unseen data.

3.3. Consensus clustering

Given the non-deterministic nature of RVA combined with the complexity of the dataset, it could produce slightly different partitions even with identical hyperparameters. To increase robustness of the cluster assignments, the neural network is trained M times for each K with its related optimal hyperparameter combination found with the procedure explained in Section 3.2. Then, a co-association matrix $\mathbf{C}$ is computed. Each element $C(i, j)$ of this matrix is calculated by counting the number of times (n_{ij}) two patients x_i and x_j are assigned to the same cluster and normalizing by the total number of runs [31]:

$$C(i, j) = \frac{n_{ij}}{M} \tag{5}$$

Subsequently, the co-association matrix is transformed into a distance matrix $\mathbf{D}$. Each element $D(i, j)$ of this matrix is computed as:

$$D(i, j) = 1 - C(i, j) \tag{6}$$

where pairs of patients frequently assigned to the same cluster have low distance values, while those rarely co-clustered have high distance values. Finally, an Agglomerative Hierarchical Clustering algorithm with average linkage is applied to $\mathbf{D}$ to obtain the final consensus partition [31].

3.4. Internal validation and optimal number of clusters

To select the optimal number of clusters K, the quality of the clustering is assessed using internal validation measures. Specifically, the Silhouette Coefficient (S), the Davies–Bouldin index (DB), and the Calinski–Harabasz index (CH) are used. The adoption of these three complementary metrics is motivated by the need for a comprehensive evaluation of each clustering structure.

The Silhouette Coefficient [32] measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation), providing a quantitative measure of the overall compactness and distinctiveness of the identified clusters. For a given clustering with K clusters, the silhouette score $s(\mathbf{X})$ for the $\mathbf{X}$-th sample in cluster G_i is defined as:

$$s(\mathbf{X}) = \frac{b(\mathbf{X}) - a(\mathbf{X})}{\max\{a(\mathbf{X}), b(\mathbf{X})\}} \tag{7}$$

where $a(\mathbf{X})$ is the mean distance between the $\mathbf{X}$-th sample and all other samples in G_i, and $b(\mathbf{X})$ is the mean distance between the $\mathbf{X}$-th sample and all samples in the nearest cluster. The total Silhouette Coefficient is the average of $s(\mathbf{X})$ over all samples.

The Davies–Bouldin index [33] assesses the average similarity between each cluster and its most similar cluster, enabling the identification of clusters with well-defined boundaries. For a pair of clusters i and j, the DB score is calculated as:

$$DB(i, j) = \frac{S(i) + S(j)}{CD_{ij}} \tag{8}$$

where $S(i)$ is the average distance between each point in cluster i and its centroid, and CD_{ij} is the distance between the centroids of clusters i and j. The final DB index is the average of the maximum ratio for each cluster.

The Calinski–Harabasz index [34] evaluates the ratio between the between-cluster dispersion and the within-cluster dispersion. For a given clustering with K clusters, the CH is defined as:

$$CH = \frac{SS_B/(K-1)}{SS_W/(n-K)} \tag{9}$$

where n is the total number of samples, k is the number of clusters, SS_B is the sum of squared distances between the cluster centroids and the global centroid (between-cluster sum of squares), and SS_W is the sum of squared distances of all samples from their respective cluster centroids (within-cluster sum of squares). Higher values of CH indicate better-defined clusters, with greater separation between clusters and greater cohesion within each cluster.

For each value of K, these three measures are computed on every latent representation across the M runs, and their mean and standard deviation are calculated. The optimal number of clusters is selected by comparing these aggregated metrics across different values of K. As an additional criterion, when the metrics have similar values, the Silhouette Coefficient (Consensus S) obtained on the final distance matrix for each K is compared, which evaluates partition consistency.

4. Experimental setup

The framework evaluation involves two stages. The first validates robustness against noise using synthetic ALS trajectories, comparing single-model and consensus predictions, fully presented in Section A.1 of the Supplementary Material. The second stage applies the methodology to real clinical data from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database to identify disease progression phenotypes and stratify patients. The following sections detail the experimental setup for this latter stage.

4.1. Dataset description

The longitudinal PRO-ACT dataset is curated by the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital (MGH), which comprises data from more than 10000 individuals diagnosed with ALS [35]. The dataset includes demographic information, family history, and extensive longitudinal clinical and laboratory measurements, making it the largest publicly accessible repository of integrated ALS clinical trial data. Multiple types of clinical assessments were routinely collected for each participant at every trial visit.

The PRO-ACT dataset aggregates clinical trials conducted between 1990 and 2010. Eligibility criteria inherent to clinical trials introduce recruitment bias, leading to a cohort that is less diverse. Compared with the general ALS population, PRO-ACT participants tend to be younger, exhibit slightly slower disease progression, and show a shorter interval between symptom onset and diagnosis.

Table 2
Baseline characteristics of PRO-ACT dataset and their average values. FVC% is the Forced Vital Capacity percent predicted.

Characteristics	Mean ± SD
Age (years)	56.2 ± 11.8
Percentage of females	40
Percentage of white	94.9
Disease duration (months)	23 ± 17
Diagnostic Delay (months)	11.6 ± 9.2
Onset site, percentage of limb onset	76
FVC%	86 ± 24
ALSFRS score	29.59 ± 5.84
ALSFRS-R score	38.37 ± 5.22
Body mass index at baseline	25.4 ± 4.4
Percentage of Riluzole use	77.5

Table 3
Input features.

