Identifying Extreme Profiles in Amyotrophic Lateral Sclerosis Patients at Diagnosis through Archetypal Analysis

I. Trescato[1], E. Tavazzi[1], M. Vettoretti[1], R. Vasta[2], A. Chiò[2], B. Di Camillo[1,3*]
[1] *Department of Information Engineering, University of Padova, Padova, Italy.*
[2] *Neuroscience Department "Rita Levi Montalcini", University of Torino, Torino, Italy.*
[3] *Department of Comparative Biomedicine and Food Science, University of Padova, Padova, Italy.*
[*] *Corresponding author: barbara.dicamillo@unipd.it.*

***Abstract*—The clinical heterogeneity that characterizes Amyotrophic Lateral Sclerosis (ALS) makes its diagnosis, prognosis, and care difficult. In this context, characterizing patients based on their clinical features or progression patterns is crucial, allowing a deeper understanding of the disease and the planning of more effective treatments. In this work, we employ Archetypal Analysis for studying a real-world ALS population based on their characteristics at diagnosis. First, we derive a set of extreme clinical types (archetypes) whose combination describes the study population, and analyze their differences in terms of clinical characteristics. Then, we cluster patients according to their similarity to the archetypes and we investigate how the so-obtained groups differ in terms of time to life-support interventions and survival.**

***Keywords*—Amyotrophic Lateral Sclerosis, Archetypal Analysis, patient stratification, unsupervised clustering.**

I. Introduction

Amyotrophic lateral sclerosis (ALS) is a progressive and degenerative disease that affects the nerve cells that control voluntary muscle movement. ALS progression impairs motor neurons in the brain and spinal cord, wasting the muscles and leading to the inability to control movements; sometimes it is also accompanied by cognitive and behavioural symptoms. ALS aetiology is still unknown, with some genetic and environmental factors possibly triggering the disease onset. The mean life expectancy is 3-5 years from onset, with death usually occurring from respiratory failure. Despite a relative uniformity during the late stages of the disease, the symptoms at the onset and the timing of the clinical manifestations are highly variable among the patients [1].

Related to this heterogeneity, one of the major needs is the identification of groups of patients with similar characteristics (*i.e.*, patient stratification), to be able to effectively predict the course of the disease in terms of speed of worsening, symptom occurrence, and the need for life-supporting interventions. For a deep understanding of this rare and incurable disease, it is also essential to investigate which are the main markers that determine its different manifestations. Clinically, patients can be stratified into phenotypes based on their characteristics or by observing their clinical progression. Alternatively, it is possible to automatically stratify patients using a data-driven approach: recently, clustering algorithms such as k-means, dimensionality reduction methods such as Uniform Manifold Approximation and Projection (UMAP), and network-based approaches were employed for this purpose [2], [3].

Archetypal Analysis (AA) is an alternative unsupervised computational approach that allows to discern a specific number of extreme and not necessarily observed points called archetypes (i.e. ideal, prototype patients), such that each archetype is constrained to be a mixture of points in the dataset and such that each point can be well represented as convex mixtures of the archetypes [4]. AA can be used on multivariate datasets as an exploratory tool since, analyzing the characteristics of each archetype, allows for highlighting the differences among groups of patients.

In this work, we aim at stratifying ALS patients by employing AA and considering the variables collected at the diagnosis of a real-world ALS cohort. First, we derive the archetypes and analyze their differences in terms of clinical features. Then, we associate each patient with their closest archetype. Finally, we investigate how the identified patients' groups differ in terms of clinical outcomes, considering: the need for non-invasive ventilation (NIV), percutaneous endoscopic gastrostomy (PEG), tracheostomy, and death.

II. Materials and methods

A. Dataset

The dataset used in this work was extracted from the Piemonte and Valle d'Aosta ALS register (PARALS) [5]: selecting the patients with the first visit from January 1st, 2007 to December 31st, 2015, we identified 924 ALS subjects. For each patient, we collected features to characterize their condition at the time of diagnosis, including:

- demographics and lifestyle: sex, marital status, educational level, smoke habits;
- dates of: birth, onset, diagnosis, PEG, NIV, tracheostomy, death;
- ALS-related variables: onset site, phenotype, mutation in ALS-linked genes (C9orf72, SOD1, TARDBP, FUS, OPT, TUB);

- comorbidities: psoriasis, epilepsy, stroke, rheumatic diseases, poliomyelitis, obstructive sleep apnea syndrome, monoclonal gammopathy of undetermined significance;
- other health-related variables: frontotemporal dementia (FTD), forced vital capacity (FVC), first/second tumour, thyroid impairments, psychiatric diseases, hypertension, diabetes, chronic obstructive pulmonary disease (COPD);
- a panel of 39 different blood exams, covering general health analysis, heart problems, and organ function.

