A Deep Temporal Consensus Clustering approach for Amyotrophic Lateral Sclerosis Patient Stratification

Miguel Pego Roque
miguelroque99@tecnico.ulisboa.pt
Instituto Superior Técnico, University of Lisbon
Lisbon, Portugal

ABSTRACT

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease that affects upper and lower motor neurons. Currently without a cure, it leads to loss of voluntary muscle movement resulting, in most cases, in death from respiratory failure within 2 to 4 years after the initial diagnosis. Non-invasive ventilation (NIV) is the most effective treatment to increase life expectancy. However, the efficacy of NIV strongly depends on the timing of the beginning of treatment. Due to the many clinical presentations of the disease, this timing is not trivial to determine. Stratification can aid this task by dividing the patients into clinically relevant subgroups according to the disease progression. The prediction of the need for NIV can then be tuned for each sub-group to achieve better performance. Inspired by state-of-the-art approaches, this work introduces a novel patient stratification method, Deep Temporal Encoded Clustering (DTEC), that takes advantage of the temporal dimension of the data to find clinically relevant patients' subgroups. Considering only the disease beginning and using promising Deep Learning mechanisms for temporal encoding, a UMAP encoder, Hierarchical Clustering, and a Consensus Clustering methodology, DTEC found subgroups that present differences across disease progression and clinical presentation. A classification pipeline was also developed and applied to each subgroup. The results were compared with the pipeline without stratification. The classifiers showed improvements, reaching values of AUC up to 83% for some subgroups. Although limited by the dataset, the results showcased the potential of a personalized prognostic prediction solution using patient stratification in the ALS context.

KEYWORDS

Temporal Stratification, Deep Learning, Consensus Clustering, Prognostic Prediction, Amyotrophic Lateral Sclerosis

1 INTRODUCTION

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease, currently without a known cure. It is characterized by a progressive degeneration of motor neurons, leading to the loss of muscle movement and, eventually, death. The life expectancy after the first symptoms varies from 2 to 4 years [6]. The most common treatments available are four approved drugs [2], which increase life expectancy in a few months [31], and non-invasive ventilation (NIV), which may increase life expectancy by over a year, depending on the starting time and clinical presentation [31]. NIV relevance is even higher when considering that the most common cause of death is due to respiratory failure.

This disease has many clinical presentations, varying not only in symptoms but also in disease progression speed. This fact leads to an increased challenge when understanding the stage of the disease and its evolution, which are crucial factors in deciding whether the patient is eligible for NIV treatment and when to start it. Therefore, a proper stratification of ALS patients is crucial to understand the disease progression and enhance the timing of the NIV start.

Although a few works on ALS patient stratification already exist, they primarily focus on clinical or traditional Machine Learning approaches [8, 22, 24] without taking full advantage of the temporal data present in ALS patient records. Most of these approaches use clinical criteria as stratification criteria and then employ traditional Machine Learning algorithms for the classification task.

Deep Learning methods for ALS patient stratification are still under-explored, despite its success in other domains and diseases [4, 17, 21, 33]. The nonlinearity of these models might offer new insights and help untangle different subgroups of patients with their own needs. By achieving a better stratification, it is possible to characterize the disease in subgroups of patients in order to uncover the underlying disease patterns. Furthermore, the prognosis can also be improved for a subgroup of patients, and personalized treatment can be offered. This specialized treatment can help improve the daily life of patients suffering from this disease and even extend their life expectancy.

In addition, Deep Learning techniques have been used to encode temporal information, which has shown to be of great value for patient stratification and prognosis in other diseases [17, 21, 28, 33]. The encoding of the temporal dimension allows a deeper exploration of the relations regarding the progression of a patient and its expected outcome. This can help determine the best starting time for NIV, ensuring a properly timed treatment and preventing a more exponential and uncontrolled decline of respiratory capacity.

The advances in these areas present a great extent of untapped potential for ALS patient stratification. Specifically, the breakthroughs in stratification using Deep Learning mechanisms offer the possibility to create methods that take advantage of the temporal dimension of the data to build clinically relevant subgroups and produce quality prognoses.

This work aims to explore temporal stratification for ALS patients, specifically using Deep Learning techniques, to characterize underlying subgroups and improve patient prognosis. Using the Portuguese ALS Dataset, it strives to build clinically relevant subgroups of patients, characterize them, and improve the prediction of disease evolution and the consequent need of NIV for these patients.

This work seeks to integrate recent relevant Deep Learning techniques to build a novel method for ALS patient stratification that

takes advantage of the temporal evolution of patients' records. We hypothesize that using this information will unveil underlying patterns of disease progression, providing a better characterization of the disease and contributing to a personalized prognosis prediction.

We aim to take a step forward in improving the quality of life of people suffering from this disease by providing clinicians with relevant insights on the different patients' subgroups that can help them decide the best treatment.

The main contributions achieved with this work include the following:

- A literature review on state-of-the-art patient stratification and prognosis prediction methods with emphasis on temporal methods.
- A novel patient stratification method, Deep Temporal Encoded Clustering, specifically tailored for the Portuguese ALS Dataset context, which unveiled distinct patient subpopulations inside our dataset.
- A detailed characterization of the different subgroups of patients obtained, regarding both disease presentation and evolution, highlighting the main clinically relevant differences between them.
- A classification pipeline for ALS patients which showed evidence of the improvement of the prognosis prediction task when applied separately to subgroups of ALS patients with similar characteristics.

2 RELATED WORK

When discussing ALS patient stratification, it is essential to mention the methods developed based on clinical criteria. These methods usually rely on a staging system, a scoring system that characterizes patients based on clinical characteristics. Two of the most relevant and recent staging systems applied to ALS are the King's ALS clinical staging system and the ALS Milano–Torino system (MiToS).

Although limited, clinical methods are a valid and relevant approach to predicting the course of ALS. In specific for MiToS, a recent study reported levels of sensibility and specificity of 82% and 63% at 12 months and 71%/68% at 18 months when predicting death, tracheotomy, or more than 23-hour NIV [29].

The growing applications of Machine Learning led to several new patient stratification and prognosis methods. Some of these methods achieved excellent results, supporting health professionals in decision-making.

One relevant work is the one of Singh et al. [28]. Their work implemented three different Machine Learning approaches for risk stratification of renal function deterioration.An interesting aspect of this work is the definition of time windows, concluding that incorporating temporal information in this context can improve the performance of the prediction models.

In the particular context of the Portuguese ALS Dataset, it is crucial to mention the work of Carreiro et al. [8], which first introduced the use of prognostic models based on time windows to predict a patient's need of NIV. A relevant aspect of its data preprocessing was the creation of snapshots through clustering of temporally-related medical exams in three different time windows (90, 180, and 365 days). The results were promising, outperforming classical approaches and pioneering in introducing a non-population-based approach for prognostic prediction in ALS.