Feature	Description	Values range
ALSFRS-R	Functional score for monitoring disease progression	[0,48]
ALSFRSb	Functional subscore for monitoring bulbar progression	[0,12]
ALSFRSul	Functional subscore for monitoring upper limbs progression	[0,8]
ALSFRSll	Functional subscore for monitoring lower limbs progression	[0,8]
ALSFRSt	Functional subscore for monitoring trunk progression	[0,8]
ALSFRSr	Functional subscore for monitoring respiratory progression	[0,12]
MiToS	Staging system based on ALSFRS-R	[0,4]

The PRO-ACT assessments are distributed across 16 data files, each corresponding to a specific data type: Adverse Events, ALSFRS(R), Concomitant Medications, Demographics, El Escorial criteria, Family History, Forced Vital Capacity, Handgrip Strength, Laboratory Data, Mortality, Muscle Strength, Riluzole Use, Slow Vital Capacity, Subject ALS History, Treatment Group, and Vital Signs. Mean baseline characteristics of the cohort are summarized in Table 2 [36].

The time duration between two consecutive trial points for ALS subjects is approximately 30 days. Therefore, patient states can be monitored on a monthly basis using the clinical trial information.

4.2. Input features

The framework utilizes disease progression measurements derived from the ALSFRS-R, a clinically validated questionnaire-based instrument for monitoring functional decline in ALS patients [4]. The ALSFRS-R comprises 12 questions that assess patient functionality across different affected domains: bulbar (Q1–Q3), upper limbs (Q4–Q5), trunk (Q6–Q7), lower limbs (Q8–Q9), and respiratory function (Q10–Q12). Each item is scored from 0 (complete loss of function) to 4 (normal function). For the model, we have considered both the total ALSFRS-R score, computed as the sum of the 12 questions, and domain-specific subscores (computed by summing the scores of the questions within each functional domain). Additionally, the Milano-Torino Staging (MiToS) system score [12] was included, a four-domain staging system that encodes as binary variables the presence of an impairment in movement, swallowing, communication, and breathing, providing an additional perspective on disease severity. Table 3 presents the complete set of input features used in the framework [21].

4.3. Data preprocessing

For the framework, ALSFRS-R functional assessment data extracted from this dataset were used, specifically focusing on the first 6 months of observation. Subjects having at least the first three evaluations recorded, each collected at approximately 3-month intervals [21], were selected, leading to an experimental cohort of 3076 individuals and 3 longitudinal observations (baseline, month 3, and month 6). For patients with longer follow-ups, trajectories were strictly truncated to these three initial time-points, ensuring no multiple samples were created for the same individual. From these longitudinal measurements, the seven input features required by the algorithm (see Section 4.2) were derived. Subsequently, longitudinal features were normalized following de Jong et al. [23], with each time point $x_{t,i}$ expressed relative to the cohort-wide baseline distribution:

$$\tilde{x}_{t,i} = \frac{x_{t,i} - \mu_{baseline,i}}{\sigma_{baseline,i}} \tag{10}$$

where $\mu_{baseline,i}$ and $\sigma_{baseline,i}$ denote the mean and standard deviation of feature i at baseline across the entire cohort, respectively.

Table 4
Internal validation metrics across different values of K. S = Silhouette score, DB = Davies–Bouldin index, CH = Calinski–Harabasz index. Values represent mean ± standard deviation across 50 runs of RVA. Consensus S is computed on the final distance matrix. Best values are highlighted in bold.

K	S (↑)	DB (↓)	CH (↑)	Consensus S (↑)
3	**0.38 ± 0.04**	0.98 ± 0.14	1983.20 ± 344.35	0.70
4	0.35 ± 0.03	0.99 ± 0.07	1921.77 ± 160.42	0.69
5	0.34 ± 0.02	**0.91 ± 0.06**	**2005.97 ± 155.90**	**0.72**
6	0.32 ± 0.03	0.92 ± 0.06	1922.22 ± 155.38	0.68

4.4. Optimization and consensus clustering

To identify the optimal number of patient subgroups, clustering was performed with $K \in \{3,4,5,6\}$ clusters. For each value of K, hyperparameter optimization was performed, as explained in Section 3.2, with two-fold cross-validation repeated $n = 10$ times. As established in Section 4.3, since each patient is represented by exactly one temporal sequence, the train–test splits of the cross-validation inherently occur at the patient level, thereby preventing any data leakage. The hyperparameter space explored is summarized in Table B.1 of the Supplementary Material.

Once the optimal hyperparameters were identified for each K, the RVA was trained $M = 50$ times to account for stochastic variability and dataset complexity. Consensus clustering was then applied to the resulting partitions to obtain a robust final clustering solution, as described in Section 3.3.

5. Results

This section presents the study findings. Simulation results, reported in Section A.2 of the Supplementary Material, show consensus clustering consistently outperforms single-model predictions across noise levels. The following sections describe the outcomes obtained using real clinical data from the PRO-ACT dataset.

5.1. Identification of the optimal number of clusters

Following the preprocessing described in Section 4.3, the clustering framework was applied to the final cohort of 3076 ALS patients. The results are shown in Table 4, which reports the internal validation metrics across different values of K.

To identify the optimal number of clusters, Student's t-tests were performed comparing $K = 3$ and $K = 5$, the two configurations showing the best overall performance. Statistical significance was determined at p-value < 0.05. The analysis revealed complementary strengths: while $K = 3$ achieves higher internal cohesion (Silhouette score: p-value < 0.05), $K = 5$ demonstrates significantly superior cluster separation

