Machine Learning and Feature Selection Methods to Predict 1-Year Disease Progression for Amyotrophic Lateral Sclerosis

Thibault Anani, Jean-François Pradat-Peyre, Francois Delbot, Pierre-François Pradat

Machine Learning and Feature Selection Methods to Predict 1-Year Disease Progression for Amyotrophic Lateral Sclerosis

Thibault Anani[1*], Jean-François Pradat-Peyre[1,2], François Delbot[1,2], Pierre-François Pradat[3,4]

[1*]Sorbonne Université, CNRS, LIP6, 4 Pl. Jussieu, Paris, 75005, France.
[2]U.F.R. SEGMI, Nanterre Université, 200 av. de la République, Nanterre, 92000, France.
[3]Département de Neurologie, Hôpital de la Pitié-Salpêtrière - APHP, 47-83 Bd de l'Hôpital , Paris, 75013, France.
[4]LIB, Sorbonne Université, 15 Rue de l'École de Médecine, Paris, 75006, France.

*Corresponding author(s). E-mail(s): thibault.anani-agondja@lip6.fr;
Contributing authors: jean-francois.pradat-peyre@lip6.fr; francois.delbot@lip6.fr; pierre-francois.pradat@aphp.fr;

Abstract

Background: This study focuses on Amyotrophic Lateral Sclerosis (ALS), a progressive neurodegenerative disease that affects motor neurons. ALS patients suffer from motor weakness, atrophy, spasticity, and difficulties in speaking, swallowing and breathing. Predicting the survival and progression of ALS is essential for optimising patient care, interventions and informed decision-making.
Methods: Using the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) database and clinical trial data from Exonhit Therapeutics, we applied machine learning techniques to predict disease progression based on ALS Functional Rating Scale (ALSFRS) scores and patient survival over one year of follow-up. Our models were validated through 10-fold cross-validation. Kaplan-Meier estimation was used to cluster patients effectively according to their profiles. To enhance the predictive accuracy of our models, we performed feature selection prior to analysis.
Results: Logistic regression combined with feature selection yielded a Balanced Accuracy score of 76% (68.6% to 79.8% in each fold) on validation data and 76.33% on test data. Integrating model results with a Kaplan-Meier estimates, we

identified four patient clusters, with a C-Index of 0.81 and an overall log-rank test p-value ≤ 0.0001. Our method demonstrated high accuracy in predicting ALSFRS score progression at 3 months (RMSE = 2.88 and adjusted $R^2 = 0.784$ with random forest). This study showcases the potential of machine learning models to provide significant predictive insights in ALS, enhancing the understanding of disease dynamics and supporting patient care.

Keywords: machine learning, feature selection, optimisation, classification, regression

1 Introduction

Amyotrophic lateral sclerosis (ALS), also known as Charcot's disease, is a progressive neurodegenerative disorder for which there is currently no cure. It is a particularly heterogeneous disease, and symptoms may vary from one individual to another, but patients generally share many similar symptoms. ALS affects motor neurons in the brain and spinal cord, leading to progressive degeneration of the body's muscles. As the disease progresses, muscles can become stiff and spastic, leading to muscle cramps and reduced mobility. Muscles begin to atrophy, leading to a loss of muscle mass. Basic functions such as speech, swallowing and breathing may also be affected, leading to difficulties in communicating and eating. In the advanced stages of the disease, progressive paralysis can develop, considerably limiting the patient's independence and quality of life. Median life expectancy at the onset of symptoms varies between 3 and 5 years. Disease progression is highly variable in terms of functional decline, body region affected and survival [1]. Consequently, establishing a reliable prognosis is a major challenge, as it conditions patient management and quality of life, e.g. by planning interventions or choosing appropriate levels of treatment and care. The evolution of a patient's physical functions is calculated using the ALS Functional Rating Scale (ALSFRS), a scale based on a series of 10 questions assessing motor abilities in several different areas: bulb, breathing, trunk, upper limbs and lower limbs [2]. Each domain is scored on a scale from 0 to 4, where 4 represents normal function and 0 a total loss of function. Thus, a score of 40 indicates normal motor functions, while a score of 0 indicates total loss of these abilities. There is also a more accurate revised version of this scale, called ALSFRS-R, which includes additional items to assess respiratory function more comprehensively [3]. Machine learning methods on large datasets, such as the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) database, have been used to exploit correlations in the data to predict disease progression. Two families of learning methods can be distinguished.

Supervised learning methods expose the algorithm to labelled data (i.e. data associated with outcomes). Several types of methods have been used to predict the evolution of a patient's health condition or their chances of survival in the mid term. These include classification methods [4–7], regression models [8, 9], deep learning methods [10, 11] as well as survival models [12, 13]. Unsupervised learning methods, on the other hand, expose the algorithm to unlabelled data (i.e. data with no associated results) and seek to identify hidden structures in this data. This type of learning is particularly

interesting for ALS patients, as it allows us to discover relationships and patterns in medical data that may not be obvious at first glance, particularly in data monitoring the progression of the disease. Clustering [6] and dimension reduction using techniques such as Principal Component Analysis (PCA) [14] or Uniform Manifold Approximation and Projection (UMAP) [15, 16] are able to group patients into clusters with similar profiles. This can help clinicians to better understand individual differences in ALS progression and to develop personalised treatment strategies.

However, because patients are so heterogeneous, prognostic models often lack precision and do not always provide information on the uncertainty and interpretability of results, which are crucial for clinicians. This issue is rarely addressed. The interpretability and complexity of machine learning models are essential, particularly in sensitive areas such as healthcare, for several reasons. Firstly, it enables healthcare professionals to understand and trust the model's predictions. If a clinician cannot explain why a model has predicted a certain prognosis, they may hesitate to base their decisions on this prediction. Interpretability is also important for model verification and validation. If a model is a "black box", it is difficult to check that it is working correctly and that it has not learned misleading correlations from the data. Finally, interpretability is essential for accountability and transparency. In areas where decisions can have far-reaching consequences, it is vital to know how and why these decisions were taken.

In this paper, we conducted a detailed analysis of clinical data from ALS patients. We applied supervised learning methods to create accurate prognostic models, capable of predicting the 1-year evolution of the ALSFRS score and the survival of a patient according to their features. Our methodology primarily involves advanced feature selection techniques that enhance the predictive accuracy and interpretability of the models while minimising their complexity. We have also sought to identify clusters of patients with similar features and survival rates according to the model outputs.

2 Materials and methods

2.1 Description of data and participants

The data used in this study come from two longitudinal datasets. The first dataset is the PRO-ACT database, available online [17] and composed of ALS clinical data [18]. It has been designed to bring together data from different ALS clinical trials to provide a comprehensive overview of their results and to facilitate research into the disease. It is freely accessible to researchers and clinicians, making it a valuable asset. The data in this database is frequently used to evaluate the effectiveness of new treatments for ALS, as well as to better understand the features of the disease and the associated risk factors. It is structured into a number of tables based on patient information such as age, gender, type of ALS, stage of disease and medical test results, as well as information on treatments received by patients, such as drugs, therapies and surgery. Although it is not representative of all patients with the disease, it is the largest benchmark database on ALS, offering a unique opportunity to develop prognostic models. The second dataset, the Exonhit database, is made up of patients from the Exonhit Therapeutics clinical trial [19]. It includes 400 patients who were followed

for 18 months, from October 2002 to August 2004. The aim of this clinical trial was to evaluate the effect of pentoxifylline on ALS patients, with a view to developing a complementary treatment to riluzole if efficacy was observed. The study demonstrated that the use of pentoxifylline was not particularly beneficial to patients and should be avoided in combination with riluzole. Variability in data collection was important, as not all patients were followed for the same length of time or at regular time intervals, with assessments occurring at different times depending on the patient. We observed the data over five distinct periods in order to develop 1-year prognostic models. The starting period denoted T_0 corresponds to the beginning of a patient's medical care, while successive periods of 3, 6, 9 and 12 months are denoted T_3, T_6, T_9 and T_{12}, respectively. These intervals were selected because they represent a reasonable trade-off that allows significant clinical changes to be captured without requiring excessive imputation, which could compromise the integrity of the data. Although a monthly approach could potentially offer finer accuracy, it would significantly increase the need for imputation given the irregularity and infrequency of assessments between patients in the datasets.

2.2 Predictors and features

The aim of this study was to predict survival as well as functional loss at one year in patients measured using the ALSFRS. In addition to the ALSFRS score, explanatory features included patients' age (year), gender (female or male), weight (kg.) and height (cm.), region of onset of first symptoms (bulbar or spinal), duration since onset of first symptoms and start of management (month), forced vital capacity or FVC (litre) which measures the maximum volume of air a patient can exhale after maximum inspiration used to assess lung function and degree of airway obstruction, pulse rate (b/min) and diastolic and systolic blood pressure (DBP and SBP) (mmHg). The revised version of the Amyotrophic Lateral Sclerosis Rating Scale (ALSFRS-R) is currently the most commonly used reference for assessing ALS. This version includes a revision of the final question on respiratory functions, which is now divided into three separate items: Dyspnoea, Orthopnoea and Respiratory Failure. This modification adds 8 points to the overall scale, bringing the total possible score to 48. Although this version is more informative, much of the data for this feature was missing for the majority of patients (69.39% missing data, as reported in Table 1). This would significantly reduce the amount of data available for our machine learning analysis. Therefore, our study mainly focused on the unrevised version of the scale. Nevertheless, we conducted several experiments on the sub-sample with ALSFRS-R scores for comparative purposes, and the results are available in the appendix (see appendix A). For patients with an ALSFRS-R score, a conversion was done by omitting the last two items of the respiratory function question, retaining only the initial dyspnoea item. This approach is consistent with the guidelines established by PRO-ACT [11, 20].