Another work, built on top of the previous and relevant to mention, is the one of Martins et al. [22]. Their approach used the same time-windows, and consisted of the use of both itemset mining and sequential pattern mining to discover patterns in disease presentation and progression that can then be used as features to prognostic models.

Regarding stratification, traditional Machine Learning methods mostly use unsupervised techniques. The applications to patient stratification are straightforward since they allow the identification of different subtypes of patients, according to the similarity of their records, even when class labels are not available [27]. Additionally, Deep Learning models are increasingly present in several stratification applications, as well. They offer high potential for stratification tasks, mainly due to their non-linearity, obtained through the use of activation functions. Despite the generalized use of linear models for disease progression prediction, studies suggest that ALS progression can be non-linear, differing according to disease severity [25], increasing the relevance of Deep Learning applications in this context.

Moreover, Machine Learning stratification methods can be divided into temporal and non-temporal, considering the existence and exploitation of temporal features.Traditional non-temporal methods are based on unsupervised techniques. Several works, in other contexts besides ALS, used such algorithms: Li and Wong's evolutionary multiobjective clustering method for patient stratification [18]; Khakabimamaghani and Ester's bayesian biclustering method for comparing different patient stratification datasets [12]; the entropy-based consensus clustering method of Liu et al. [20].

Specifically for the Portuguese ALS Dataset, it is pertinent to mention Pires' work [24] on the matter of non-temporal methods for patient stratification and prognostic prediction. One of the key conclusions of this study was that using separate prognostic models for each subgroup of patients can improve the prediction results.

The use of deep-learning methods for patient stratification is present in several applications [3, 4, 14]. A work worth mentioning is the one of Lin et al. [19] where the "Skip-Gram" architecture (a method typically used in Natural Language Processing - NLP) is used for pediatric patient risk stratification. The authors concluded that the Deep Learning model outperforms the traditional methods.

In the ALS field, an interesting work using traditional Machine Learning models is the one of Berry et al. [5]. In this work, the authors compare traditional stratification with stratification using predicted survival in the context of small trials' randomization. The study concluded that the second method achieved better performance, highlighting the applicability to other neurological diseases and the potential of reducing the size of medical trials to obtain a more efficient representation.

Another relevant work in this context is the crowd-sourcing approach for the stratification of ALS patients of Kueffner et al. [15] By gathering the results obtained by the several approaches and forming a consensus clustering, this approach enables the identification of the best-defined subgroups of patients and the features that distinguish each.

Still concerning more traditional Machine Learning models explicitly applied to ALS, it is interesting to acknowledge the work

of Ramamoorthy et al. [26]. In order to find patterns in disease progression in ALS patients, the authors developed a method based on a Mixture of Gaussian Processes (MoGP). This approach allowed to produce a non-linear model able to obtain better performance than linear methods, possibly to the non-linearity of ALS progression. In another subsequent work, the authors experimented using Sequential Pattern Mining to extract features that encode temporal dependency [7]. One of the main challenges found in this dataset was the limited time entries for each patient, which negatively impacted the performance of the models implemented.

Regarding temporal Deep Learning approaches, Huang et al. [11] developed a Deep Learning approach to predict patients' lung cancer risk and unveil subgroups with similar risk.

Madiraju [21] proposed a deep temporal clustering technique, with promising results, that uses a temporal autoencoder to map the input temporal data into an effective latent representation, followed by a temporal clustering layer.

Furthermore, in the field of neurodegenerative diseases, Zhen et al. [33] developed a deep stratification network (DPS-Net) for Alzheimer's patients in the context of neuroimaging data.

Another key work in temporal stratification is the AC-TPC method by Lee et al. [17]. In this work, the authors propose a novel architecture that finds clusters and their centroids based not only on the similarity of the temporal observations but also on their future outcomes.

Another recent work worth mentioning is the one of Landi et al. [16], where the authors developed a Patient Stratification framework, for both multi-disease and disease-specific cohorts of patients. The results achieved were promising, identifying patients across several complex diseases and obtaining relevant subgroups of patients within a disease (distinguished by factors such as disease progression, comorbidities, and symptom severity).

Specifically for ALS, few works on temporal patient stratification using Deep Learning methods were found which represents both a challenge and a motivation for the work to be developed.

Nonetheless, one relevant work to mention is the one of van der Burgh et al. [30] regarding the use of Deep Learning techniques for survival prediction of ALS patients. The records used corresponded mostly to deceased or terminal patients labeled as short, medium, or long survivors. Important to note however, that the most relevant improvements were obtained only when joining information from magnetic resonance imaging (MRI) to the clinical data.

3 DATA AND DATA PREPROCESSING

The dataset used in this work is the Portuguese ALS Dataset. This dataset contains both demographic and clinical data for ALS patients followed at the ALS clinic of the Translational Clinic Physiology Unit, Hospital de Santa Maria, IMM, Lisbon.

Particularly in this work, we will use a version of the dataset after the preprocessing method described by Carreiro et al. [8] has been applied, to obtain patients snapshots, and considering the three different time-windows (90,180 and 365 days).

Because of the interest in exploring the data's temporal dimension and the patients' corresponding evolution, it is not relevant to consider patients with one or very few records. The threshold of at least 3 records was chosen, taking into consideration that a higher value would exclude more than half of the patients.

Furthermore, features with more than 60% of missing values are excluded, resulting in a total of 27 columns of the original dataset being used, representing both temporal and static variables, of several types (categorical, DateTime, and numerical - both discrete and continuous). These features include both Demographic data, Onset Evaluations, Respiratory Tests and ALS Functional Scores.

The common Data Preprocessing pipeline applied to the data, before being fed to the patient stratification and prognosis prediction methods, consists on three main stages: Missing Value Imputation (MVI); Encoding/Discretization of the categorical variables; and Normalization between [-1,1].

After the preprocessing steps, the dataset contains a total of 38 features, comprising data from a cohort of 561 ALS patients, totaling 3472 records, and with observations ranging from August 1996 to December 2021.

4 STRATIFICATION METHODOLOGY

The Deep Temporal Encoded Clustering method (DTEC) was developed to capture the temporal dimension of patients' records and use it to obtain medically relevant groups.

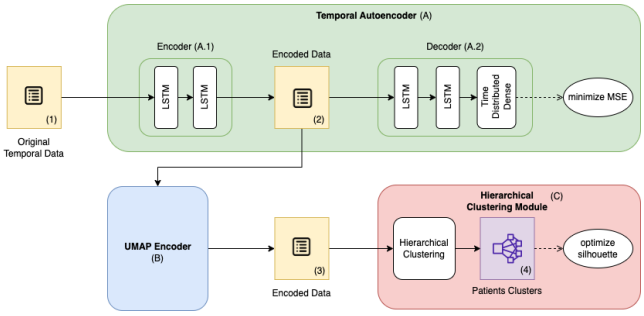


Figure 1: The proposed DTEC Architecture

The DTEC architecture is composed of three main modules, as shown on Figure 1. The Temporal Autoencoder - TAE (Figure 1-A), the UMAP Encoder (Figure 1-B) and the Hierarchical Clustering Module - HCLUST (Figure 1-C).