B. Data preprocessing

A preprocessing phase was necessary to solve some typical issues that may arise with real-world data. First, we aggregated the variables presenting low occurrences and belonging to the same category, creating the binary variables: "genetics", which indicates if a mutation occurs in at least one tested gene; "comorbidities", indicating that at least one of the concurrent conditions listed above occurs; "tumour", which indicates that the patient had at least one cancer. Other more frequent and already binary conditions, such as hypertension or diabetes, were not aggregated. Variables assuming multiple possible values were either converted into binary features, if their occurrence had low frequency in the data (such as psychiatric diseases, that were transformed into new features indicating the presence/absence of at least one psychiatric disease), or coded as dummy features (such as FTD and smoke habits).

Then, we derived some variables coding the time passed from the ALS onset to the other clinical events for which a date was available, obtaining the following new features: age at onset, diagnostic delay, and time to PEG/NIV/tracheostomy/death. Next, we filtered out the blood test variables presenting more than 30% of missing values. In total, 16 of the 39 available tests were removed, 10 of which had more than 50% missing values. We imputed the remaining missing values in the preprocessed data using the *mice* R package [6] with default parameters. To check the robustness of the imputation process, we compared the distribution of each variable before and after the imputation.

Tables I and II provide an overview of the data after these preprocessing steps. Lastly, we scaled all variables in the range [0,1] to balance the contribution of features to the analysis.

C. Method

The goal of the analysis is the unsupervised stratification of the patients according to their clinical characteristics at diagnosis and the comparison of the obtained groups in terms of clinical outcomes, namely time to NIV, time to PEG, time to tracheostomy, and time to death. The analysis consists of three steps: (1) the identification of a number of archetypes from the data, using all the available variables except those corresponding to the outcomes; (2) the unsupervised classification of the subjects into different clusters according to their similarity to the archetypes; (3) the comparison of the clusters in terms of the four outcomes of interest.

1) *Archetypal Analysis.* Given an $n \times m$ matrix X representing a multivariate dataset, where n is the number of observations, m is the number of variables, and k is the number of archetypes chosen by the user, AA allows determining a $k \times m$ matrix Z of archetypes such that:

a) the data are best approximated by convex combinations of the archetypes Z, *i.e.*, they minimize the residual sum of squares (RSS):

$$RSS = ||X - \alpha Z^T||_2, \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1 \quad (1)$$

b) the archetypes are convex combinations of the data points:

$$Z = X^T \beta, \quad \beta_i \geq 0, \quad \sum_i \beta_i = 1 \quad (2)$$

where α are the coefficients of the archetypes and β are those of the dataset, respectively.

Since the identification of the archetypes is based on an iterative process that, starting from a set of *k* randomly chosen points in the features' space, minimizes the approximation error between the original data points and those reconstructed as a combination of the archetypes, it is recommended to repeat the identification algorithm several times to avoid falling into local minima [4]. Here, we tested a number of archetypes *k* from 2 to 8, repeating the procedure 15 times for each *k*. The optimal number of archetypes was chosen by detecting the elbow in the scree plot reporting on the x-axis the value of *k* and on the y-axis the RSS. To reinforce the choice of the best *k*, we also analyzed the minimum, mean and standard deviation of the RSS over the 15 repetitions for each *k* [7].

The archetypes, by construction, lie on the boundary of the convex hull of the data and can be therefore easily influenced by outliers. Thus, to avoid incorrect or skewed results, we employed the identification method of *robust archetypes*, which reduces the influence of outliers by using M-estimators instead of least squares estimators when performing the optimization procedure [8]. All analyses were performed using the *archetypes* R package [9].

For the sole purpose of visualizing the data and archetypes in two dimensions, principal component analysis (PCA) was used. The two-dimensional representation allows to easily check the position of the archetypes with respect to the data and to verify that they are actually on the convex hull. Radar plots were used to investigate the characteristics of each archetype individually; for a more effective rendering, we computed for each variable its standard deviation (sd) across the *k* archetypes and then only included in the radar plots those with sd>0.13 (sd_{min}=0.00004679, sd_{max}=0.4158768), i.e. those variables that differentiate the most among archetypes.