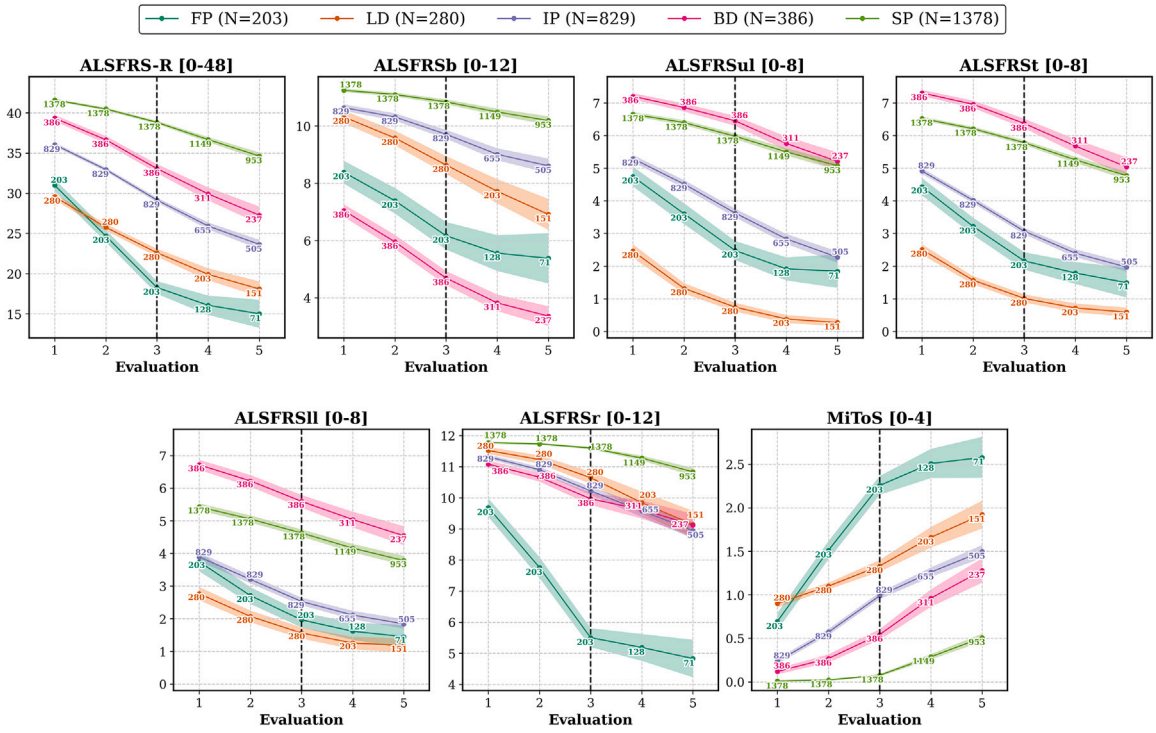


Fig. 2. Disease progression of ALS patient groups obtained, by functional domain (range between brackets) and MiToS. Patient groups: FP = Fast Progressor, LD = Limb-dominant, IP = Intermediate Progressor, BD = Bulbar-dominant, SP = Slow Progressor. Lines represent average temporal feature trajectories for which no imputation of missing values was performed, and shaded areas correspond to the 95% confidence interval. Note that the black dashed line indicates which data were used for algorithm development, corresponding to the first three evaluations.

(Davies–Bouldin index: p-value < 0.05). The Calinski–Harabasz index shows no significant difference (p-value > 0.05).

$K = 5$ was finally selected based on four key considerations. First, it achieves the best Davies–Bouldin index, indicating optimal balance between cluster separation and compactness. Second, $K = 5$ exhibits the lowest variability across metrics, demonstrating superior stability, a critical requirement for clinical applications where reproducibility is essential. Third, the Consensus Silhouette score reaches its maximum at $K = 5$, confirming robust agreement across multiple clustering runs. Fourth, from a clinical perspective, the higher resolution provided by five clusters enables the identification of clinically relevant patient subgroups that would remain undetected with $K = 3$, offering finer stratification for personalized treatment strategies and prognostic assessment. Although the Silhouette score decreases marginally from $K = 3$ to $K = 5$, this trade-off is justified by the gains in stability, separation quality, and clinical granularity. The optimal hyperparameters for $K = 5$ are:

$$n_{\text{hidden layers}} = 2, n_{\text{neurons}} = [64, 2], \eta = 10^{-4}, \text{Batch size} = 64$$

5.2. Progression profiles

Analyzing the obtained clusters, the 5 patient subgroups effectively reveal markedly different functional profiles. Fig. 2 shows the trajectories of the ALSFRS-R score, subscores, and MiToS over the first 12 months of observation, computed by averaging the individual trajectories of patients assigned to the same cluster by the algorithm. It is crucial to note that the algorithm exclusively utilized the first 3 longitudinal points (spanning the initial 6 months) to compute the cluster assignments. The subsequent data points (months 6 to 12) were not seen by the model during training or clustering; they are plotted purely for visualization purposes to demonstrate how the identified clinical phenotypes continue to diverge over time.

The first cluster (N = 203) corresponds to the most aggressive phenotype, with relatively low baseline ALSFRS-R score and the steepest global decline over follow-up; respiratory function drops early and sharply, paralleled by consistent losses in limb and trunk subscores, and patients rapidly progress to higher MiToS stages. Based on this rapid progression across all domains, this subgroup is designated as Fast Progressor (FP). The second cluster (N = 280) also starts from a markedly impaired profile but is characterized by a specific pattern: upper-limb, lower-limb and trunk subscores are already low at baseline and approach floor values over time, whereas bulbar and especially respiratory scores are comparatively preserved. Given this predominant impairment in limb function, this subgroup is designated as Limb-dominant (LD). The third cluster (N = 829) displays an intermediate baseline ALSFRS-R score level and follows a moderate decline that is more pronounced in limb and trunk domains compared to bulbar and respiratory functions. This limb-accentuated pattern is milder than the severe limb-dominant profile of LD group and distinctly more progressive than the stable course observed in the mildest subgroup. Based on this moderate progression rate, this cluster is designated as Intermediate Progressor (IP). The fourth cluster (N = 386) is characterized by predominantly bulbar involvement, while the other domains show only mild decline, and is therefore designated as Bulbar-dominant (BD). Finally, the fifth cluster (N = 1378) represents the mildest phenotype, with the highest ALSFRS-R score and near-normal subscores in all domains and only modest declines over the observation period, mirrored by minimal progression in MiToS stage. Given this favorable progression profile, this subgroup is designated as Slow Progressor (SP) and shows a milder phenotype compared to the IP group.