The existing features, while essential, may not be sufficient to capture the underlying relationships in the data. Therefore within the experiment several derived features have been added to the existing data. These are created from the initial raw data in order to increase the ability of the models to capture significant patterns and

Table 1: Distribution of missing data as they were in the original data at T_0 before cleaning. For each feature, the corresponding percentage of missing data is indicated. Q1, Q2, ..., Q10, are the ten questions used to calculate the ALSFRS score. Q10 is replaced by Q10a, Q10b and Q10c to calculate the ALSFRS-R score.

Features	**PRO ACT**	**Exonhit**	**Overall**
Initial sample size	10,723	400	11,123
Gender	0.00	0.00	0.00
Age	28.17	0.00	27.16
Weight	39.15	2.50	37.83
Height	37.54	0.25	36.20
Onset	12.40	0.00	11.96
ALSFRS	36.22	1.75	34.98
ALSFRS-R	65.18	100.00	66.43
Q1 Speech	39.29	0.50	37.89
Q2 Salivation	39.30	0.25	37.89
Q3 Swallowing	39.29	0.50	37.89
Q4 Handwriting	39.30	0.75	37.91
Q5 Cutting	39.34	0.25	37.93
Q5a Gastrotomy handling	36.21	0.25	34.92
Q6 Dressing and hygiene	39.29	0.50	37.89
Q7 Turning in bed	39.29	0.50	37.89
Q8 Walking	39.29	0.25	37.89
Q9 Climbing stairs	39.29	0.25	37.89
Q10 Respiration	39.29	0.25	37.89
Q10a Dyspnea	68.25	100.00	69.39
Q10b Orthopnea	68.25	100.00	69.39
Q10c Respiratory Insufficiency	68.25	100.00	69.39
Symptom Duration	35.87	0.00	34.58
Forced vital capacity	22.93	0.50	22.13
Pulse	32.09	1.00	30.97
Diastolic blood pressure	32.03	1.00	30.92
Systolic blood pressure	32.03	1.00	30.92

improve the performance of the learning methods. Firstly, we calculated the Body Mass Index (BMI) from weight and height to measure the corpulence of the different patients (< 18.5 means underweight while > 40 means morbid obesity). We then calculated 4 additional subscores from the ALSFRS scale relating to the bulbar areas ($Q1 + Q2 + Q3$), the upper limbs ($Q4 + Q5$), the trunk ($Q6 + Q7$) and the lower limbs ($Q8 + Q9$). The ALSFRS/ALSFRS-R scales are useful for measuring overall loss of function, but they do not take into account differences in the location and severity of symptoms between patients. These differences can have a major impact on patients' quality of life and on the treatment strategies to be implemented. To overcome these limitations, several clinical staging methods have been proposed for ALS (see Table 2). King's system is not directly based on the ALSFRS-R subscores, but it can be approximated with a concordance of over 90% [22]. It categorises patients into five stages based on the areas of the central nervous system affected by the disease and the speed of progression, focusing on the bulbar area, arms and legs. Stages 1 to 3 represent the number of areas affected, stage 4 represents the need for a feeding tube and/or

Table 2: The different ALS staging systems as indicated in [21]. NIV=non-invasive ventilation.

Stage	King's	MiToS	Fine'till 9
0	No region affected	No functional loss	No ALSFRS-R subscore ≤ 9
1	One region affected	Loss of one region	One subscore ≤ 9
2	Two regions affected	Loss of two regions	Two subscores ≤ 9
3	Three regions affected	Loss of three regions	Three subscores ≤ 9
4	Need for NIV or feeding tube	Loss of four regions	Four subscores ≤ 9
5	Death	Death	Death

the need for non-invasive respiratory assistance (NIV), and stage 5 represents death. The Milano-Torino Staging System (MiToS) is derived from the ALSFRS-R scale [23]. It subdivides disease progression into six clinical stages, ranging from 0, meaning no major functional loss, to 5, corresponding to death. The stage is assigned on the basis of patients' functional needs: loss of mobility, communication, ability to swallow and breathe independently. Each stage corresponds to the need for specific clinical assistance, making it possible to standardise the assessment of ALS progression, guide clinical interventions and facilitate communication between healthcare professionals and patients. While the ALSFRS-R directly measures functional ability in several domains, the King's score and the MiToS score offer complementary perspectives for understanding the progression of ALS, with King's offering better accuracy in the early and middle stages of the disease, while MiToS is more accurate at the end of the disease [24]. The final system, Fine'till 9 (FT9), initially breaks down the ALSFRS-R score into 4 distinct subscores, each with a maximum of 12 points, corresponding to the bulbar area, fine motor skills (upper limbs), gross motor skills (lower limbs) and respiratory [21]. The stage is assigned according to the number of subscores less than or equal to 9. Unlike the MiToS score, which focuses on patients' functional needs, and the King's score, which assesses progression according to the anatomical regions affected, the FT9 score attempts to combine elements of both approaches. We adapted these 3 systems to the ALSFRS score in order to include them in the experiment. The ability to breathe appropriately for MiTos and the need for NIV for King's were determined by Q5a and Q10. For FT9 we reported the respiratory subscore at Q10 (4 points) and considered a failure when the subscore was below 3, keeping the same ratio of 0.75 (9/12) as in the original system. Besides, the stage of the disease is not the same for each patient at T_0. We therefore added an additional feature measuring the decline rate, defined as follows:

$$decline\ rate = \frac{ALSFRS_{Max} - ALSFRS_{T_0}}{Symptom\ Duration} \tag{1}$$

With $ALSFRS_{Max} = 40$, the score for a patient in good health, the $ALSFRS_{T_0}$ score at the start of the patient's medical care and *Symptom Duration* the time in months between the appearance of the first symptoms and T_0.

Table 3: Distribution of the main features in the dataset after cleaning at T_0. For each quantitative feature, the mean, standard deviation and range of its values are shown, while for each of the qualitative features the percentage distribution between the different classes is indicated. In addition to the 11 explanatory features, the ALSFRS feature is broken down into 11 different features. There is one feature per question, except for the fifth question which contains two. We also have 21 derived features. This gives a total of 43 explanatory features. The target features are in bold and the number of patients present in brackets.

Features	**PRO-ACT** (4659)				**Exonhit** (384)				**Overall** (5043)			
	Avg	**Std**	**Min**	**Max**	**Avg**	**Std**	**Min**	**Max**	**Avg**	**Std**	**Min**	**Max**
Gender (0: Female)	0.62	–	–	–	0.64	–	–	–	0.63	–	–	–
Age (years)	55.57	11.73	18	84	55.26	11.96	21.96	77.91	55.55	11.74	18	84
Weight (kg)	75.53	18.09	39.1	263	70.12	14.12	41	130	75.12	17.87	39.1	263
Height (cm)	170.3	9.98	131	205	170.03	8.6	147	192	170.28	9.88	131	205
BMI (kg/m^2)	25.97	5.55	14.77	91.64	24.17	4.12	13.7	44.98	25.83	5.48	13.7	91.64
Onset (0: Spinal)	0.79	–	–	–	0.77	–	–	–	0.79	–	–	–
ALSFRS (/40)	29.76	5.72	7	40	27.52	6.53	9	39	29.59	5.81	7	40
Bulbar Score (/12)	10.13	2.32	0	12	9.64	2.54	2	12	10.09	2.34	0	12
Upper Limbs Score (/8)	5.63	2.14	0	8	5.01	2.41	0	8	5.58	2.17	0	8
Trunk Score (/8)	5.6	1.88	0	8	4.98	2.19	0	8	5.55	1.92	0	8
Lower Limbs Score (/8)	4.74	2.29	0	8	4.29	2.32	0	8	4.7	2.29	0	8
Mitos Total (/4)	0.42	0.72	0	4	0.62	0.77	0	2	0.43	0.72	0	4
Kings Total (/4)	2.13	0.88	0	4	2.28	0.78	0	3	2.14	0.87	0	4
FT9 Total (/4)	1.91	1.08	0	4	2.19	1.06	0	4	1.93	1.08	0	4
Decline Rate	0.6	0.5	0	9.43	0.62	0.44	0.03	3.14	0.6	0.49	0	9.43
S. Duration (months)	22.34	14.59	0.53	287.19	24.78	11.94	4	58	22.53	14.42	0.53	287.19
FVC (litres)	3.28	1.12	0	7	2.53	1.01	0.24	6.13	3.22	1.13	0	7
Pulse	77.05	12.24	42	135	78.73	12.21	46	120	77.18	12.24	42	135
DBP (mmHg)	81.69	10.49	30	130	83.7	11.49	54	130	81.84	10.58	30	130
SBP (mmHg)	131.19	17.2	80	210	138.03	18	93	200	131.71	17.36	80	210
ALSFRS T_{12} (/40)	18.59	11.88	0	40	16.45	12.17	0	39	18.42	11.91	0	40
Survived (0: Death)	0.8	–	–	–	0.72	–	–	–	0.79	–	–	–

2.3 Managing missing data

In this study, we had a significant amount of missing data in several features (see Table 1). We took the decision to remove patients with missing data in order to ensure the quality and robustness of the prediction models. Indeed, missing data may introduce biases and affect the performance of machine learning algorithms. Several methods exist for dealing with missing data, such as imputation, which involves replacing missing values with estimates. However, even the most effective imputation methods can fail by increasing bias and/or decreasing the interpretability of the models [25], particularly in complex and heterogeneous datasets such as those used in our case. By removing patients with missing data, we ensure that models are trained on complete and consistent data, minimising the risks of bias and overfitting. This approach also simplifies the process of modelling and interpreting results, avoiding the complications associated with multiple imputation and the advanced analysis techniques required to

deal with incomplete data. Table 3 indicates the number of patients remaining and the different features present at T_0 in the cleaned dataset.