The TAE receives the original temporal data and returns a latent space corresponding to the encoding of the temporal information of each patient. The UMAP Encoder receives the encoded data and, using graph representations, obtains a new optimized encoding. Lastly, the HCLUST module is responsible for grouping the patients into subgroups based on previously obtained representations.

This architecture was inspired by promising literature architectures [16, 21] but adapted and optimized to the specific context of ALS and the Portuguese ALS Dataset. One key difference is the use of UMAP to optimize the encoding space used for the clustering without reducing the dimensionality of the representations. More regarding this module will be presented in its specific subsection further in this section (see 4.3).

The use of non-deterministic methods, both in the TAE and the UMAP Encoder, results in variations in the representations obtained. This leads to a challenge in terms of the stability of choosing

the best number of clusters. Since the encoding space varies, the ideal number of clusters also changes from run to run. To improve this aspect, a consensus clustering methodology was implemented, inspired by state-of-the-art approaches like the works of Kueffner et al. [15], Liu et al.[20] and Fred et al. [9].

In the following subsections, the three main modules will be discussed in more detail as well as the consensus clustering methodology.

4.1 Additional Data Preprocessing

To be used as input of the TAE, the data needs to be grouped by patient, being shaped as a matrix of dimension:

$$N \times M, \tag{1}$$

where N is the total number of patients, and M is a matrix with dimension $R \times F$ (where R is a fixed number of records per patient, and F is the total number of features - columns excluding the target).

Figure 2 showcases this additional preprocessing performed.

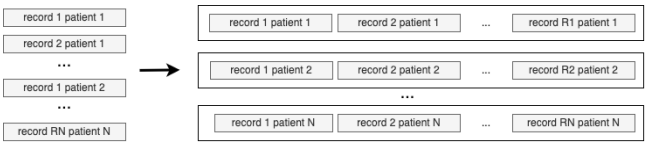


Figure 2: Data Transformation

4.2 Temporal Autoencoder - TAE

The Temporal Autoencoder is responsible for, given the original temporal data (after the preprocessing described in the previous section), learning a lower dimensionality space representation by capturing the most relevant information out of the data (temporal or non-temporal patterns, correlations, etc.).

A more detailed scheme showcasing this module and including the dimensions of each layer and the inputs and outputs obtained can be seen on Figure 3. The optimizer used to train our network is the Adam optimizer [13].

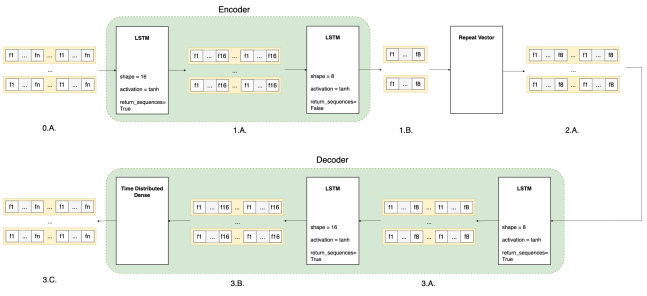


Figure 3: TAE detailed architecture

As an autoencoder, the network can be divided into two primary submodules, the encoder and the decoder, linked by a Repeat Vector layer.

The encoder consists of two LSTMs layers. It receives the original data and, using LSTMs, learns temporal changes across the records. Similarly, the decoder is composed by two LSTM layers but appearing in the reverse order and with the addition of a Time Distributed Dense at the end of the last one.

The learning process of TAE is done through the minimization of the mean-squared error (MSE) that penalizes how much the representation obtained by the decoder differs from the data received by the encoder.

Different configurations for the network were tested, with the combinations of hyperparameters (dimensions, activation functions, etc.) being empirically tuned to maximize the silhouette (which is the metric used to evaluate the quality of clusters, defined in 4.4) obtained by the HCLUST module. Furthermore, some configurations performed worst or increased the computation time without practical advantages on the silhouettes obtained and were therefore excluded.

4.3 UMAP

The UMAP Encoder module uses the UMAP method to encode the data into a new latent space using non-linear and non-stochastic transformations. The objective is to obtain an optimized, more "clusterable" representation of the data.

The use of UMAP to optimize clustering is backed up by current state-of-the-art works such as the works of McConville et al. [23] and Allaoui et al. [1], which state the use of UMAP as a way of improving clustering accuracy. UMAP's preservation of both the local and global structure potentiates the learning of an optimized, more clusterable embedding manifold, which leads to better clusters. Specifically, in the work of Allaoui et al. [1], UMAP proved to improve performance in different clustering algorithms, from density-based approaches, such as HDBSCAN, to hierarchical clustering methods, such as the one used in the DTEC architecture, Agglomerative Clustering.

Although, traditionally, UMAP is used to reduce the dimension of the latent space, we maintain the latent space dimension obtained by the TAE (n_components=8). The rationale behind this choice is the fact that the TAE is already reducing the dimensionality of the data, and further reduction could lead to loss of information (considering all the particularities of the dataset). This was experimentally observed by the lack of improvement in the silhouette score of the clusters obtained with HCLUST when experimenting with lower values.

Similarly to the TAE, the remainings hyperparameters choices were made by considering the improvements on the silhouette score of the clusters later obtained with HCLUST but also considering if the increase of runtime for some configurations was or was not accompanied by significant improvements.

4.4 Hierarchical Clustering - HCLUST

The HCLUST module is the third and last step of the proposed DTEC method. To stratify the patients, represented by the latent space obtained previously, we use a Hierarchical Clustering method, more specifically Agglomerative Clustering using Ward linkage and Euclidean distance as the affinity metric.

Agglomerative Clustering is a hierarchical clustering algorithm that uses a bottom-up approach to group the data. It starts by making each point a cluster and then recursively grouping the two

nearest clusters into one larger cluster. Some of the advantages of using a hierarchical approach are that it also does not require initializing the number of clusters, and instead, it provides a dendrogram to visualize the clusters allowing to cut it at any desired number of clusters. In addition, the dendrogram itself provides a way of finding the optimal number of clusters acting as one of the main advantages of this method when compared with others such as Gaussian Mixtures Models (GMM).

In order to evaluate the optimal number of clusters and its quality, the dendrogram was used allied with the silhouette score.

The silhouette score is a metric of both cohesion and separation, evaluating how much a cluster point is similar to its cluster when compared to its similarity to other clusters. It ranges from -1 to 1, with negative values representing a wrong cluster assignment, values near 0 representing overlapping clusters, and values closer to 1 representing clearly distinguished clusters.

The linkage method chosen was Ward, a method for hierarchical clustering analysis which is considered to perform well on noisy data and is widely used in applications with real-world data.