5.3. Baseline clinical features

To characterize the clusters identified from longitudinal data, features at disease onset and at 1st visit were examined to determine which were most discriminative between the resulting patient subgroups (see Table 5). The Chi-Square two-sided test was applied to categorical variables (sex and disease onset site) to assess differences

Table 5
Disease onset and 1st visit characterization. For categorical variables, values represent number of patients and percentage within cluster. For continuous variables, the first row shows mean, the second row shows 95% confidence interval [lower; upper], and the third row indicates number of patients with available data. P-values are from Chi-square two-sided tests (categorical variables) and Kruskal–Wallis two-sided tests (continuous variables), assessing for differences across all clusters. Patient groups: FP = Fast Progressor, LD = Limb-dominant, IP = Intermediate Progressor, BD = Bulbar-dominant, SP = Slow Progressor.

Characteristic	FP N = 203	LD N = 280	IP N = 829	BD N = 386	SP N = 1378	*p*-value
Sex						1.06×10^{-9}
Male	126, 62.1%	173, 61.8%	519, 62.6%	189, 49.0%	937, 68.0%	
Female	77, 37.9%	107, 38.2%	310, 37.4%	197, 51.0%	441, 32.0%	
Onset site						0
Limb	56, 46.7%	106, 69.7%	356, 71.9%	31, 12.4%	609, 69.8%	
Bulbar	36, 30.0%	11, 7.2%	42, 8.5%	207, 83.1%	97, 11.1%	
Other	28, 23.3%	35, 23.0%	97, 19.6%	11, 4.4%	167, 19.1%	
Age at onset (years)	57.28 [55.82;58.73] 201	52.21 [50.84;53.57] 273	54.28 [53.50;55.06] 813	59.04 [58.07;60.01] 372	53.75 [53.13;54.35] 1327	1.24×10^{-17}
Diagnostic Delay (months)	10.29 [9.16;11.41] 178	9.70 [8.87;10.52] 259	10.98 [10.38;11.57] 763	10.78 [9.90;11.65] 353	12.05 [11.56;12.54] 1267	7.31×10^{-6}
ALSFRS-R at 1st visit	30.99 [30.27;31.70] 203	29.57 [29.17;29.97] 280	36.05 [35.80;36.29] 829	39.34 [38.95;39.73] 386	41.61 [41.45;41.77] 1278	0
Slope ALSFRS-R at 6 months	−2.12 [-2.27;−1.96] 203	−1.16 [-1.25;−1.06] 280	−1.14 [-1.20;−1.09] 829	−1.04 [-1.12;−0.96] 386	−0.46 [-0.49;−0.43] 1278	0
FVC% at 1st visit	75.56 [72.99;78.13] 111	81.68 [79.23;84.13] 140	85.94 [84.53;87.35] 469	83.68 [81.59;85.77] 229	92.94 [91.86;94.01] 856	0
Baseline creatinine (μmol/L)	63.34 [60.52;66.16] 174	55.59 [53.51;57.67] 249	62.33 [61.03;63.63] 729	70.88 [69.12;72.64] 319	71.49 [70.45;72.52] 1178	0

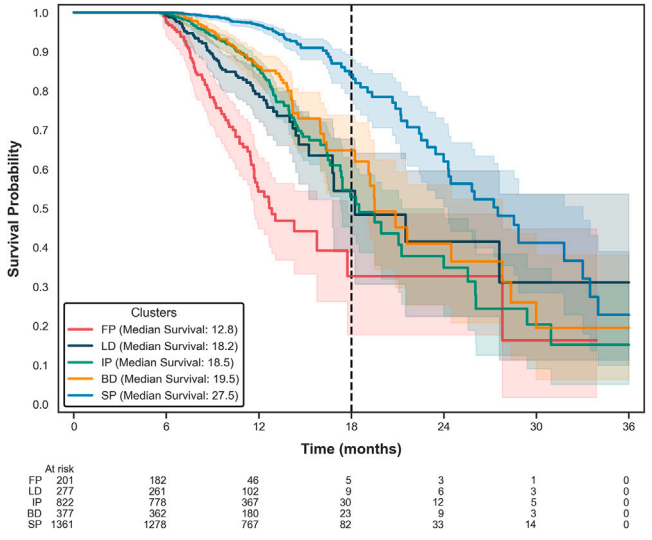


Fig. 3. Kaplan–Meier survival curves for the five patient subgroups over a 36-month follow-up period. The legend reports the median survival time (in months) for each subgroup. Shaded areas represent the 95% confidence intervals. The dashed vertical line at 18 months indicates the threshold beyond which estimates should be interpreted with caution due to the limited number of remaining subjects. The table below the horizontal axis displays the number of patients at risk for each cluster at specific time points. Patient clusters: FP = Fast Progressor, LD = Limb-dominant, IP = Intermediate Progressor, BD = Bulbar-dominant, SP = Slow Progressor.

in their distribution across clusters. For continuous features, normality was evaluated using the Kolmogorov–Smirnov two-sided test within each cluster. Since age at onset, diagnostic delay, slope of ALSFRS-R at 6 months, ALSFRS-R at 1st visit, Forced Vital Capacity percent predicted (FVC%) at 1st visit, and baseline creatinine, a proxy for global muscle mass [37], did not consistently follow a normal distribution across all clusters, we applied the non-parametric Kruskal–Wallis two-sided test to identify significant differences across clusters. Pairwise comparisons were then performed using Bonferroni correction. Statistical significance was determined at a *p*-value threshold of 0.05. The continuous features exhibited varying levels of significance when comparing different cluster pairs:

- Age at disease onset: FP differed from LD, IP, SP (p-value < 0.05), while BD showed significant differences compared to LD, IP, SP (p-value < 0.05).
- Diagnostic delay: SP differed from FP, LD, IP (p-value < 0.05).
- FVC% at 1st visit: no differences observed between BD and LD, IP (p-value > 0.05).
- ALSFRS-R at 1st visit: all pairwise combinations showed significant differences.
- Slope of ALSFRS-R score at 6 months: LD, IP, and BD showed no mutual differences (p-value > 0.05).
- Baseline creatinine levels: FP and IP, as well as BD and SP, showed no differences (p-value > 0.05).