2.4 Model evaluation criteria

The predictive quality of our classification models was assessed by calculating the number of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). Based on these criteria, several metrics were derived:

- Sensitivity: $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FP}$
- Precision: $\frac{TP}{TP+FP}$
- Balanced accuracy: $\frac{Sensitivity+Specificity}{2}$

For our regression models, we evaluated the error rate using:

- Root Mean Square Error (RMSE): $\sqrt{\frac{\sum_{i=1}^{N}(pred_i - real_i)^2}{N}}$
- Coefficient of determination (R^2): $1 - \frac{\sum_{i=1}^{N}(real_i - pred_i)^2}{\sum_{i=1}^{N}(real_i - \overline{real})^2}$
- Adjusted coefficient of determination (R^2_{adj}): $1 - \frac{(1-R^2)\times(N-1)}{N-k-1}$
- Pearson's correlation coefficient (PCC): $\frac{\sum_{i=1}^{N}(pred_i - \overline{pred})\times(real_i - \overline{real})}{\sigma_{pred}\times\sigma_{real}}$

With N, the number of patients, $pred_i$, the value predicted by the model, $real_i$, the actual value for patient i, $\overline{real}$ represents the average of the values, k the number of features used in the model and σ represents the standard deviation.

2.5 Feature selection

In the medical context, the data used for classification or regression presents a particular challenge in terms of quantity and quality. It is common to have few complete data, but many features associated with this data. This can make the models obtained by learning difficult to generalise (overfitting) without additional precautions. Reducing the number of features to exactly what is needed to predict the value of the target feature is a possibility well suited to this context. We then looked for an optimal solution, i.e. the smallest subset that would give the best-performing model. Generally speaking, and this is the case in the medical field, the selection of features increases the interpretability of the model while reducing the amount of information needed to be collected, which in turn reduces the effort involved in collecting and entering information. Ideally, all possible combinations of subsets of features should be evaluated to find the best solution. However, this approach is impossible because of the combinatorial explosion that results when the number of features is large. Indeed, with our 43 features, there are 2^{43} different subsets to explore, which in the current state of knowledge cannot be achieved in a reasonable time. The feature selection problem is considered NP-hard [26, 27], i.e. its complexity increases exponentially with the number of features. Nevertheless, several approximation methods are available.

2.5.1 Embedded methods

Embedded methods integrate feature selection directly into the model learning process, using techniques like Lasso and Elastic Net in linear classification methods (e.g., logistic regression, ridge regression, SVMs) to regularise or penalise certain features, potentially reducing their weights or eliminating them. Decision tree-based methods, such as random forests, assess feature importance by evaluating their impact on prediction accuracy and calculating the average information gain during tree construction [11, 28].

2.5.2 Filter methods

Filter methods evaluate the relationship between explanatory features and the target feature using statistical techniques, selecting features based on calculated scores. Techniques include the Spearman Correlation Coefficient (SCC) [29], which assesses monotonic relationships, and ANOVA, which compares group means to identify significant features [30]. Mutual Information (MI) measures the shared information between features [31], and Minimum Redundancy Maximum Relevance (MRMR) selects features with high relevance to the target while minimising redundancy [32–34]. ReliefF algorithm evaluates feature weights by distinguishing between nearby instances of different classes [35–37].

2.5.3 Metaheuristics

Metaheuristics are optimisation algorithms that efficiently solve complex problems like feature selection. These include Genetic Algorithm [38], Population Based Incremental Learning [39], Differential Evolution (DE) [40], Particle Swarm Optimisation [41], Tabu Search [42], and Simulated Annealing [43]. DE, which performed best in our studies, mimics natural evolutionary processes, using diversity among individuals for iterative improvement [44]. The main steps of the DE method are explained in detail in the Appendix B. Feature subsets are treated as individuals, and optimisation continues until a stopping criterion, like low diversity or a time limit, is met.

2.6 Model building process

First, we merged our two datasets into a single dataset for a total of 5043 patients. We then divided it into two distinct subsets: 75% of the patients (3782 patients, 3002 of whom survived the first year of treatment) were placed in a dataset reserved for training, and the remaining 25% (1261 patients, 1001 of whom survived one year of treatment) in an independent dataset reserved for testing. The features in the dataset have different scales, which can have a negative impact on the learning methods by slowing down their speed of convergence or biasing the results. In this particular case, it is common practice to carry out standardisation to ensure that each feature contributes equally to the model training and that features with larger scales do not dominate the others. We applied standardisation to transform and replace each X_i

feature into a new Z_i feature as follows:

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{2}$$

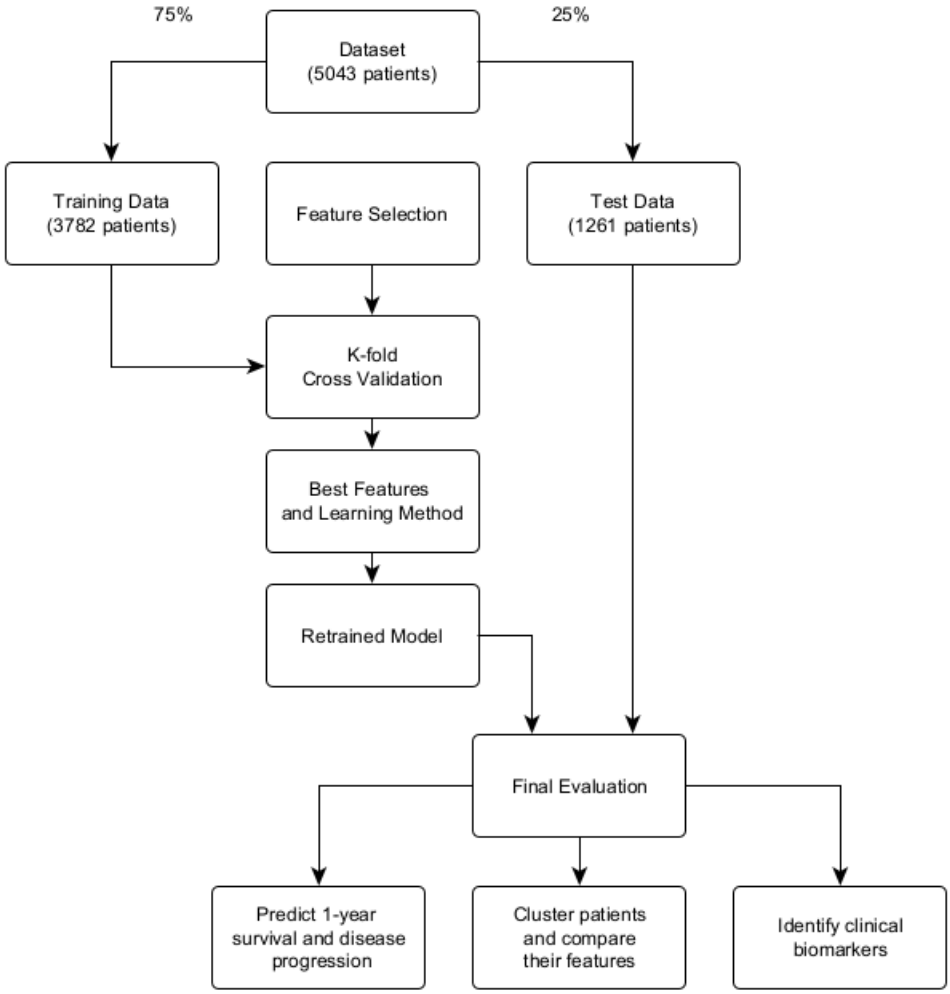
With μ the mean of X_i, and σ its standard deviation. By performing this transformation, all the features have a zero mean and a variance equal to one, bringing them to a common scale. To select the best model, we applied a k-fold cross-validation with $k = 10$. Unlike a simple split where a fixed portion of the data is reserved for validation, this method involves dividing the training set into k subsets (or folds). The training and validation process is then repeated iteratively on these folds. At each iteration, a different fold is used as the validation set, while the others are used for training. This cycle is repeated for each fold, allowing each observation to be used for both training and validation, but never at the same time. Model performance, measured by Balanced Accuracy in our study, is then averaged across all folds to provide an overall assessment. For the survival model, the folds were stratified to retain the percentage of patients in each class. This method minimises bias, variation and the risk of overfitting in estimating model performance, which is particularly important when few data is available [45, 46]. In order to avoid any potential data leakage, standardisation was applied to the training data at each fold (see equation 2). The means and standard deviations calculated for each explanatory feature in the training data were then used to transform the validation data appropriately. Full training was then performed on the 3782 patients and tested on the test data using the same data scaling procedure. The aim was to design the model with the most appropriate combination of learning and feature selection methods. Several learning methods were used in the experiment: logistic regression (LR), ridge regression (RR), k-nearest neighbours (KNN), gaussian naive bayes (GNB), random forest (RF), linear discriminant analysis (LDA) and Light Gradient Boosting Machine (LGBM). These learning methods were implemented using the Python libraries scikit-learn (1.3.2) [47] and lightgbm (3.3.5) [48]. We found a significant imbalance in the distribution of patients between the two classes in our dataset. Indeed, we have 79.38% of surviving patients (see Table 3) and only 20.62% of deceased patients. When classes are unbalanced, the model can be biased towards the majority class, which can lead to inaccurate predictions and poor performance for the minority class. For this reason, we adjusted the patient weights in the machine learning methods used for the minority class according to the following formula:

$$W_i = \frac{N}{C \times N_i} \tag{3}$$

Where W_i represents the weight assigned to patients in class i, N is the total number of patients in the dataset, C is the number of classes present in the data, and N_i is the number of patients belonging to class i. As a result, all the classes are considered with the same importance during the learning phase of the machine learning methods.