4.5 Consensus Clustering

As already stated, hierarchical clustering methods provides a dendrogram to visualize the best number of clusters, allowing one to cut the dendrogram at any K number. However, due to the varying representations obtained, the approach still faces issues regarding stability, with the optimal value K changing across runs.

Consensus Clustering is an approach that tackles this challenge by providing a robust methodology to evaluate the stability of the clusters obtained and choose the clustering parameters, such as K. Instead of using only one iteration, Consensus Clustering considers the results of multiple iterations across different latent spaces and with different numbers of clusters.

The methodology used was based on evidence accumulation clustering approaches such as the one proposed by Fred et al. [9]. For each run, clusters are registered in an incremental co-association matrix C, varying the number of clusters from 2 to 4. Then, this matrix is divided by $N \times n_{runs}$ (where N is the number of parameters tested - in this case three, K number of clusters varying from 2 to 4 - and n_{runs} is the total number of runs), to obtain a similarity matrix S. We subtract this similarity matrix S from an all-ones matrix obtaining a distance matrix M. The distance matrix M is later used in Hierarchical Clustering to obtain a consensus dendrogram.

The best number of clusters is chosen by considering the silhouette score, dendrogram, and, additionally, a heatmap representation based on the distance matrix.

Consensus clustering was applied to different configurations in this work, namely, considering the complete feature set, considering only temporal features, and considering only the first record of each patient.

5 STRATIFICATION RESULTS

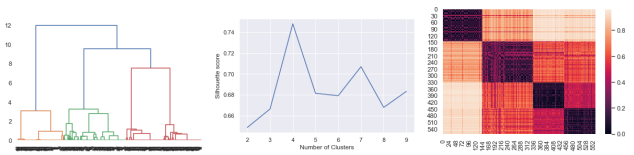
In order to find the optimal number of clusters for each run, the corresponding dendrogram, obtained using Ward linkage, is used together with the silhouette score. However, after performing several runs, its observable that the optimal number varies between several values. The lack of stability in the best K number of clusters at each run of DTEC, causing difficulties in choosing a fixed K optimal number of clusters, motivated the pursuit of a Consensus Clustering approach. This approach is not dependent on a single run, increasing the robustness of the solution and providing a more straightforward way of choosing the best number of clusters.

Three consensus clustering configurations were applied, namely, considering the complete feature set, considering only temporal features, and considering only the first record of each patient. The results of each approach are summarized next.

5.1 Complete Feature Set

In this first consensus clustering configuration, the complete feature set obtained after the preprocessing (described on Section 3) were used, both static and temporal.

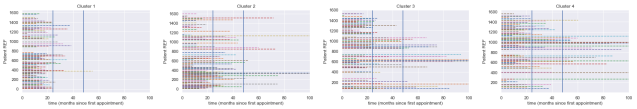
To access the optimal number of clusters for the Consensus Clustering the experiments of DTEC were conducted considering a range of values for optimal K number of clusters (i.e. [2,3,4]). After obtaining the computed distance matrix, three visualizations were used to decide the K optimal number of clusters, namely: the dendrogram, the silhouette scores for different K values, and the corresponding distance matrix heatmap visualization(some of these visualizations are present on Figure 4). The K chosen was 4, presenting a silhouette of 0.75.



(a) Dendrogram (b) Silhouette evolution (c) Distance Matrix Heatmap for K=4

Figure 4: Visualizations for choosing the best number of clusters - Complete Feature Set configuration

The clusters obtained contained differences regarding the duration distribution of its patients (Figure 5), the temporal features evolution (Figure 6) and the static features distributions (Figure 7).

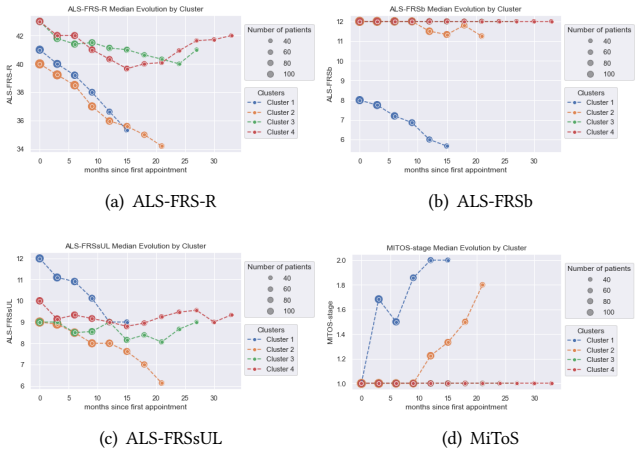


(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4

Figure 5: Duration of each Cluster Record - Complete Feature Set configuration

This differences can be summarized in the following characterization:

- Cluster 1 seems to correspond to patients with, mostly, bulbar onset but also includes patients with an Axial/Respiratory onset. The bulbar onset is medically associated with a slightly faster



(a) ALS-FRS-R (b) ALS-FRSb

(c) ALS-FRSsUL (d) MiToS

Figure 6: ALS-FRSb, ALS-FRSsUL and MiTos Median Evolution across Clusters - Complete Feature Set configuration

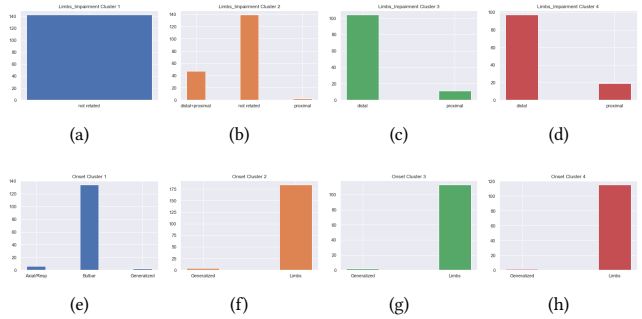
progression, and a shorter duration [10], compared with limbs onset. This insight is also observed in the results found. Furthermore, the patients of this Cluster present an older age at onset and, in contrast with the other clusters, have a slight female predominance.

- Cluster 2 appears to correspond to patients with an average/fast progression and duration associated with limbs onset. A considerable number of patients shows both distal and proximal limbs impairment. There are also many "not related" values at limbs impairment that are most certainly a result of the MVI, holding no clinical meaning due to the fact that the onset is limbs-related. In fact, although not possible to confirm, it is possible that these patients also correspond to distal+proximal limbs impairment due to the generally faster progression.
- Patients in clusters 3 and 4 present limbs onset and are associated with a slower progression and higher duration. Contrasting with cluster 2, patients in these clusters only present one type of impairment (only distal or proximal, but predominantly distal). They also show only one type of limbs affected (either upper or lower).
- The main difference between clusters 3 and 4 is the side of limbs affected. While cluster 3 contains only patients with right limbs affected, Cluster 4 includes patients with either left or, some times, both limbs affected.