5.4. Survival analysis

To evaluate the proposed stratification, Kaplan–Meier survival curves were generated for a 36-month follow-up (Fig. 3). However, due to a substantial decrease in patients at risk after 18 months, formal statistical testing was restricted to an 18-month interval. Pairwise log-rank tests were evaluated at a 0.05 significance level, with the Bonferroni correction applied for multiple comparisons.

The analysis revealed distinct survival patterns among the subgroups. The extreme phenotypes were highly distinct: the FP cluster exhibited the most pronounced decline in survival (median survival: 12.8 months; p-value < 0.001 vs. others), whereas the SP cluster

maintained significantly higher survival rates (median survival: 27.5 months; p-value < 0.001 vs. others). In contrast, the LD, IP, and BD groups followed intermediate, frequently overlapping trajectories with no statistically significant differences observed among them.

6. Discussion

By applying the proposed deep learning framework to the PRO-ACT dataset, we identified five clinically meaningful subgroups that capture the highly heterogeneous nature of ALS progression. Crucially, these data-driven clusters align closely with established clinical phenotypes, demonstrating that the developed methodology successfully reproduces and refines known subtypes [38].

Cluster FP represents the most severe group, characterized by aggressive progression (ALSFRS-R decline of −2.12/month), rapid early respiratory deterioration and the shortest median survival (12.8 months). These patients present with low baseline ALSFRS-R scores (30.99) and FVC% (75.56%), alongside low creatinine (63.34 μmol/L). This profile strongly mirrors the established respiratory phenotype, which is recognized in the literature for its highly aggressive course, early ventilatory failure, and extremely poor prognosis [38,39]. In contrast, cluster SP characterizes the mildest disease form, exhibiting the slowest decline (−0.46/month) and longest median survival (27.5 months). Due to minimal early functional decline, SP patients experience the longest diagnostic delay (12.05 months). This group, characterized by a younger age at onset (mean 53.75 years) and male predominance (68.0%), maintains the highest baseline ALSFRS-R (41.61), FVC% (92.94%), and creatinine (71.49 μmol/L), reflecting preserved muscle mass and respiratory function. This group's slow and proportional functional decline across all clinical domains, coupled with the younger age at onset, aligns with the slowly progressive variant of the classic spinal phenotype [40,41].

Between these extremes, three intermediate clusters (LD, BD, and IP) display distinct domain involvement but share comparable survival times, baseline respiratory functions, and overall decline rates. This suggests their differences lie primarily in the anatomical distribution of motor neuron loss rather than the overall rate of neurodegeneration. Cluster LD shows severe limb-predominant involvement, presenting the lowest baseline ALSFRS-R (29.57) and creatinine (55.59 μmol/L), indicating substantial peripheral muscle wasting [38]. Specifically, this marked and isolated limb impairment reflects lower motor neuron (LMN) predominant syndromes (such as flail arm or flail leg phenotypes), which are characterized by severe peripheral involvement and sustained preservation of vital functions [38,41]. Cluster BD demonstrates a bulbar-dominant phenotype; aligning with known bulbar-onset ALS demographics [38,40], it features the highest bulbar-onset proportion (83.1%), highest female prevalence (51.0%), and oldest onset age (59.04 years). Finally, cluster IP exhibits moderate progression across all domains and low creatinine levels (62.33 μmol/L), representing a clinical profile milder than LD but more aggressive than SP, which corresponds to the typical classic spinal phenotype with an intermediate disease course [41].

In summary, the developed methodology demonstrates the capacity to recover recognized clinical phenotypes (bulbar, respiratory, LMN-predominant) while refining the classic spinal presentation into precise, data-driven severity categories (IP and SP). This convergence between unsupervised algorithmic stratification and classical clinical taxonomy significantly strengthens confidence in the validity of the cluster assignments and highlights their potential translational impact.

Combining ALSFRS-R subscores with the total score and MiToS staging was essential to capture the multidimensional nature of ALS progression. Unlike previous approaches relying on single-timepoint assessments or assuming predefined linear trajectories [7], our RVA framework learns complex temporal dependencies across multiple features. This enables the identification of non-linear progression patterns and multivariate interactions that traditional staging systems or parametric models may miss, revealing distinct trajectories even among patients who appear similar when evaluated on individual features.

Simulation experiments provide critical validation of the methodology's robustness. The consensus clustering approach demonstrated superior performance compared to single-model predictions, maintaining high cluster purity even under substantial noise conditions (see Table A.2 of the Supplementary Material). This robustness is particularly relevant for clinical applications where measurement errors and natural inter-patient variability are common challenges. Crucially, the method maintained reliable performance on noisy data despite being trained under optimal conditions, demonstrating strong generalization capabilities.