For DE, the choice of learning method has been integrated directly by associating a learning method with each feature subset (or individuals), in addition to the chromosomes that make it up. This increases the complexity of the search space but allows DE

Fig. 1: Diagram showing the main steps taken from the cleaned data to the construction of the final machine learning model.



to converge on the best performing method among those available, thus guaranteeing better predictive quality. For each filter method, the features were sorted according to the calculated relevance score (SCC, ANOVA, etc.) from the most relevant to the least relevant. Next, a learning was performed with each learning method for each possible value of k number of features ($k \in \{1, 43\}$), in order to retain the model offering the best predictive quality on average over all folds. MRMR was implemented using the Python library mrmr-selection (0.2.8) [49] and ReliefF using skrebate (0.62) [50]. Secondly, we stratified the patients in the test data into several clusters according to their survival rate estimated by the best performing model. We applied Kaplan-Meier

estimator to these patients, a statistical method used to estimate the survival function of a population from a sample of life-time data [51, 52]. The Kaplan-Meier curve is a graphical representation of this estimate as a function of time [11]. We used it in this study to determine whether there was a significant difference in the rate of disease progression between the clusters of patients identified. Finally, a prediction model of the evolution of the ALSFRS score T_3 to T_{12} was built from the same feature subset. This regression model was developed by carrying out a new cross-validation with 10 folds to select the best model from the set of learning methods. A summary of all the major steps is available in Figure 1.

3 Results

3.1 Prediction of 1-year survival

Table 4: Performance of the best model for predicting 1-year survival for each of the 6 feature selection methods in cross-validation; SCC, ANOVA, MI, MRMR, ReliefF, and DE. Abbreviations: LM = Learning Method, TN = True Negatives, FP = False Positives, FN = False Negatives, TP = True Positives, Sens. = Sensitivity, Spec. = Specificity, Prec. = Precision, Balanced = Balanced Accuracy and k = number of features selected. For the Balanced Accuracy score, the minimum and maximum scores obtained during cross-validation are indicated in brackets.

Methods	**LM**	**TN**	**FP**	**FN**	**TP**	**Sens.**	**Spec.**	**Prec.**	**Balanced**	**k**
w/selection	RR	584	196	783	2219	73.92	74.86	91.86	74.39 (66.8:77.7)	43
SCC	RR	584	196	783	2219	73.92	74.86	91.86	74.39 (66.8:77.7)	43
ANOVA	LR	580	200	767	2235	74.47	74.39	91.80	74.40 (67.3:77.9)	41
MI	RR	585	195	776	2226	74.13	75.00	91.92	74.58 (66.3:77.9)	32
MRMR	RR	585	195	784	2218	73.93	75.00	91.92	74.47 (66.9:78.5)	41
ReliefF	RR	584	196	783	2219	73.92	74.86	91.86	74.39 (66.8:77.7)	43
DE	LR	**603**	**177**	**757**	**2245**	**74.90**	**77.28**	**92.71**	**76.05 (68.6:79.8)**	**19**

The best performing features for maximising the Balanced Accuracy score were selected using 6 different feature selection methods (SCC, ANOVA, MI, MRMR, ReliefF and DE). For the DE metaheuristic, the algorithm was run for one hour with a population of 100 individuals, $F = 1$ and $CR = 0.5$. An analysis of the methods listed in the Table 4 suggests that the DE method with LR is more effective across all evaluation metrics, particularly in terms of the Balanced Accuracy score. This method was able to predict a patient's survival with a score of 76.05% (almost 80% in the best fold), compared with 74.39% without it, representing an improvement of 1.7 percentage points and 2 points in the best fold. In addition, the DE method considerably reduced the number of features required to build the model, with 19 features selected. We also obtained a Balanced Accuracy score on the test data of 76.33% (Sensitivity = 73.87%, Specificity = 79.23%, Precision = 93.16%), which is quite close to the values obtained during cross-validation, showing that the model can be generalised. Overall,

parametric regression methods (LR and RR) appear to be more effective than other types of learning methods on this type of data.

Reducing the number of features in a machine learning model also has the advantage of making it easier to interpret. Figure 2 suggests that the feature representing patients' FVC at T_0 is the most important of the features selected. The lower the FVC at T_0, the higher the risk of not surviving the first year of the disease (survival and FVC being strongly correlated [54]). Features such as age, duration since first symptoms, weight, height and ALSFRS (decline rate) also appear to have a significant impact on survival. Some of these features have already demonstrated their importance in the literature (e.g. FVC, age, duration since first symptoms) for anticipating the speed of patient decline [54, 55]. However, features such as pulse rate and blood pressure have not, to our knowledge, been reported as relevant factors in disease progression.

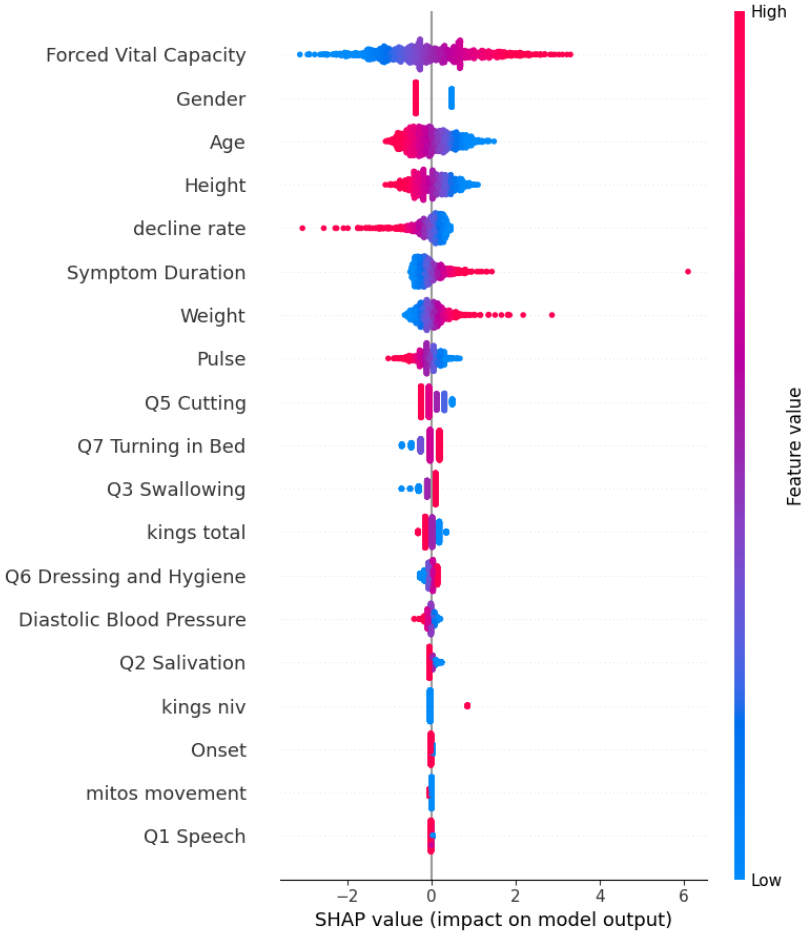
Table 5: Performances of the survival probabilities identified by the model. The Min Zone and Max Zone columns represent the lower and upper bounds of the different survival zones, e.g. the first row (0 to 20) represents patients for whom the model predicted a survival probability between 0% and 20%. The Accuracy column shows the proportion of correct model predictions in each survival zone and N samples the number of patients in each survival zone.

Min Zone (%)	**Max Zone (%)**	**Accuracy**	**N samples**
0	20	66.40	125
20	40	41.20	216
40	60	53.99	263
60	80	92.42	330
80	100	98.47	327

To further clarify the performance of our LR model we calculated the rate of correct predictions in different survival zones. The results displayed in Table 5 suggest that the model is very accurate for high survival zones (60% to 100%), which is a positive sign of its ability to correctly identify patients with a high survival probability. On the other hand, performance is much more moderate for medium (40% to 60%) and low (0% to 40%) survival zones.

The statistics provided in Table 6 allow us to analyse the features of patients in different survival zones predicted by our model. By examining how the means and standard deviations of the features change across the survival zones, it is possible to better understand which features influence the model's predictions. The data in Table 6 and Figure 2 highlight notable disparities between patients in different clusters. In particular, there is a trend towards higher BMI (height and weight) and FVC corresponding to higher survival rates (23.41 and 1.86 on average for patients with high mortality versus 27.93 and 4.23 on average for those with high survival). In addition, a decrease in age, pulse and decline rate is associated with improved survival rates (61.82, 84.66 and 1.18 for patients with high mortality versus 46.73, 73.58 and 0.34).

Fig. 2: The features with the greatest impact on the model according to the test data. The present data was calculated using the Shap (SHapley Additive exPlanation) value proposed by Lundberg et al. [53]. The x-axis represents the impact of the feature on the model output, while the y-axis represents the names of the most important features for the model. Each point represents a Shap value and the thickness represents the density. The colour represents whether a value is high (red) or low (blue) depending on the value interval of a feature e.g. FVC is the feature that has the most impact in calculating the model output.