5.2 Using Only Temporal Features

For this configuration, only features with temporal evolution were used. Using the same methods described in the "Complete Feature Set" configuration, the K optimal number of clusters chosen was 3 with a silhouette of 0.67. Important to state that values such as 2 are too small to be medically relevant, as well as values too high (such as 5, 6, etc.) considering the relatively small number of patients.

By performing the same analysis regarding duration, temporal features evolution and static features distribution (although those last features are not used for the stratification), some key differences are once again observed.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7: Limbs Impairment (a-d) and Onset (e-h) Distribution by Cluster, some of the most relevant static features in terms of observed differences for the Complete Feature Set configuration

(a) Dendrogram (b) Silhouette evolution (c) Distance Matrix Heatmap for K=3

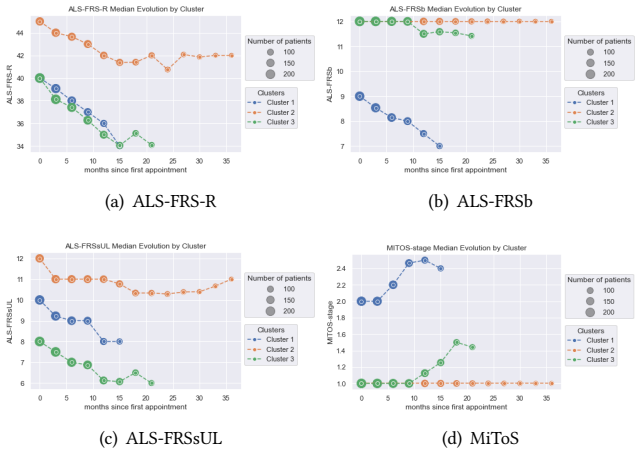
Figure 8: Visualizations for choosing the best number of clusters - Temporal Only configuration

Resuming the results obtained:

- Patients on cluster 1 appear to be relatively older at onset, presenting a slight majority of female patients (in contrast with the other clusters). Patients appear to be associated with a faster progression, bulbar onset, and primarily unrelated to limbs impairment.
- Cluster 2 patients can be associated with a slower progression and limbs onset, especially regarding lower limbs and a distal impairment.
- Cluster 3 seems to contain mostly younger patients with limbs onset, especially associated with upper limbs. The limbs impairment distribution is similar to the one of cluster 2. The progression in this last cluster also appears to be fast, but not as fast as cluster 1.

5.3 Using Only the First Record

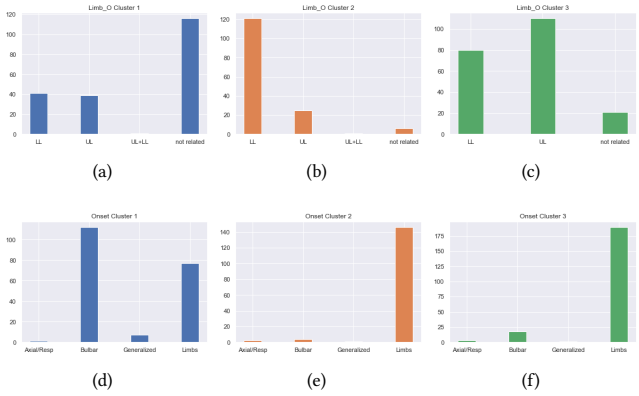
An adjustment is made in the DTEC method to use only the first record of each patient. Specifically, the TAE module is replaced by a Dense Autoencoder (DAE). The key differences in this autoencoder are the use of a Flatten layer at the beginning and the replacement of LSTM layers by Dense layers. All encoding dimensions and activation functions are the same, as well as the optimizer and loss functions used. All the features are used for this configuration, both static and temporal(note, however, that only the first record of temporal features is used as well). Once again, using the same methodology, an analysis is performed considering the dendrogram,



(a) ALS-FRS-R (b) ALS-FRSb

(c) ALS-FRSsUL (d) MiToS

Figure 9: ALS-FRS-R, ALS-FRSb, ALS-FRSsUL and MiTos Median Evolution across Clusters - Temporal Only configuration

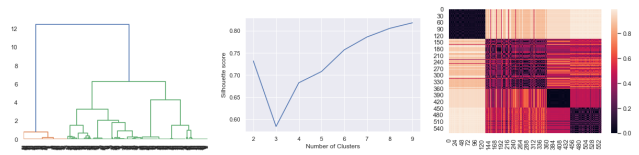


(a) (b) (c)

(d) (e) (f)

Figure 10: Limbs Onset (a-d) and Onset (e-h) Distribution by Cluster, some of the most relevant static features in terms of observed differences for the Temporal Only configuration

silhouette evolution, and distance matrices heatmaps for this configuration. The best medically relevant number of clusters achieved was 4 with a silhouette score of 0.68.

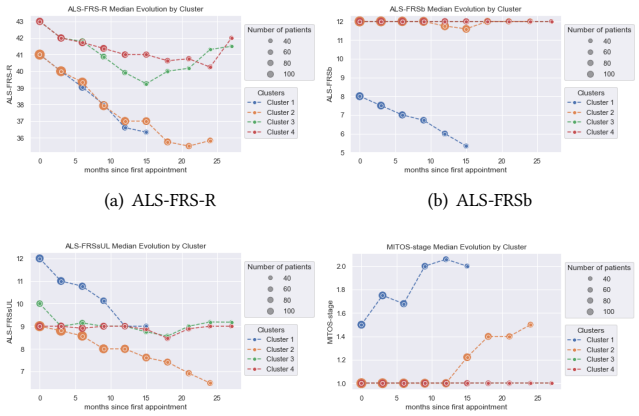


(a) Dendrogram (b) Silhouette evolution (c) Distance matrix Heatmap for K=4

Figure 11: Visualizations for choosing the best number of clusters - First Record Only configuration

Performing the analysis of duration, temporal features evolution and static features distribution of the clusters the main takeaways are:

- Cluster 1 corresponds to bulbar onset patients, generally older, with a slight female predominance, and presenting a faster progression which leads to a shorter duration.
- Cluster 3 and 4 present a similar slower progression and longer duration, diverging in the side of the limbs affected. Cluster 3 contains patients with left limbs onset, while cluster 4 presents patients with right or both limbs affected.
- Cluster 2 lies between the other clusters, presenting a relatively fast progression but with a longer duration when compared with cluster 1. It contains limbs onset patients with distinct sides affected. Regarding limbs, impairment, it contains all the distal+proximal patients but also many "not related" patients. As already explained in other configurations, the "not related" impairment on limbs onset patients is a result of MVI and holds no meaning. Once again, since the evolution analysis showed a faster progression and the cluster also contains distal+proximal patients, it is more likely that the patients marked as "not related" are also distal+proximal patients.