To evaluate the relative performance of the RVA approach against recently developed pattern-mining methodologies, a direct empirical comparison was conducted with the state-of-the-art triclustering framework (ClusTric) by Amaral et al. [21]. The comprehensive analysis is reported in Appendix C of the Supplementary Material. When evaluated on the exact same cohort, RVA optimally identifies five distinct subgroups, providing a more comprehensive characterization of ALS heterogeneity than the optimal partition found for ClusTric (three clusters). A head-to-head clinical comparison at an equivalent granularity (five clusters) demonstrates that RVA extracts more clinically representative profiles, successfully isolating true fast progressors with rapid respiratory decline and a cohesive bulbar-dominant phenotype, both of which ClusTric fails to accurately detect. Furthermore, survival analysis reveals that RVA achieves superior prognostic resolution by stratifying the patient cohort into three statistically distinct risk profiles (high, intermediate, and low risk), whereas ClusTric's partitions effectively collapse into only two statistically distinct macro-groups. Overall, these analyses empirically suggest that the proposed deep learning-based approach offers greater phenotypic resolution and a potentially greater clinical benefit for the stratification of patients with ALS. For an in-depth quantitative and clinical discussion of this benchmark, the reader is referred to Appendix C of the Supplementary Material.

Beyond these empirical clinical findings, the developed framework introduces several methodological benefits. First, the approach requires substantially less data preprocessing, operating on minimally preprocessed data with only standard baseline normalization. Second, the end-to-end learning method offers a fundamental architectural advantage: once hyperparameters are selected, all model parameters, including the encoder weights, decoder weights, and Gaussian mixture components, are jointly optimized through backpropagation to maximize a single unified objective function (the ELBO). This joint optimization ensures that the learned latent representations are specifically tailored to produce well-separated, clinically meaningful clusters. In contrast, ClusTric employs a sequential pipeline where triclustering and hierarchical clustering operate as independent modules with separate optimization objectives. Since these stages are optimized independently, there is no guarantee that the patterns optimal for triclustering homogeneity are also optimal for patient stratification, potentially resulting in suboptimal final clustering solutions. Third, the recurrent nature of the encoder–decoder architecture makes it possible to predict future performance decline; although we have not explored this application of the proposed approach in this work, it undoubtedly represents a highly promising direction for the further development and application of the method. By contrast, such predictive capability is inherently absent in traditional modeling approaches, which are limited to the retrospective identification of patterns. Finally, the consensus clustering approach provides uncertainty quantification by aggregating multiple model runs: the co-association matrix naturally captures the stability of patient assignments, where patients consistently clustered together across runs indicate high confidence, while variable assignments reflect boundary cases requiring closer clinical attention. This advantage is missing from deterministic single-run methods like ClusTric.

This study has several limitations that we have to take into consideration. First, the PRO-ACT dataset, while comprehensive, represents a clinical trial population that may not fully reflect the broader ALS population, as evidenced by the younger age, slower progression rates, and under-representation of certain phenotypes compared to hospital-based cohorts [42]. Second, the analysis focused on ALSFRS-R-based features only, but incorporating additional biomarkers, genetic data, or neuroimaging could potentially improve stratification accuracy. Third, while simulation experiments demonstrate robustness to noise, the simplified synthetic data may not capture all sources of variability present in real clinical settings. Fourth, the assumption that patients remain in their assigned cluster throughout disease progression may not hold for all individuals, as some patients may exhibit phenotypic transitions as the disease evolves.

Future research should address these limitations by considering complementary approaches. First, validation studies in diverse hospital-based cohorts are needed to establish the generalizability of these findings beyond clinical trial populations. Second, incorporating multimodal data including fluid biomarkers, neurophysiological measures, and genetic markers could enhance the precision of phenotypic stratification. Third, longitudinal analysis examining cluster stability over time would help determine whether patients maintain consistent phenotypic characteristics throughout disease progression. Finally, prospective validation studies should evaluate whether this stratification approach can effectively guide clinical decision-making and improve patient outcomes.

7. Conclusion

In this work, we have developed and validated a novel deep learning framework for stratifying ALS patients based on longitudinal disease progression patterns. By combining a Recurrent Variational Autoencoder with consensus clustering, the approach successfully identified five clinically distinct subgroups from routine ALSFRS-R assessments collected over six months. These subgroups demonstrated significantly different functional trajectories, baseline characteristics, and survival outcomes, consistent with established ALS phenotypes described in medical literature, thus confirming the clinical relevance of our stratification methodology.

The key contributions of this work include: (1) the first application of RVA to ALS patient stratification, demonstrating the potential of deep learning to capture complex, multivariate, non-linear disease progression patterns; (2) a robust hyperparameter optimization strategy based on prediction strength and two-fold cross-validation, providing a reliable solution for hyperparameter selection in unsupervised settings where ground truth is unavailable; (3) a robust consensus clustering framework that maintains high accuracy even under challenging noise conditions; (4) the identification of clinically meaningful patient subgroups using only routine clinical assessments, making the approach practical for real-world implementation; and (5) comprehensive validation using both simulation experiments and analyses of real-world clinical data.

The findings highlight the heterogeneous nature of ALS progression and the importance of data-driven approaches for precision medicine. The identified subgroups provide a foundation for personalized prognosis, optimized clinical trial design, and tailored therapeutic strategies. As we advance towards more individualized ALS care, this stratification framework offers a valuable tool for improving patient management and accelerating the development of effective treatments. Future integration with emerging biomarkers and validation across diverse populations will further enhance the clinical utility of this approach, ultimately contributing to better outcomes for individuals living with ALS.

CRediT authorship contribution statement

Federico De Mori Bajolin: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Erica Tavazzi:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Conceptualization. **Anna M. Bianchi:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Conceptualization. **Martin O. Mendez:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research of Federico De Mori Bajolin is supported by a PNRR Cycle 39 scholarship (nr. 4556, I.4.1 PNRR Public Administration).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2026.105059.

Data availability

Data are publicly available. Access to the dataset can be obtained by submitting a request directly to the PRO-ACT platform.