It is important to note that although some features appear to be strongly correlated with survival (e.g. ALSFRS score), they were not selected by our feature selection step, suggesting redundancy with other features in our data.

The Kaplan-Meier curves show the survival probabilities for the patients, divided into clusters according to the survival probabilities predicted by the LR model (Figure 3). These curves clearly indicate that the clusters of patients predicted by the model

Table 6: Statistics for features calculated on patients from the test data present in the different survival zones defined by our model, e.g. the column labelled "0:20 (106)" indicates the zone where the model identifies a survival rate of less than 20%, 106 being the number of patients in this zone; the average (Avg) height is 169.64cm with a standard deviation (Std) of 9.08cm.

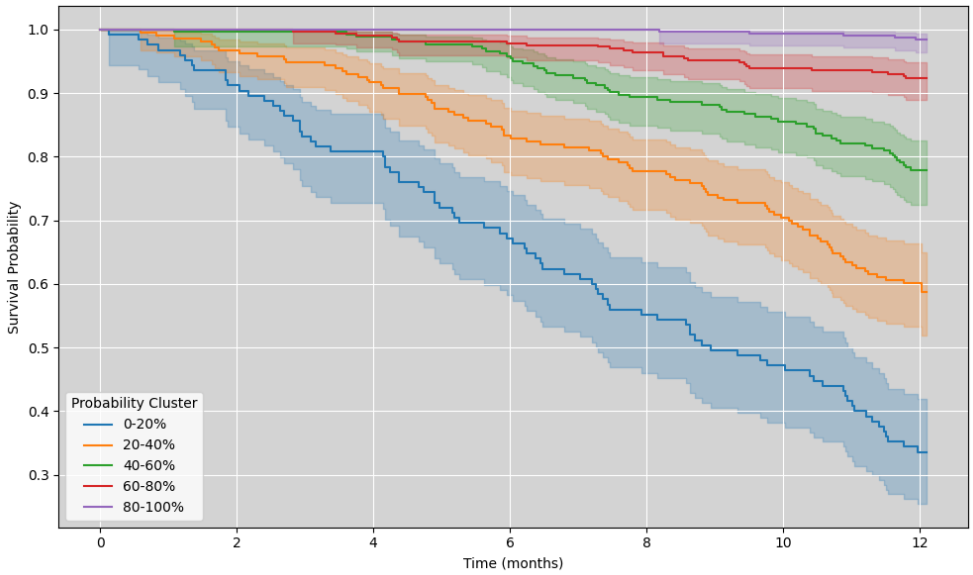
Feature	Survival rates (Number of patients)									
	0:20 (125)		20:40 (216)		40:60 (263)		60:80 (330)		80:100 (327)	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
Gender (0:Female)	0.65	0.48	0.57	0.50	0.58	0.49	0.63	0.48	0.72	0.45
Age (years)	61.82	10.16	60.84	10.09	59.87	10.37	55.02	10.43	46.73	10.22
Weight (kg)	67.04	12.37	69.4	12.68	71.97	13.81	76.63	15.4	84.07	22.84
Height (cm)	169.19	9.14	168.9	9.24	167.96	9.26	170.66	9.67	173.48	9.89
Onset (0:Spinal)	0.65	0.48	0.72	0.45	0.75	0.43	0.78	0.42	0.87	0.33
Q1 Speech (/4)	2.23	1.27	2.89	1.07	3.04	1.09	3.31	0.88	3.57	0.71
Q2 Salivation (/4)	2.83	1.17	3.25	0.92	3.27	0.92	3.46	0.79	3.72	0.57
Q3 Swallowing (/4)	2.73	1.05	3.27	0.80	3.44	0.70	3.55	0.62	3.78	0.44
Q5 Cutting (/4)	1.99	1.25	2.48	1.30	2.52	1.24	2.91	1.13	2.87	1.12
Q6 Dressing (/4)	1.86	1.18	2.32	1.16	2.42	1.07	2.73	0.96	2.82	0.98
Q7 Turning in Bed (/4)	2.31	1.16	2.70	1.14	2.91	0.95	3.21	0.82	3.39	0.75
S. Duration (months)	18.82	12.16	18.67	10.47	22.67	12.33	23.14	13.60	27.40	21.21
FVC (litres)	1.86	0.82	2.52	0.73	2.83	0.72	3.40	0.83	4.23	1.00
Pulse (b/min)	84.66	14.01	80.12	11.78	76.85	12.13	76.11	11.26	73.58	10.66
DBP (mmHg)	83.90	11.55	82.52	10.22	82.11	10.44	82.26	10.57	81.65	10.55
Mitos Movement (0:No)	0.42	0.49	0.29	0.45	0.24	0.43	0.12	0.32	0.09	0.29
Kings NIV (0:No)	0.08	0.27	0.03	0.18	0.02	0.14	0.02	0.14	0.02	0.14
Kings Total (/4)	2.80	0.61	2.44	0.74	2.22	0.79	2.02	0.86	1.76	0.90
Decline Rate	1.18	0.74	0.80	0.40	0.60	0.34	0.48	0.31	0.34	0.23

have distinct survivals. Indeed, the blue curve (0-20%) drops the fastest, indicating that patients in this cluster have the lowest probability of survival, while the purple curve (80-100%) is always the highest, showing that patients in this cluster have the highest survival probability. Although the model has an increasing level of uncertainty as the survival rate decreases, we obtain a concordance index (C-Index) of 0.81, which means that our model is able to correctly predict the order of survival events in 81% of cases. This indicates good discrimination between patients who survive longer and those who survive shorter. Moreover, the clear differences between the curves show that the probability clusters predicted by the model are well separated in terms of actual survival, which testifies to the discriminating capacity of the model (Global log-rank test p-value ≤ 0.0001).

3.2 Prediction of 1-year ALSFRS score

As with the survival model, only the data available at baseline were taken into account (i.e. at T_0) in order to predict the evolution of the ALSFRS score. The features selected are similar to those shown in Figure 2. From T_0, we predicted patients' ALSFRS every 3 months until T_{12}. Consequently, only patients who survived the first year of the disease were included in this model, which differs from the approach used for the

Fig. 3: Kaplan-Meier curves calculated from the test data. Each curve represents one of 5 survival clusters (see Table 5). The coloured areas around the curves represent confidence intervals. The wider the zones, the greater the uncertainty.



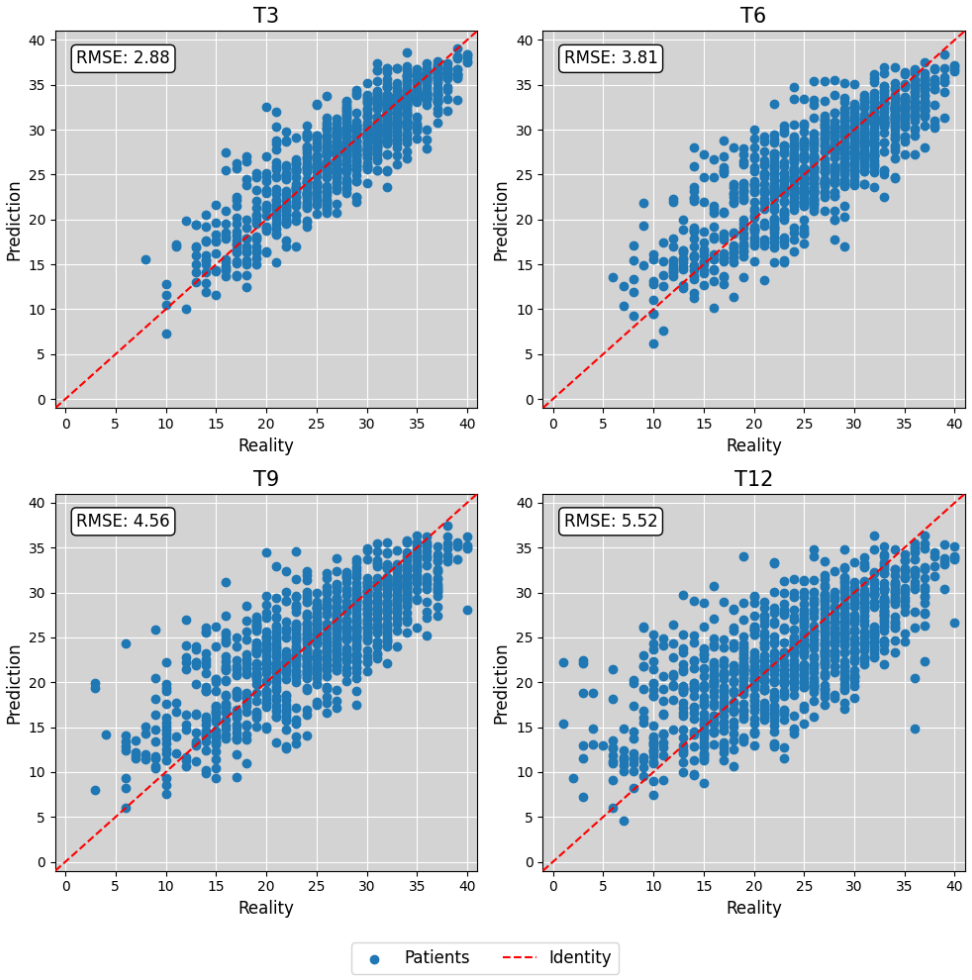
classification model. By cross-validation, the RF method was the one that allowed us to obtain the best performing model. The method was selected on the basis of the global RMSE obtained over all the periods.