(a) ALS-FRS-R (b) ALS-FRSb

(c) ALS-FRSsUL (d) MiToS

Figure 12: ALS-FRS-R, ALS-FRSb, ALS-FRSsUL and MiTos Median Evolution across Clusters - First Record Only configuration

5.4 Discussion and Conclusions

Initially, the DTEC method raised a few concerns regarding the difficulty of choosing the optimal number of clusters due to the variations in the data representations obtained. These variations were likely a result of the use of non-stochastic steps at DTEC combined with a high heterogeneity and small size of the dataset.

This issue was overcome with the Consensus Clustering approach, which increased the robustness of our solution by providing a clear way of choosing the best number of clusters. This method is not dependent on a single run, using the results obtained across several runs and with different numbers of clusters to calculate the

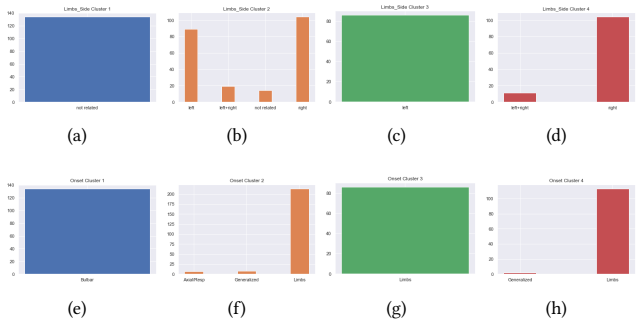


Figure 13: Limbs Side (a-d) and Onset (e-h) Distribution by Cluster, some of the most relevant static features in terms of observed differences for the First Record Only configuration

distance between points and, with that, choose the optimal number of clusters that maximizes the silhouette.

Between the three Consensus Clustering configurations experimented, the clusters that obtained a better silhouette score were the ones of the "Complete Feature Set" configuration, reaching a silhouette of 0.75. This configuration was also the one where it was more evident which was the best number of clusters (4 in this case).

Regarding cluster characterization, a few similarities across configurations can be found. Generally, bulbar onset patients are grouped in one cluster, always associated with an apparent faster progression/shorter duration, which is also supported by clinical studies [10]. One or two clusters appear to group patients with a slower progression in all configurations, associated with limbs onset, sometimes divided by the side of the limbs affected. Finally, the remaining patients appear to be grouped in a cluster with a more intermediate progression.

Furthermore, the clusters obtained using only the first record appears to present similarities with the ones of the first configuration ("Complete Feature Set"). This similarity could point to the fact that more than three records might be required to find more temporal patterns on the data. However, as seen previously, the size of the data when considering patients with more than 3 records is very small (only around 400 patients, a drop of around 25-30% of the current 561 patients, which is already a small number). Furthermore, the clusters obtained using only three records do present differences in the evolution of the disease as observed on the results obtained. An alternative hypothesis is simply that the onset observations and the disease evolution have a strong correlation. Important to note that although the clusters obtained presented similarities, the "Complete Feature Set" configuration obtained a silhouette score 5% higher than the "First Record Only" configuration.

Interestingly, only three clusters were defined when removing the static features ("Temporal Only" configuration). This fact might reflect the importance of static features in separating some of the clusters in the "Complete Feature Set" configuration. In fact, for this configuration, the most significant difference in clusters 3 and 4 is indeed a static feature: the side of the limbs affected.

A medical study by Zhang et al. [32] suggests that the side of the limbs affected appears to be related to the area of the brain involved and to the severity of the disease. This medical study [32] also notes that patients with right limb onset have an older onset age and lower disability severity (measured by ALS-FRS-R) compared with left limb onset patients. In our data, these differences are not so easily noted. The age gap between cluster 3 (right-side patients) and 4 (primarily left-side patients) of the "Complete Feature Set" configuration is relatively trim (60 vs. 59), and the ALS-FRS-R evolution is similar. Nonetheless, the correlation between the limbs side and the disease presentation/brain regions affected [32] justifies the clinical relevance of the clusters obtained.

With this in mind, the configuration considered to obtain the most relevant results was the "Complete Feature Set" configuration. The clusters obtained with this configuration will be used for testing a personalized prognostic prediction approach (see Section 6), which can also be viewed as a validation mechanism.

6 PROGNOSTIC PREDICTION

Inspired by the promising results obtained from previous stratification works when applying prognosis prediction to each separate subpopulation (such as the one of Pires et al. [24]), a classification pipeline was developed. Using the medically relevant clusters obtained by the DTEC, the classification pipeline is applied to each to deliver a more personalized and targeted prediction of the patients' need of NIV.

The general methodology pipeline of the prediction task is present on Figure 14.

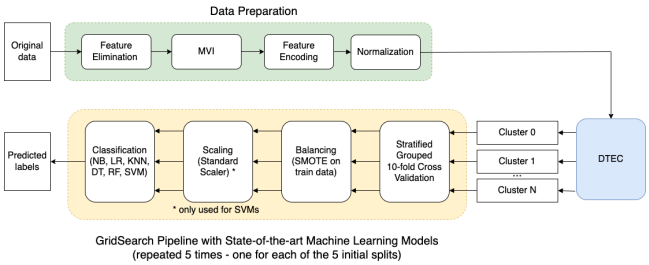


Figure 14: Classification task workflow

After the initial data preparation (described on Section 3), the data is grouped in clusters by the DTEC method (described on Section 4). The clusters obtained by the best configuration of DTEC (Section 5) are fed separately to the Prediction Pipeline. For each cluster of patients C, the data fed to the Classification Pipeline corresponds to the records of all patients in cluster C after the initial Data Preparation is applied.

The Classification Pipeline uses a Gridsearch and a cross-validation methodology to train state-of-the-art classifiers, obtaining a predicted label for each record. It includes a Balancing step (and, in the case of SVMs, Scaling also) to preprocess the data and improve the results. Several classifiers were used, namely, Naive Bayes (NB), Logistic Regression (LR), K-nearest neighbors (KNN), Decision Trees (DT), Random Forests (RF), and Support Vector Machines (SVM).

In order to decide the best time-window to use from the three available (90, 180 and 365), an evaluation regarding target variable distribution on the whole dataset, as well as on each separate cluster, was performed. The 365 days time-window was chosen, Although

still containing some clusters with a high imbalance of the target variable, for the 90 and 180 time-windows this imbalance was even worse, with some clusters containing only 1-2% records of the minority class.

The evaluation and training strategy of the classifiers is based on a 5×10 cross-validation methodology. The evaluation metric used for the grid search optimization is the area under the ROC curve (AUC). We use Specificity (Spe.), Sensitivity (Sen.), and AUC to evaluate the classifiers. It is important to note that all the records are considered independent for evaluation purposes.

6.1 Results

A summary of the best results obtained for each cluster and for the no stratification configuration is present on Table 1.

Table 1: Best Prediction results using Patient Stratification (DTEC) by Cluster and using No Stratification (365 days time-window). For each cluster it is indicated the number of patients and records (and the percentage of the total data they represent in terms of patients and records, respectively).