References

[1] L.C. Wijesekera, P.N. Leigh, Amyotrophic lateral sclerosis, Orphanet J. Rare Dis. 4 (2009) 3, http://dx.doi.org/10.1186/1750-1172-4-3.

[2] J. Morris, Amyotrophic lateral sclerosis (ALS) and related motor neuron diseases: An overview, Neurodiagnostic J. 55 (3) (2015) 180–194, http://dx.doi.org/10.1080/21646821.2015.1075181.

[3] S.R. Pfohl, R.B. Kim, G.S. Coan, C.S. Mitchell, Unraveling the complexity of amyotrophic lateral sclerosis survival prediction, Front. Neuroinformatics 12 (2018) 36, http://dx.doi.org/10.3389/fninf.2018.00036.

[4] J.M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function, J. Neurol. Sci. 169 (1) (1999) 13–21, http://dx.doi.org/10.1016/S0022-510X(99)00210-5, URL https://www.sciencedirect.com/science/article/pii/S0022510X99002105.

[5] R.S. Bedlack, T. Vaughan, P. Wicks, J. Heywood, E. Sinani, R. Selsov, E.A. Macklin, D. Schoenfeld, M. Cudkowicz, A. Sherman, How common are ALS plateaus and reversals? Neurology 86 (9) (2016) 808–812, http://dx.doi.org/10.1212/WNL.0000000000002251.

[6] S. Pires, M. Gromicho, S. Pinto, M. de Carvalho, S.C. Madeira, Patient stratification using clinical and patient profiles: Targeting personalized prognostic prediction in ALS, in: I. Rojas, O. Valenzuela, F. Rojas, L.J. Herrera, F. Ortuño (Eds.), Bioinformatics and Biomedical Engineering, Springer International Publishing, Cham, 2020, pp. 529–541.

[7] E. Tavazzi, et al., Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review, Artif. Intell. Med. 142 (2023) 102588, http://dx.doi.org/10.1016/j.artmed.2023.102588, URL https://www.sciencedirect.com/science/article/pii/S0933365723001021.

[8] D. Ramamoorthy, K. Severson, S. Ghosh, K. Sachs, Answer ALS, J.D. Glass, C.N. Fournier, Pooled Resource Open-Access ALS Clinical Trials Consortium, ALS/MND Natural History Consortium, T.M. Herrington, J.D. Berry, K. Ng, E. Fraenkel, Identifying patterns in amyotrophic lateral sclerosis progression from sparse longitudinal data, Nat. Comput. Sci. 2 (9) (2022) 605–616, http://dx.doi.org/10.1038/s43588-022-00299-w.

[9] S. Pires, M. Gromicho, S. Pinto, M. Carvalho, S.C. Madeira, Predicting non-invasive ventilation in ALS patients using stratified disease progression groups, in: 2018 IEEE International Conference on Data Mining Workshops, ICDMW, 2018, pp. 748–757, http://dx.doi.org/10.1109/ICDMW.2018.00113.

[10] M. Gromicho, T. Leão, M. Oliveira Santos, S. Pinto, A.M. Carvalho, S.C. Madeira, M. De Carvalho, Dynamic Bayesian networks for stratification of disease progression in amyotrophic lateral sclerosis, Eur. J. Neurol. 29 (8) (2022) 2201–2210, http://dx.doi.org/10.1111/ene.15357.

[11] J.C. Roche, R. Rojas-Garcia, K.M. Scott, W. Scotton, C.E. Ellis, R. Burman, L. Wijesekera, M.R. Turner, P.N. Leigh, C.E. Shaw, A. Al-Chalabi, A proposed staging system for amyotrophic lateral sclerosis, Brain 135 (3) (2012) 847–852, http://dx.doi.org/10.1093/brain/awr351.

[12] A. Chiò, E.R. Hammond, G. Mora, V. Bonito, G. Filippini, Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis, J. Neurol. Neurosurg. & Psychiatry 86 (1) (2015) 38–44, http://dx.doi.org/10.1136/jnnp-2013-306589.

[13] N.J. Thakore, B.R. Lapin, T.G. Kinzy, E.P. Pioro, Deconstructing progression of amyotrophic lateral sclerosis in stages: a Markov modeling approach, Amyotroph. Lateral Scler. Front. Degener. 19 (7–8) (2018) 483–494, http://dx.doi.org/10.1080/21678421.2018.1484925, PMID: 30001159.

[14] R. Balendra, A. Jones, N. Jivraj, C. Knights, C.M. Ellis, R. Burman, M.R. Turner, P.N. Leigh, C.E. Shaw, A. Al-Chalabi, Estimating clinical stage of amyotrophic lateral sclerosis from the als functional rating scale, Amyotroph. Lateral Scler. Front. Degener. 15 (3–4) (2014) 279–284, http://dx.doi.org/10.3109/21678421.2014.897357, PMID: 24720420.

[15] I. Tramacere, E. Dalla Bella, A. Chiò, G. Mora, G. Filippini, G. Lauria, The MITOS system predicts long-term survival in amyotrophic lateral sclerosis, J. Neurol. Neurosurg. & Psychiatry 86 (11) (2015) 1180–1185, http://dx.doi.org/10.1136/jnnp-2014-310176.

[16] R. Kueffner, N. Zach, M. Bronfeld, et al., Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach, Sci. Rep. 9 (2019) 690, http://dx.doi.org/10.1038/s41598-018-36873-4.

[17] R. Gomeni, M.F. and, Amyotrophic lateral sclerosis disease progression model, Amyotroph. Lateral Scler. Front. Degener. 15 (1–2) (2014) 119–129, http://dx.doi.org/10.3109/21678421.2013.838970, PMID: 24070404.

[18] M.-L. Ong, P.F. Tan, J.D. Holbrook, Predicting functional decline and survival in amyotrophic lateral sclerosis, PLOS ONE 12 (2017) e0174925, http://dx.doi.org/10.1371/journal.pone.0174925.