Table 7: Performance of the model to predict the evolution of ALSFRS by keeping the selected features present in the 2 section on the validation and test data. Abbreviations: RMSE = Root Mean Squared Error, R^2 = Coefficient of determination, R^2_{ajd} the adjusted version and PCC = Pearson's correlation coefficient.

Dataset	**Scores**	$\mathbf{T_3}$	$\mathbf{T_6}$	$\mathbf{T_9}$	$\mathbf{T_{12}}$
Validation	RMSE	3.143 (2.90:3.29)	4.069 (3.86:4.25)	4.976 (4.73:5.29)	5.835 (5.47:6.17)
	R^2	0.765	0.673	0.590	0.503
	R^2_{ajd}	0.764	0.671	0.587	0.500
	PCC	0.875	0.820	0.768	0.701
Test	RMSE	2.880	3.808	4.563	5.516
	R^2	0.784	0.697	0.628	0.529
	R^2_{ajd}	0.774	0.683	0.612	0.508
	PCC	0.887	0.836	0.794	0.729

The results shown in Table 7 and Figure 4 suggest that the ALSFRS evolution prediction model works well for short-term predictions (T_3 and T_6), but that the accuracy progressively decreases for longer prediction horizons (T_9 and T_{12}). The RMSE,

Fig. 4: The calibration curve of the model on the test data for each period. The blue points are the ALSFRS scores of the patients placed according to their actual value and that predicted by the model. The red line represents the identity function (Prediction=Reality).



R^2, and PCC all confirm this trend. Despite this decrease in accuracy, the predictions remain correlated with the actual values, which shows that the model captures the general trends in the evolution of the ALSFRS well, but that it could benefit from further refinement for longer-term predictions. Previous studies [6, 11] have also reached similar conclusions using other learning methods such as recursive neural networks [6] and random forests [11] to calculate the slope of ALSFRS between T_3 and T_{12}. This may indicate some limitations for predicting the ALSFRS score. Nevertheless, we obtained better scores than those reported in the literature, with a higher PCC (0.729 vs. 0.472) as well as a higher R^2 (0.529 vs. 0.219) while we only used data available at

T_0 indicating a better accuracy of our model [6, 11]. Random forest models are also easier to interpret than neural networks. Overall, our model is able to explain more than 53% of the variations in the ALSFRS score from T_0 to T_{12}.

4 Discussion

Our study reinforces the idea that the use of machine learning methods can provide significant help in predicting survival and disease progression in ALS. Feature selection is often barely addressed or neglected in the literature. Nevertheless, we have shown that the choice of an effective feature selection method can have a significant impact on the quality of machine learning models. Indeed, by identifying the features with the greatest impact and using logistic regression, we obtained a Balanced Accuracy of 76.05% with 19 features in predicting the 1-year survival probabilities of patients at T_0 instead of 74.39% with 43 features without it. This method also reduces the "black box" aspect of some classification and regression models and improve their interpretability. With this methodology, it is easier to identify the specific features of patients in the zones compared to our previous research using the UMAP method on the same dataset [15, 16]. This underlines the crucial importance of feature selection in the modelling process and reinforces the idea that even with the limited information available at T_0, it is possible to make accurate predictions. Using a logistic regression model and the Kaplan-Meier estimator we were able to identify clusters among patients with similar profiles. It is also conceivable that features not included in our data, such as specific co-morbidity factors, individual genetic variations or quality of life aspects, could have a significant influence on patient survival.

Our results in ALSFRS prediction showed accurate results at 3 months and 6 months, but not at longer term. Although we use fewer features, our results align with previous studies [6, 11] using learning methods such as recursive neural networks [6] and random forests [11, 56] to calculate the slope of the ALSFRS between T_3 and T_{12}, which may indicate some limitations in predicting the ALSFRS score. It is important to recognise that the variability observed in disease progression as shown in Figure 4 reflects the challenges inherent in modelling with subjective data, such as ALSFRS scores. Our model produces reliable predictions for the majority of the population at 3 months, suggesting its potential utility in clinical settings for predicting short-term decline. Our models have been designed to be applied at different stages of a patient's disease progression, allowing them to be adjusted and recalibrated as more data becomes available. This iterative approach refines predictions, increases statistical power [57] and improves the applicability of the model to personalised predictive medicine over time.

However, the continuing complexity of predicting this metric can be explained by a number of factors. Data from the PRO-ACT and Exonhit databases, while valuable, may not be sufficient to establish a robust prognostic model. In addition, the fact that some ALSFRS subscores are based on the subjective perception of the patient and clinician introduces inter-individual variability which may influence the

results. Aggregation of subscores raises issues, notably a loss of sensitivity to subtle changes in specific areas of motor function. This approach can attenuate variations in a particular subdomain when combined with others, making it difficult to detect specific changes. ALS, with its great heterogeneity of symptoms and progression, accentuates this complexity. In the advanced stages of the disease, when scores may stabilise, this variability in the way patients perceive and report their symptoms may influence the results. This heterogeneity, combined with the inherent limitations of the metrics, can have significant negative consequences on the accuracy of ALSFRS score prediction, particularly in the long term. To overcome these difficulties, the use of objective biomarkers and complementary measures such as genetic features [58, 59] and imaging [60], is essential to monitor the progression of ALS. A multidimensional approach, integrating various clinical assessments and biological parameters would provide a more complete picture of the disease and its impact on motor function and quality of life in patients. This enriched approach would also facilitate the application of machine learning techniques for prediction.

It should be noted that comparison with other similar approaches is complex due to the specificities of the PRO-ACT and Exonhit databases, notably the absence of genetic data and the methodology we employ. The metrics used may also differ, as may the preprocessing of the data. In addition, PRO-ACT and Exonhit are based on a specific population, those taking part in clinical trials, which may influence the quality of our models when applied to real-life data. The addition of real-life data could reduce the current bias and significantly increase the robustness and predictive quality of our various models. Nevertheless, our methodology differs in that it incorporates features that are not taken into account by the most recent survival models. Furthermore, whereas some studies use backward propagation, such as the ENCALS survival prediction model [13] or the random forest used by Pancotti et al [11] or the Origent survival model [12] for feature selection, our study uses a heuristic-based approach. Our previous research has demonstrated the effectiveness of differential evolution over sequential feature selection methods, such as backward propagation, and embedded methods, such as random forest, in capturing the complex relationships between features in ALS progression [61, 62]. Although the inclusion of a feature selection cycle, in particular using our approach slows down the construction of the model, its application to new patients requires only a few milliseconds of computation once the relevant data for prediction is provided. This remains true even on low-capacity devices. The complexity lies solely in building the model, not in using it. This approach has considerable potential to make a significant contribution to the medical community by providing a better understanding of individual variations in the progression of ALS. It paves the way for the development of more personalised treatment strategies, tailored to the specific features of each patient. The use of heuristics means that relevant features can be selected efficiently, while ensuring that they are easily accessible to clinicians when assessing patients.

Furthermore, it is important to mention that our approach is not limited to ALS. Its flexibility allows it to be applied to other diseases, other medical issues or other

prediction problems, provided that a sufficient amount of structured data is available. This generalisation of the methodology increases its relevance and value, providing a reproducible and adaptable tool for widespread applications in the medical field. As a result, our methodology can make a significant contribution to the advancement of medical knowledge and practice, encouraging research and innovation beyond the specific context of ALS.

Declarations

Data availability statement

The study used training data from the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trial) database. The PRO-ACT database contains data provided by members of the PRO-ACT Consortium, and access is granted via registration on their website. The cleaned dataset is available in our github repository [63].

Code availability statement

All our data and algorithms are available on github [63, 64]. A web application, where it is possible to use our models in real time, was developed during this experiment and is available on the heroku servers [65].

Acknowledgements

Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organisations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: ALS Therapy Alliance; Cytokinetics, Inc.; Amylyx Pharmaceuticals, Inc.; Knopp Biosciences; Neuraltus Pharmaceuticals, Inc.; Neurological Clinical Research Institute, MGH; Northeast ALS Consortium; Novartis; Prize4Life Israel; Regeneron Pharmaceuticals, Inc.; Sanofi; Teva Pharmaceutical Industries, Ltd.; The ALS Association. The findings reported in this paper follow the guidelines recommended by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [66].

Author contributions statement

T.A. designed and carried out the experiment, analysed the data and wrote the first draft of the manuscript. J-F.P-P., F.D. contributed to the design of the study, the analysis of the data and the drafting of the manuscript. P-F.P. provided medical expertise and helped analyse the data. All the authors have read and approved the final version of the manuscript.

Declaration of interest

No conflict of interests.

References

[1] Swinnen B, Robberecht W. The phenotypic variability of amyotrophic lateral sclerosis. Nature Reviews Neurology. 2014 Nov;10(11):661–670. https://doi.org/10.1038/nrneurol.2014.184.

[2] Cedarbaum JM, Stambler N. Performance of the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) in multicenter clinical trials. Journal of the Neurological Sciences. 1997;152:s1–s9. https://doi.org/https://doi.org/10.1016/S0022-510X(97)00237-2.

[3] Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. Journal of the Neurological Sciences. 1999;169(1):13–21. https://doi.org/https://doi.org/10.1016/S0022-510X(99)00210-5.

[4] Antoniadi AM, Galvin M, Heverin M, Hardiman O, Mooney C. Prediction of caregiver burden in amyotrophic lateral sclerosis: a machine learning approach using random forests applied to a cohort study. BMJ Open. 2020;10(2). https://doi.org/10.1136/bmjopen-2019-033109.

[5] Huang Z, Zhang H, Boss J, Goutman SA, Mukherjee B, Dinov ID, et al. Complete hazard ranking to analyze right-censored data: An ALS survival study. PLOS Computational Biology. 2017 12;13(12):1–21. https://doi.org/10.1371/journal.pcbi.1005887.