	Specificity	Sensitivity	AUC	Classifier
Cluster 1	70.0 +- 2.49	69.23 +- 0.0	69.62 +- 1.24	LR
Cluster 2	66.78 +- 0.38	62.28 +- 1.98	64.53 +- 0.8	LR
Cluster 3	73.33 +- 3.95	71.43 +- 0.0	72.38 +- 1.98	LR
Cluster 4	81.43 +- 1.6	85.71 +- 0.0	83.57 +- 0.8	SVM
No Stratification	73.25 +- 0.22	63.41 +- 0.0	68.33 +- 0.11	LR

As observed on Table 1, the classification task obtained better AUCs for 3 out of 4 clusters compared with the best value without stratification (68.33% with Logistic Regression). Particularly, we can see that:

- Cluster 1 obtained a slightly better AUC of 69.62% with Logistic Regression. The Specificity obtained was slightly worse, while the Sensitivity was better.
- Cluster 2 had slightly worse results than the no stratification approach (AUC = 64.53% with Logistic Regression).
- Cluster 3 achieved a considerable improvement in the Sensitivity score, which also reflects on the increase of the AUC score to 72.38% with Logistic Regression.
- Cluster 4 was the one with more significant improvements, with the best algorithm being SVMs. It achieved an improvement of 8.18% on Specificity, 22.3% on Sensitivity, and 15.24% on AUC when compared with the no-stratification results (AUC = 83.57%).

7 CONCLUSIONS

This work proposed a stratification method using deep learning techniques to unveil the temporal disease patterns of ALS and produce clinically relevant clusters of patients. Using a processed subset of the Portuguese ALS dataset, this work introduced one of the first specifically tailored solutions for ALS temporal patient stratification, the Deep Temporal Encoded Clustering (DTEC) method.

To increase the robustness of this solution, a Consensus Clustering approach was implemented, using the accumulated results obtained by running DTEC several times and saving the outcomes with different numbers of clusters.

Using only the first three records of each patient, the method was able to obtain four clinically relevant sub-populations within ALS patients, distinguished by their different evolution and clinical presentations of the disease and achieving a silhouette score of 0.75. A detailed characterization of these subgroups was presented and compared with the subpopulations obtained with different configurations of this same approach.

Some key insights obtained include that bulbar onset patients showed to be related to faster progression. In contrast, limbs onset patients varied their progression from slow to fast, with the main onset differences between sub-populations being the limbs side and type of impairment. Notably, patients that showed distal limb impairment are associated with a slower progression.

The main focus of this work is on the temporal patient stratification methodology in order to obtain medically relevant subgroups; nonetheless, a prognosis prediction pipeline was also implemented. The classification pipeline was mainly developed as a validation mechanism to test the possible improvements regarding the prediction of the need of NIV when considering each cluster separately. These improvements were evident, particularly in some of the clusters found, reaching AUC values up to 83.57%, contrasting to the 68.33% AUC without stratification. Although limited by some factors, which resulted in the improvements not being so evident in some of the clusters, the classification pipeline contributed to showcasing the potential of personalized prognostic prediction.

7.1 Future Work

Further exploration of prognostic prediction methods using the obtained clusters might produce even more significant improvements. For example, an approach that considers each patient instead of each record would be highly relevant. This can be done, for example, by using the DTEC's TAE + UMAP encoding or by using a new encoding module specific to the classification task. It is also pertinent to explore other classification mechanisms besides the ones already used, for example, Deep Learning classification methods.

Another relevant follow-up task would be to integrate the optimization process of DTEC with the prediction pipeline, producing an end-to-end mechanism. This way, DTEC would be optimized by considering the improvements of the classification metrics (Specificity, Sensitivity, and AUC) and not only clustering metrics such as the silhouette.

Alternative solutions to improve the stability of DTEC results, besides the implemented Consensus Clustering mechanism, can also be studied. An additional proposed future work is to explore data augmentation techniques to increase the quantity of training data available since the small size of the dataset is one of the factors considered to impact the results obtained negatively.

Finally, it would be interesting to see how this method performs in other contexts, using a different cohort of ALS patients or even adapting it to another disease.

REFERENCES

[1] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In *Lecture Notes in Computer Science.*

Springer International Publishing, 317–325. https://doi.org/10.1007/978-3-030-51935-3_34

[2] ALS Association [n.d.]. ALS Association "Understanding ALS: What is ALS?" page. https://www.als.org/understanding-als/what-is-als. Last accessed 17 Dec 2021.

[3] ARCAS [n.d.]. ARCAS Homepage. https://arcas.ai/. Last accessed 29 Nov 2021.

[4] Turgay Ayer, Oguzhan Alagoz, Jagpreet Chhatwal, Jude W Shavlik, Charles E Kahn, Jr, and Elizabeth S Burnside. 2010. Breast cancer risk estimation with artificial neural networks revisited. *Cancer* 116, 14 (April 2010), 3310–3321. https://doi.org/10.1002/cncr.25081

[5] James D Berry, Albert A Taylor, Danielle Beaulieu, Lisa Meng, Amy Bian, Jinsy Andrews, Mike Keymer, David L Ennist, and Bernard Ravina. 2018. Improved stratification of ALS clinical trials using predicted survival. *Ann. Clin. Transl. Neurol.* 5, 4 (April 2018), 474–485. https://doi.org/10.1002/acn3.550

[6] Carolyn A Brown, Cathy Lally, Varant Kupelian, and W Dana Flanders. 2021. Estimated prevalence and incidence of amyotrophic lateral sclerosis and SOD1 and C9orf72 genetic variants. *Neuroepidemiology* 55, 5 (July 2021), 342–353. https://doi.org/10.1159/000516752

[7] A Carreiro. 2016. *An integrative mining approach for prognostic prediction in neurodegenerative diseases.* Ph.D. Dissertation. Instituto Superior Técnico, Lisbon,Portugal.

[8] André V Carreiro, Pedro M T Amaral, Susana Pinto, Pedro Tomás, Mamede de Carvalho, and Sara C Madeira. 2015. Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis. *J. Biomed. Inform.* 58 (Dec. 2015), 133–144. https://doi.org/10.1016/j.jbi.2015.09.021

[9] Ana Fred and Anil K. Jain. 2002. Evidence Accumulation Clustering Based on the K-Means Algorithm. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 442–451. https://doi.org/10.1007/3-540-70659-3_46

[10] Jordan R. Green, Yana Yunusova, Mili S. Kuruvilla, Jun Wang, Gary L. Pattee, Lori Synhorst, Lorne Zinman, and James D. Berry. 2013. Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 14, 7-8 (July 2013), 494–500. https://doi.org/10.3109/21678421.2013.817585