2) *Subject clustering based on archetypes.* As mentioned, AA allows representing the data as a linear combination of archetypes, each multiplied by a non-negative coefficient α_i. We decided to assign each subject to the archetype with higher α_i, following the rationale of the nearest prototype classifier [7], to inspect how patients cluster based on their extreme behaviours, obtaining *k* distinct clusters.

3) *Comparison of the cluster in terms of clinical outcomes.* Finally, we compared the different clusters using the Wilcoxon test to assess any statistically significant differences in terms of

TABLE I: Preprocessed categorical variables.

Feature	Levels	% of subjects
sex	0: male	52 %
	1: female	48 %
marital status	0: combined	77 %
	1: living alone	23 %
education	0: illiterate	3 %
	1: primary school	35 %
	2: 8th grade	32 %
	3: short diploma	7 %
	4: high school	16 %
	5: graduated	7 %
smoke habits *	0: never	48 %
	1: ex	35 %
	2: current	17 %
genetics	0: no gene mutations	89 %
	1: gene mutation	11 %
onset site	0: bulbar	33 %
	1: not bulbar	67 %
FTD *	0: ftd	18 %
	1: cognitive	20 %
	2: behavioral	16 %
	3: non-executive	3 %
	4: healthy	49 %
tumour	0: never	95 %
	1: at least one	5 %
comorbidities	0: none	88 %
	1: at least one	12 %
thyroid	0: healthy	88 %
	1: impairment	12 %
psychiatric dis.	0: healthy	95 %
	1: impairment	5 %
hypertension	0: healthy	49 %
	1: impairment	51 %
diabetes	0: healthy	89 %
	1: impairment	11 %
COPD	0: healthy	93 %
	1: impairment	7 %

TABLE II: Preprocessed continuous variables.

Feature	25^{th}-50^{th}-75^{th} percentile
age at onset [years]	60 - 67 - 74
diagnostic delay [months]	5 - 9 - 14
time to PEG ◇ [months]	16 - 24 - 33
time to NIV ◇ [months]	14 - 23 - 34
time to tracheo ◇ [months]	21 - 27 - 43
time to death ◇ [months]	19 - 30 - 47
FVC at diagnosis [L]	70 - 87 - 104
white blood cells [10^9/L]	5.2 - 6.2 - 7.5
neutrophils [10^9/L]	2.8 - 3.6 - 4.6
lymphocytes [10^9/L]	1.4 - 1.7 - 2.2
monocytes [10^9/L]	0.4 - 0.5 - 0.6
ESR [1] [mm/h]	4 - 9 - 19
creatinine [mg/dL]	0.6 - 0.7 - 0.9
uric acid [mg/dL]	3.8 - 4.7 - 5.6
albumin [g/dl]	4 - 4.3 - 4.6
glucose [mg/dL]	81 - 88 - 99
triglycerides [mg/dL]	73 - 94 - 128
$cholesterol_{tot}$ [mg/dL]	174 - 199 - 230
$cholesterol_{hdl}$ [mg/dL]	49 - 58 - 70
$cholesterol_{ldl}$ [mg/dL]	96 - 117 - 144
$cholesterol_{ldl/hdl}$	1.5 - 2 - 2.7
$bilirubin_{tot}$ [mg/dL]	0.5 - 0.7 - 0.9
$bilirubin_{dir}$ [mg/dL]	0.2 - 0.2 - 0.3
$bilirubin_{indir}$ [mg/dL]	0.3 - 0.5 - 0.6
alkaline phosphatase [U/L]	56 - 68 - 86
creatine phosphokinase [U/L]	90 - 153 - 251
sodium [mEq/L]	140 - 142 - 143
potassium [mmol/L]	3.9 - 4.2 - 4.4
chlorine [mmol/L]	101 - 103 - 105
TSH [2] [mU/L]	0.9 - 1.6 - 2.5

* categorical variables coded as dummy
◇ outcome variables
[1] erythrocyte sedimentation rate
[2] thyroid stimulating hormone

the time of occurrence of the clinical outcomes, by performing a multiple pairwise comparison between groups.

III. Results

Fig. 1 shows the scree plot for the tested k, while Tab. III shows the minimum, average and standard deviation value of the RSS calculated over the 15 repetitions for each k. These methods identify $k = 7$ as the optimal number of archetypes.