[6] Tang M, Gao C, Goutman SA, Kalinin A, Mukherjee B, Guan Y, et al. Model-Based and Model-Free Techniques for Amyotrophic Lateral Sclerosis Diagnostic Prediction and Patient Clustering. Neuroinformatics. 2019 Jul;17(3):407–421. https://doi.org/10.1007/s12021-018-9406-9.

[7] Hothorn T, Jung HH. RandomForest4Life: A Random Forest for predicting ALS disease progression. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2014;15(5-6):444–452. PMID: 25141076. https://doi.org/10.3109/21678421.2014.893361. https://doi.org/10.3109/21678421.2014.893361.

[8] Schuster C, Hardiman O, Bede P. Survival prediction in Amyotrophic lateral sclerosis based on MRI measures and clinical characteristics. BMC Neurology. 2017 Apr;17(1):73. https://doi.org/10.1186/s12883-017-0854-x.

[9] Westeneng HJ, Debray TPA, Visser AE, van Eijk RPA, Rooney JPK, Calvo A, et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. The Lancet Neurology. 2018 May;17(5):423—433. https://doi.org/10.1016/s1474-4422(18)30089-9.

[10] van der Burgh HK, Schmidt R, Westeneng HJ, de Reus MA, van den Berg LH, van den Heuvel MP. Deep learning predictions of survival based on MRI in

amyotrophic lateral sclerosis. NeuroImage: Clinical. 2017;13:361–369. https://doi.org/https://doi.org/10.1016/j.nicl.2016.10.008.

[11] Pancotti C, Birolo G, Rollo C, Sanavia T, Di Camillo B, Manera U, et al. Deep learning methods to predict amyotrophic lateral sclerosis disease progression. Scientific Reports. 2022 Aug;12(1):13738. https://doi.org/10.1038/s41598-022-17805-9.

[12] Danielle B, James D B, Sabrina P, Jonathan D G, Christina F, Jonavelle C, et al. Development and validation of a machine-learning ALS survival model lacking vital capacity (VC-Free) for use in clinical trials during the COVID-19 pandemic. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2021;22(sup1):22–32. PMID: 34348539. https://doi.org/10.1080/21678421.2021.1924207. https://doi.org/10.1080/21678421.2021.1924207.

[13] Westeneng HJ, Debray TPA, Visser AE, van Eijk RPA, Rooney JPK, Calvo A, et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. The Lancet Neurology. 2018 May;17(5):423–433. https://doi.org/10.1016/S1474-4422(18)30089-9.

[14] Taguchi Yh, Iwadate M, Umeyama H. Heuristic principal component analysis-based unsupervised feature extraction and its application to gene expression analysis of amyotrophic lateral sclerosis data sets. In: 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2015. p. 1–10. Available from: https://doi.org/10.1109/CIBCB.2015.7300274.

[15] Grollemund V, Pradat PF, Querin G, Delbot F, Le Chat G, Pradat-Peyre JF, et al. Machine Learning in Amyotrophic Lateral Sclerosis: Achievements, Pitfalls, and Future Directions. Frontiers in Neuroscience. 2019;13:135. https://doi.org/10.3389/fnins.2019.00135.

[16] Grollemund V, Chat GL, Secchi-Buhour MS, Delbot F, Pradat-Peyre JF, Bede P, et al. Development and validation of a 1-year survival prognosis estimation model for Amyotrophic Lateral Sclerosis using manifold learning algorithm UMAP. Scientific Reports. 2020 Aug;10(1):13378. https://doi.org/10.1038/s41598-020-70125-8.

[17] : Pooled Resource Open-Access ALS Clinical Trials Database. Available from: https://ncri1.partners.org/proact.

[18] Atassi N, Berry JD, Shui AM, Zach N, Sherman A, Sinani E, et al. The PROACT database. Neurology. 2014;83:1719 – 1725. https://doi.org/10.1212/WNL.0000000000000951.

[19] Meininger V, Asselain B, Guillet P, Leigh P, Ludolph A, Lacomblez L, et al. Pentoxifylline in ALS: a double-blind, randomized, multicenter, placebo-controlled

trial. Neurology. 2006 January;66(1):88—92. https://doi.org/10.1212/01.wnl.0000191326.40772.62.

[20] : ALS Prediction Prize4Life Challenge. Available from: https://ncri1.partners.org/ProACT/Home/ALSPrize.

[21] Nimish J Thakore TGK Brittany R Lapin, Pioro EP. Deconstructing progression of amyotrophic lateral sclerosis in stages: a Markov modeling approach. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2018;19(7-8):483–494. PMID: 30001159. https://doi.org/10.1080/21678421.2018.1484925. https://doi.org/10.1080/21678421.2018.1484925.

[22] Balendra R, Jones A, Jivraj N, Steen IN, Young CA, Shaw PJ, et al. Use of clinical staging in amyotrophic lateral sclerosis for phase 3 clinical trials. Journal of Neurology, Neurosurgery & Psychiatry. 2015;86(1):45–49. https://doi.org/10.1136/jnnp-2013-306865. https://jnnp.bmj.com/content/86/1/45.full.pdf.

[23] Chiò A, Hammond ER, Mora G, Bonito V, Filippini G. Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis. Journal of Neurology, Neurosurgery & Psychiatry. 2015;86(1):38–44. https://doi.org/10.1136/jnnp-2013-306589. https://jnnp.bmj.com/content/86/1/38.full.pdf.

[24] Fang T, Al Khleifat A, Stahl D, Torre C, Murphy C, Young C, et al. Comparison of the King's and MiToS staging systems for ALS. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2017 01;18:1–6. https://doi.org/10.1080/21678421.2016.1265565.

[25] Kaushal S. Missing data in clinical trials: Pitfalls and remedies. International journal of applied & basic medical research. 2014 09;4:S6–7. https://doi.org/10.4103/2229-516X.140707.

[26] Chen B, Hong J, Wang Y. The minimum feature subset selection problem. Journal of Computer Science and Technology. 1997 Mar;12(2):145–153. https://doi.org/10.1007/BF02951333.

[27] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence. 1997;97(1):245–271. Relevance. https://doi.org/https://doi.org/10.1016/S0004-3702(97)00063-5.

[28] Biau G, Scornet E.: A Random Forest Guided Tour. Available from: https://arxiv.org/abs/1511.05741.

[29] Tsanas A, Little MA, McSharry PE.: A Simple Filter Benchmark for Feature Selection. Available from: https://api.semanticscholar.org/CorpusID:55130315.

[30] Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 2020 Mar;143:106839. https://doi.org/10.1016/j.csda.2019.106839.

[31] Brown G, Pocock AC, Zhao MJ, Luján M.: Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. Available from: https://api.semanticscholar.org/CorpusID:6621217.

[32] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Computational Systems Bioinformatics CSB2003 Proceedings of the 2003 IEEE Bioinformatics Conference CSB2003. 2003;p. 523–528. https://doi.org/10.1109/CSB.2003.1227396.

[33] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005;27(8):1226–1238. https://doi.org/10.1109/TPAMI.2005.159.

[34] Zhao Z, Anand R, Wang M. Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2019;p. 442–452. https://doi.org/10.1109/DSAA.2019.00059.

[35] Kira K, Rendell LA. The Feature Selection Problem: Traditional Methods and a New Algorithm. In: AAAI Conference on Artificial Intelligence; 1992. Available from: https://api.semanticscholar.org/CorpusID:46457448.

[36] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. Machine Learning: ECML-94. 1994;p. 171–182. https://doi.org/10.1007/3-540-57868-4_57.

[37] Robnik-Šikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning. 2003 Oct;53(1):23–69. https://doi.org/10.1023/A:1025667309714.

[38] McCall J. Genetic algorithms for modelling and optimisation. Journal of Computational and Applied Mathematics. 2005 Dec;184(1):205–222. https://doi.org/10.1016/j.cam.2004.07.034.

[39] Baluja S. Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning; 1994. Available from: https://api.semanticscholar.org/CorpusID:14799233.

[40] Chakravarty K, Das D, Sinha A, Konar A. Feature selection by Differential Evolution algorithm - A case study in personnel identification; 2013. p. 892–899. Available from: https://api.semanticscholar.org/CorpusID:19608187.

[41] Marandi A, Afshinmanesh F, Shahabadi M, Bahrami F. Boolean Particle Swarm Optimization and Its Application to the Design of a Dual-Band Dual-Polarized Planar Antenna; 2006. p. 3212 – 3218. Available from: https://doi.org/10.1109/CEC.2006.1688716.

[42] Zhang H, Sun G. Feature selection using tabu search method. Pattern Recognition. 2002;35(3):701–711. https://doi.org/https://doi.org/10.1016/S0031-3203(01)00046-2.

[43] Mafarja MM, Mirjalili S. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. Neurocomputing. 2017;260:302–312. https://doi.org/https://doi.org/10.1016/j.neucom.2017.04.053.

[44] Storn R, Price K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. Journal of Global Optimization. 1997 Dec;11(4):341–359. https://doi.org/10.1023/A:1008202821328.

[45] Prusty S, Patnaik S, Dash SK. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology. 2022;4. https://doi.org/10.3389/fnano.2022.972421.

[46] Szeghalmy S, Fazekas A. A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. Sensors 2023. 2023;23(4). https://doi.org/10.3390/s23042333.

[47] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.: Scikit-learn: Machine Learning in Python. arXiv. Available from: http://arxiv.org/abs/1201.0490.

[48] Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30 (NIP 2017); 2017. Available from: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/.