[19] H.-J. Westeneng, et al., Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model, Lancet Neurol. 17 (5) (2018) 423–433, http://dx.doi.org/10.1016/S1474-4422(18)30089-9.

[20] R. Steinbach, N. Gaur, A. Roediger, T.E. Mayer, O.W. Witte, T. Prell, J. Grosskreutz, Disease aggressiveness signatures of amyotrophic lateral sclerosis in white matter tracts revealed by the D50 disease progression model, Hum. Brain Mapp. 42 (3) (2021) 737–752, http://dx.doi.org/10.1002/hbm.25258.

[21] D.M. Amaral, D.F. Soares, M. Gromicho, M. de Carvalho, S.C. Madeira, P. Tomás, H. Aidos, Temporal stratification of amyotrophic lateral sclerosis patients using disease progression patterns, Nat. Commun. 15 (5717) (2024) http://dx.doi.org/10.1038/s41467-024-49954-y.

[22] D.F. Soares, R. Henriques, M. Gromicho, M. de Carvalho, S.C. Madeira, Triclustering-based classification of longitudinal data for prognostic prediction: targeting relevant clinical end points in amyotrophic lateral sclerosis, Sci. Rep. 13 (2023) 6182.

[23] J. de Jong, M.A. Emon, P. Wu, R. Karki, et al., Deep learning for clustering of multivariate clinical patient trajectories with missing values, GigaScience 8 (11) (2019) giz134, http://dx.doi.org/10.1093/gigascience/giz134, URL https://doi.org/10.1093/gigascience/giz134.

[24] F.A. Gers, N.N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, J. Mach. Learn. Res. 3 (2002) 115–143, URL https://www.jmlr.org/papers/volume3/gers02a/gers02a.pdf.

[25] Z. Jiang, Y. Zheng, H. Tan, B. Tang, H. Zhou, Variational deep embedding: An unsupervised and generative approach to clustering, 2017, arXiv:1611.05148.

[26] M.P. Roque, A.S. Martins, M. Gromicho, M. de Carvalho, et al., Deep temporal consensus clustering for patient stratification in amyotrophic lateral sclerosis, ESANN 2024 Proc. (2024).

[27] Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, Learning to diagnose with LSTM recurrent neural networks, 2016, arXiv preprint arXiv:1511.03677.

[28] I.M. Baytas, C. Xiao, X. Zhang, F. Wang, A.K. Jain, J. Zhou, Patient subtyping via time-aware LSTM networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, ACM, New York, NY, USA, 2017, pp. 65–74, http://dx.doi.org/10.1145/3097983.3097997.

[29] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, 2022, arXiv:1312.6114.

[30] R. Tibshirani, G.W. and, Cluster validation by prediction strength, J. Comput. Graph. Statist. 14 (3) (2005) 511–528, http://dx.doi.org/10.1198/106186005X59243, URL https://doi.org/10.1198/106186005X59243.

[31] A.L. Fred, A.K. Jain, Combining multiple clusterings using evidence accumulation, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 835–850, http://dx.doi.org/10.1109/TPAMI.2005.113.

[32] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, http://dx.doi.org/10.1016/0377-0427(87)90125-7, URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[33] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (2) (1979) 224–227, http://dx.doi.org/10.1109/TPAMI.1979.4766909.

[34] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Comm. Statist. Theory Methods 3 (1974) 1–27, URL https://api.semanticscholar.org/CorpusID:122217223.

[35] Prize4Life, Pooled Resource Open-Access ALS Clinical Trials Database, https://ncri1.partners.org/ProACT/Home/Index.

[36] N. Atassi, J. Berry, A. Shui, N. Zach, A. Sherman, E. Sinani, J. Walker, I. Katsovskiy, D. Schoenfeld, M. Cudkowicz, M. Leitner, The PRO-ACT database, Neurology 83 (19) (2014) 1719–1725, http://dx.doi.org/10.1212/WNL.0000000000000951.

[37] R.P.A. van Eijk, et al., Monitoring disease progression with plasma creatinine in amyotrophic lateral sclerosis clinical trials, J. Neurol. Neurosurg. & Psychiatry 89 (2) (2018) 156–161, http://dx.doi.org/10.1136/jnnp-2017-317077.

[38] A. Chiò, A. Calvo, C. Moglia, L. Mazzini, G. Mora, P. study group, Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study, J. Neurol. Neurosurg. & Psychiatry 82 (7) (2011) 740–746, http://dx.doi.org/10.1136/jnnp.2010.235952.

[39] G. Gautier, A. Verschueren, A. Monnier, S. Attarian, E. Salort-Campana, J. Pouget, ALS with respiratory onset: Clinical features and effects of non-invasive ventilation on the prognosis, Amyotroph. Lateral Scler. 11 (4) (2010) 379–382, http://dx.doi.org/10.3109/17482960903426543, PMID: 20001486.

[40] A. Chiò, C. Moglia, A. Canosa, et al., ALS phenotype is influenced by age, sex, and genetics: A population-based study, Neurology 94 (8) (2020) e802–e810, http://dx.doi.org/10.1212/WNL.0000000000008869.

[41] P. Couratier, G. Lautrette, J. Luna, P. Corcia, Phenotypic variability in amyotrophic lateral sclerosis, Rev. Neurol. 177 (5) (2021) 536–543, http://dx.doi.org/10.1016/j.neurol.2021.03.001, URL https://www.sciencedirect.com/science/article/pii/S0035378721005051, SFN 2020.

[42] A. Chiò, A. Canosa, S. Gallo, S. Cammarosano, C. Moglia, G. Fuda, et al., ALS clinical trials: do enrolled patients accurately represent the ALS population? Neurology 77 (2011) 1432–1437.