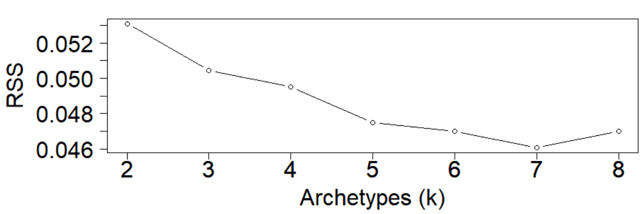

Fig. 1: Scree plot for a number of archetypes k from 2 to 8. The RSS value displayed for each k is the minimum RSS obtained over the 15 repetitions of the algorithm.

TABLE III: Minimum, mean, and sd of the RSS computed over 15 repetitions, for a number of archetypes k from 2 to 8.

Archetypes	k=2	k=3	k=4	k=5	k=6	k=7	k=8
min	0.0531	0.0504	0.0495	0.0475	0.0470	**0.0461**	0.0470
mean	0.0531	0.0511	0.0502	0.0482	0.0479	**0.0474**	0.0476
sd	<0.0001	0.000642	0.00045	0.001264	0.000715	**0.00064**	0.000426

Fig. 2 shows how the archetypes (in red) are positioned amongst the subjects (in green): as expected, they position

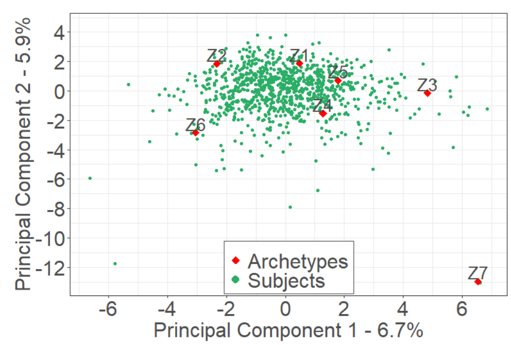

Fig. 2: Two-dimensional representation of the subjects (green dots) and archetypes (red diamonds) using PCA.

on the borders of the dataset, identifying types who are not necessarily observed in the data but who are extreme in their characteristics.

The 7 mined archetypes are reported in Fig. 3 as radar plots. The variables were filtered according to their standard deviation, ultimately retaining 24 features out of a total of 44.

By analyzing the resulting archetypes Z_i, we can describe the extreme behaviors characterizing the data. Z_1 represents a male subject with non-bulbar onset and medial age at onset, as well as Z_2, which is a female subject almost free of other diseases. Z_3 delineates an individual who is living alone, with mainly no FTD issues and high bilirubin values. Z_4 represents a subject with rather advanced age at onset and FTD at diagnosis. Z_5 outlines an ex-smoker subject, with older age at onset and affected by hypertension; this is, in general, also the archetype placing greater emphasis on the presence of other comorbidities, such as diabetes. Z_6 is mainly similar to Z_1, presenting in addition relatively high values of triglycerides and cholesterol. Finally, Z_7 describes a female subject, ex-smoker, with pronounced cognitive FTD impairment and with high ESR and albumin values; this is also the only archetype with a low chlorine value.

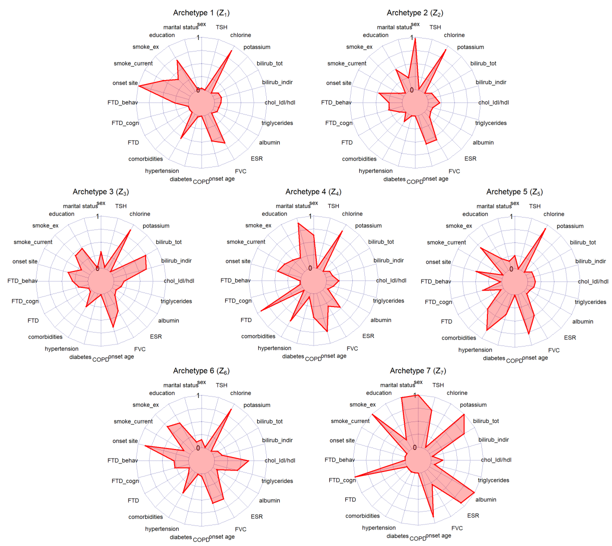

Fig. 3: Radar plots representing the archetypes, including the 24 variables with sd>0.13 only. For each variable, the innermost circle represents a standardised value equal to 0, the outermost circle equal to 1. The levels of the categorical variables correspond to those of Tab. I, then normalized in the range [0,1].

We then assigned each subject to their most representative archetype, getting seven clusters C_i with 60 subjects in C_1, 286 in C_2, 59 in C_3, 79 in C_4, 259 in C_5, 180 in C_6, and only 1 subject in C_7. Based on this, we decided to perform the following analysis only considering clusters C_1-C_6.