[11] Peng Huang, Cheng T Lin, Yuliang Li, Martin C Tammemagi, Malcolm V Brock, Sukhinder Atkar-Khattra, Yanxun Xu, Ping Hu, John R Mayo, Heidi Schmidt, Michel Gingras, Sergio Pasian, Lori Stewart, Scott Tsai, Jean M Seely, Daria Manos, Paul Burrowes, Rick Bhatia, Ming-Sound Tsao, and Stephen Lam. 2019. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *Lancet Digit Health* 1, 7 (Nov. 2019), e353–e362. https://doi.org/10.1016/s2589-7500(19)30159-1

[12] Sahand Khakabimamaghani and Martin Ester. 2016. Bayesian biclustering for patient stratification. *Pac. Symp. Biocomput.* 21 (2016), 345–356. https://doi.org/10.1142/9789814749411_0032

[13] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. (2014). https://doi.org/10.48550/ARXIV.1412.6980

[14] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13 (2015), 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[15] Robert Kueffner, Neta Zach, Maya Bronfeld, Raquel Norel, Nazem Atassi, Venkat Balagurusamy, Barbara Di Camillo, Adriano Chio, Merit Cudkowicz, Donna Dillenberger, Javier Garcia-Garcia, Orla Hardiman, Bruce Hoff, Joshua Knight, Melanie L Leitner, Guang Li, Lara Mangravite, Thea Norman, Liuxia Wang, ALS Stratification Consortium, Jinfeng Xiao, Wen-Chieh Fang, Jian Peng, Chen Yang, Huan-Jui Chang, and Gustavo Stolovitzky. 2019. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Sci. Rep.* 9, 1 (Jan. 2019), 690. https://doi.org/10.1038/s41598-018-36873-4

[16] Isotta Landi, Benjamin S. Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T. Dudley, Cesare Furlanello, and Riccardo Miotto. 2020. Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digital Medicine* 3, 1 (July 2020). https://doi.org/10.1038/s41746-020-0301-z

[17] Changhee Lee and Mihaela van der Schaar. 2020. Temporal Phenotyping using Deep Predictive Clustering of Disease Progression. (2020). https://doi.org/10.48550/ARXIV.2006.08600

[18] Xiangtao Li and Ka-Chun Wong. 2019. Evolutionary multiobjective clustering and its applications to patient stratification. *IEEE Trans. Cybern.* 49, 5 (May 2019), 1680–1693. https://doi.org/10.1109/tcyb.2018.2817480

[19] Enju Lin, Jennifer L. Hefner, Xianlong Zeng, Soheil Moosavinasab, Thomas Huber, Jennifer Klima, Chang Liu, and Simon M. Lin. 2019. A deep learning model for pediatric patient risk stratification. *The American journal of managed care* 25 10 (2019), e310–e315.

[20] H Liu, R Zhao, H Fang, F Cheng, Y Fu, and Y Y Liu. 2017. Entropy-based consensus clustering for patient stratification. *Bioinformatics* 33, 17 (2017), 2691–2698. https://doi.org/10.1093/bioinformatics/btx167

[21] N S Madiraju. 2018. *Deep temporal clustering: Fully unsupervised learning of time-domain features.* Ph.D. Dissertation. Arizona State University.

[22] Andreia S Martins, Marta Gromicho, Susana Pinto, Mamede de Carvalho, and Sara C Madeira. 2021. Learning prognostic models using DiseaseProgression patterns: Predicting the need forNon-invasive Ventilation in amyotrophic LateralSclerosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* PP (May 2021), 1–1. https://doi.org/10.1109/tcbb.2021.3078362

[23] Ryan McConville, Raul Santos-Rodriguez, Robert J Piechocki, and Ian Craddock. 2020. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding. (2020). arXiv:1908.05968 [cs.LG]

[24] Sofia Pires, Marta Gromicho, Susana Pinto, Mamede de Carvalho, and Sara C Madeira. 2020. Patient stratification using clinical and patient profiles: Targeting personalized prognostic prediction in ALS. In *Bioinformatics and Biomedical Engineering*. Springer International Publishing, Cham, 529–541. https://doi.org/10.1007/978-3-030-45385-5_47

[25] Divya Ramamoorthy, Kristen Severson, Soumya Ghosh, Karen Sachs, Jonathan D Glass, Christina N Fournier, James Berry, Kenney Ng, Ernest Fraenkel, Answer ALS, and Pooled Resource Open-Access ALS Clinical Trials Consortium. 2021. Identifying patterns of ALS progression from sparse longitudinal data. (May 2021).

[26] Divya Ramamoorthy, Kristen Severson, Soumya Ghosh, Karen Sachs, Jonathan D Glass, Christina N Fournier, James Berry, Kenney Ng, Ernest Fraenkel, Answer ALS, and Pooled Resource Open-Access ALS Clinical Trials Consortium. 2021. Identifying patterns of ALS progression from sparse longitudinal data. (May 2021). https://doi.org/10.1101/2021.05.13.21254848

[27] Cátia M Salgado and Susana M Vieira. 2020. Machine learning for patient stratification and classification part 2: Unsupervised learning with clustering. In *Leveraging Data Science for Global Health*. Springer International Publishing, Cham, 151–168. https://doi.org/10.1101/2021.05.13.21254848

[28] Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B Ellis, Erwin P Bottinger, and John V Guttag. 2015. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *J. Biomed. Inform.* 53 (Feb. 2015), 220–228. https://doi.org/10.1016/j.jbi.2014.11.005

[29] Irene Tramacere, Eleonora Dalla Bella, Adriano Chiò, Gabriele Mora, Graziella Filippini, Giuseppe Lauria, and EPOS Trial Study Group. 2015. The MITOS system predicts long-term survival in amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* 86, 11 (Nov. 2015), 1180–1185. https://doi.org/10.1136/jnnp-2014-310176

[30] Hannelore K van der Burgh, Ruben Schmidt, Henk-Jan Westeneng, Marcel A de Reus, Leonard H van den Berg, and Martijn P van den Heuvel. 2017. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *NeuroImage Clin.* 13 (2017), 361–369. https://doi.org/10.1016/j.nicl.2016.10.008

[31] Bart Vrijsen, Dries Testelmans, Catharina Belge, Wim Robberecht, Philip Van Damme, and Bertien Buyse. 2013. Non-invasive ventilation in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Frontotemporal Degener.* 14, 2 (March 2013), 85–95. https://doi.org/10.3109/21678421.2012.745568

[32] Qiuli Zhang, Cuiping Mao, Jiaoting Jin, Chen Niu, Lijun Bai, Jingxia Dang, and Ming Zhang. 2014. Side of Limb-Onset Predicts Laterality of Gray Matter Loss in Amyotrophic Lateral Sclerosis. *BioMed Research International* 2014 (2014), 1–11. https://doi.org/10.1155/2014/473250

[33] Andrew Zhen, Minjeong Kim, and Guorong Wu. 2021. Disentangling The Spatio-Temporal Heterogeneity of Alzheimer's Disease Using A Deep Predictive Stratification Network. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 46–49. https://doi.org/10.1109/ISBI48211.2021.9433903