Fig. 4 reports the comparison of the time of occurrence of the four considered outcomes in the different clusters. The clusters differ in terms of mean time as well as in the order of occurrence of the outcomes. Clusters C_1 and C_3 have time-to-event values comparable with those of the other clusters, for all outcomes but tracheostomy, where C_3 has the shortest time of occurrence among all clusters. Statistically significant differences are observed in C_2 vs C_4, C_2 vs C_5, and C_4 vs C_6 for PEG, NIV and death (all p-values$<$0.015). This confirms that archetypes-based clustering can be a useful tool to identify groups of patients with different characteristics both in terms of covariates at diagnosis and clinical outcomes.

IV. Conclusion

In this work, we applied AA to a multivariate dataset of ALS patients to study disease heterogeneity and outline extreme behaviors at diagnosis. We identified 7 archetypes that were first compared in terms of clinical characteristics, and then used to define 7 clusters of patients. By analyzing their distributions in terms of time to PEG, NIV, tracheostomy,

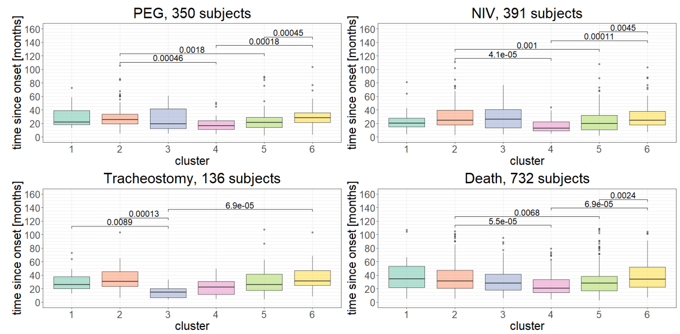

Fig. 4: Box plots comparing the times of occurrence of the outcomes (PEG, NIV, tracheostomy, and death) among the clusters. For each outcome, the number of subjects who experienced the event and statistically different p-values (Wilcoxon test, threshold = 0.01) are reported.

and death, we assessed how stratifying patients based on their similarity to the mined archetypes can be a valid criterion to identify groups characterized by different progression timing and patterns. In future works, we aim to further characterize the identified clusters to explore intra-cluster variability (e.g. by comparing the characteristics of the subjects belonging to the same cluster) and inter-cluster variability (e.g. considering the differences in progression patterns and comparing the archetypes with the clinical phenotypes described in the literature).

Acknowledgement

This research was supported by the University of Padova project C94I19001730001, by the Italian Ministry of Health (Ricerca Finalizzata) grant RF-2016-02362405, and by the Italian Ministry of Education, University and Research (PRIN) grant 2017SNW5MB, and by the BRAINTEASER project (contract n. GA101017598), as a part of the European Union's Horizon 2020 research and innovation programme. ET was founded by the Department of Information Engineering, University of Padova (Research Grant B junior).

References

[1] Chiò, A., Calvo, A., Moglia, C., Mazzini, L. & Others Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study. *Journal Of Neurology, Neurosurgery & Psychiatry*. **82**, 740-746 (2011)

[2] Tang, M., Gao, C., Goutman, S. & Others Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering. *Neuroinformatics*. **17** pp. 407-421 (2019)

[3] Tavazzi, E., Daberdaku, S., Zandonà, A., Vasta, R. & Others Predicting functional impairment trajectories in amyotrophic lateral sclerosis: a probabilistic, multifactorial model of disease progression. *Journal Of Neurology*. **269**, 3858-3878 (2022)

[4] Cutler, A. & Breiman, L. Archetypal analysis. *Technometrics*. **36**, 338-347 (1994)

[5] Chiò, A., Mora, G., Moglia, C., Manera, U., Canosa, A. & Others Secular trends of amyotrophic lateral sclerosis: the Piemonte and Valle d'Aosta register. *JAMA Neurology*. **74**, 1097-1104 (2017)

[6] Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal Of Statistical Software*. **45** pp. 1-67 (2011)

[7] Ragozini, G., Palumbo, F. & D'Esposito, M. Archetypal analysis for data-driven prototype identification. *Statistical Analysis And Data Mining: The ASA Data Science Journal*. **10**, 6-20 (2017)

[8] Eugster, M. & Leisch, F. Weighted and robust archetypal analysis. *Computational Statistics & Data Analysis*. **55**, 1215-1225 (2011)

[9] Eugster, M. & Leisch, F. From spider-man to hero-archetypal analysis in R. (2009)