[49] Mazzanti S.: mrmr. GitHub. Available from: https://github.com/smazzanti/mrmr.

[50] Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH.: Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining. arXiv. Available from: https://arxiv.org/abs/1711.08477.

[51] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457–481. https://doi.org/https://doi.org/10.1007/978-1-4612-4380-9_25.

[52] Rich J, Neely J, Paniello R, Voelker C, Nussenbaum B, Wang E. A practical guide to understanding Kaplan-Meir curves. Otolaryngology–head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery. 2010 09;143:331–6. https://doi.org/10.1016/j.otohns.2010.05.007.

[53] Lundberg SM, Erion GG, Lee SI.: Consistent Individualized Feature Attribution for Tree Ensembles. arXiv. Available from: https://arxiv.org/abs/1802.03888.

[54] Daghlas S, Govindarajan R. Relative effects of forced vital capacity and ALSFRS-R on survival in ALS. Muscle & Nerve. 2021 06;64. https://doi.org/10.1002/mus.27344.

[55] Kjældgaard AL, Pilely K, Olsen KS, Jessen AH, Lauritsen AØ, Pedersen SW, et al. Prediction of survival in amyotrophic lateral sclerosis: a nationwide, Danish cohort study. BMC Neurology. 2021 Apr;21(1):164. https://doi.org/10.1186/s12883-021-02187-8.

[56] Taylor AA, Fournier C, Polak M, Wang L, Zach N, Keymer M, et al. Predicting disease progression in amyotrophic lateral sclerosis. Annals of Clinical and Translational Neurology. 2016 Sep;3(11):866–875. https://doi.org/10.1002/acn3.348.

[57] Zhou N, Manser P. Does including machine learning predictions in ALS clinical trial analysis improve statistical power? Annals of Clinical and Translational Neurology. 2020 Aug;7(10):1756–1765. https://doi.org/10.1002/acn3.51140.

[58] Byrne S, Elamin M, Bede P, Shatunov A, Walsh C, a B, et al. Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: A population-based cohort study. The Lancet Neurology. 2012 03;https://doi.org/10.1016/S1474-4422(12)70014-5.

[59] Witzel S, Frauhammer F, Steinacker P, Devos D, Pradat PF, Meininger V, et al. Neurofilament light and heterogeneity of disease progression in amyotrophic lateral sclerosis: development and validation of a prediction model to improve interventional trials. Translational Neurodegeneration. 2021 Aug;10(1):31. https://doi.org/10.1186/s40035-021-00257-y.

[60] Bede P, Iyer PM, Finegan E, Omer T, Hardiman O. Virtual brain biopsies in amyotrophic lateral sclerosis: Diagnostic classification based on in vivo pathological patterns. NeuroImage: Clinical. 2017;15:653–658. https://doi.org/https://doi.org/10.1016/j.nicl.2017.06.010.

[61] Anani T, Delbot F, Pradat-Peyre JF. Experimental Comparison of Metaheuristics for Feature Selection in Machine Learning in the Medical Context. In: Maglogiannis I, Iliadis L, Macintyre J, Cortez P, editors. Artificial Intelligence Applications and Innovations. Cham: Springer International Publishing; 2022. p. 194–205. Available from: https://doi.org/10.1007/978-3-031-08337-2_17.

[62] Anani T, Delbot F, Pradat-Peyre JF. An optimised version of differential evolution heuristic for feature selection. In: Dorronsoro B, Yalaoui F, Talbi EG, Danoy G, editors. Metaheuristics and Nature Inspired Computing. Marrakech: Springer International Publishing; 2024. Available from: https://link.springer.com/book/9783031692567.

[63] Anani T.: ALSML. GitHub. Available from: https://github.com/thibaultanani/ALSML.

[64] Anani T.: Tournament in Differential Evolution. GitHub. Available from: https://github.com/thibaultanani/TiDE.

[65] Anani T, Delbot F, Pradat-Peyre JF, Pradat PF.: ALSML. Heroku. Available from: https://alsml-78a86daadd83.herokuapp.com/.

[66] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Medicine. 2015 Jan;13(1):1. https://doi.org/10.1186/s12916-014-0241-z.

Appendix A Results with ALSFRS-R

A.1 Model building process

The model building method used is similar in all respects to that presented in section 2.6. The dataset consists of a sub-sample of 1598 patients with an ALSFRS-R score from PRO-ACT. We then divided it into two distinct subsets: 75% of the patients (1198 patients, 1024 of whom survived to the first year of follow-up) were placed in a training set, and the remaining 25% (400 patients, 342 of whom survived to one year of follow-up) in an independent test set. The total number of features was 48 (the 43 original features, the ALSFRSR score, Q10a, Q10b, Q10c and a respiratory score which is the sum of the last 3).

A.2 Results

In a similar way to the ALSFRS score study, DE remains the method that obtains the best results on the vast majority of evaluation metrics. The results on the validation data seem at first sight to be better, with a Balanced Accuracy score of 78.57% (Table A1) compared with 76.05% (Table 4), an increase of 2.52% percentage points. However, the model selected is composed of more features than in the previous experiment (29 compared with 19), making it more complex. The main features selected are broadly similar to those of the first experiment on the unrevised score (see Figure 2). The selected features include FVC, age, gender, duration since symptom onset, weight and height (BMI), ALSFRS score, pulse rate and blood pressure. Although the features specific to the revised version of the scale (Q10a, Q10b, Q10c) were selected, they did not appear to have a significant impact on the model's decision. We also obtained a Balanced Accuracy score on the test data of 71.59% (Sensitivity = 72.41%, Specificity

Table A1: Performance of the best model for predicting 1-year survival for each of the 6 feature selection methods in cross-validation; SCC, ANOVA, MI, MRMR, ReliefF, and DE. Abbreviations: LM = Learning Method, TN = True Negatives, FP = False Positives, FN = False Negatives, TP = True Positives, Sens. = Sensitivity, Spec. = Specificity, Prec. = Precision, Balanced = Balanced Accuracy and k = number of features selected. For the Balanced Accuracy score, the minimum and maximum scores obtained during cross-validation are indicated in brackets.

Methods	**LM**	**TN**	**FP**	**FN**	**TP**	**Sens.**	**Spec.**	**Prec.**	**Balanced**	**k**
w/selection	LR	130	44	257	767	74.93	74.71	94.57	74.76 (63.85:84.64)	48
SCC	LR	131	43	258	766	74.81	75.30	94.69	74.85 (63.85:84.64)	46
Anova	RR	138	36	280	744	72.65	79.31	95.38	75.98 (70.22:83.17)	41
MI	RR	137	37	279	745	72.89	78.73	95.26	75.69 (67.82:87.42)	39
MRMR	LR	130	44	257	767	74.93	74.71	94.57	74.76 (63.85:84.64)	48
ReliefF	RR	138	36	282	742	72.45	79.31	95.36	75.89 (67.48:84.15)	43
DE	RR	**146**	**28**	**274**	**750**	**73.24**	**83.91**	**96.39**	**78.57 (69.76:86.93)**	29

Table A2: Performance of the model to predict the evolution of the ALSFRS-R by keeping the selected features present in the section A1 on the validation and test data. The best model was obtained with Ridge regression.

Dataset	**Scores**	**T_3**	**T_6**	**T_9**	**T_{12}**
Validation	RMSE	3.242 (2.68:3.50)	4.640 (4.35:4.91)	5.663 (5.43:5.98)	6.820 (6.19:7.16)
	R^2	0.751	0.619	0.552	0.456
	R^2_{ajd}	0.743	0.608	0.539	0.440
	PCC	0.866	0.787	0.743	0.676
Test	RMSE	2.866	3.849	5.134	6.570
	R^2	0.814	0.725	0.608	0.499
	R^2_{ajd}	0.783	0.680	0.544	0.417
	PCC	0.902	0.853	0.784	0.708

= 70.76%, Accuracy = 93.80%), which is far from the results obtained on the validation data and could indicate a low generalisation capacity of the model. On the other hand, the scores obtained (indicated in Table A2 and Figure A2) on the prediction of the evolution of the disease are lower than or equivalent in the best case to the scores obtained in the previous experiment (see Table 7 and Figure 4), both on the validation data (RMSE of 6.820 vs 5.835) and test data (RMSE of 6.570 vs to 5.516). The addition of more features and poorer scores could be explained by the small amount of data available (1589 patients, meaning only 31.69% of the original data). Overall, the inclusion of the ALSFRS-R score in addition to the ALSFRS does not appear to make a significant positive difference in this study.

Appendix B Differential Evolution

Each chromosome of an individual represented a feature and whether or not it would be taken into account during the learning process. An individual was represented by a Boolean vector (i.e. made up of 0s and 1s). So, depending on the value of a chromosome, it was possible to know whether a feature was taken into account when the value was 1 or ignored in the opposite case. A score was then assigned to an individual by training with the feature subset that defined it. The individual with the highest Balanced Accuracy score was then considered to be the best performing individual, as shown in Figure B3.

At the start of each generation, new individuals called mutants were generated from individuals in the population. As many mutants as individuals were created using a mutation strategy. To form each mutant, 3 individuals $(r1, r2, r3)$ different from each other, but also different from the individual at position i, in the population were chosen at random. Each new mutant was calculated using the following formula:

$$Mut_i = Ind_{r1} + F \times (Ind_{r2} - Ind_{r3}) \tag{B1}$$

With F a parameter to be selected upstream of the algorithm between 0 and 2 which controls the amplification of the differential variation $(Ind_{r2} - Ind_{r3})$. If the formula B1 gave a number that was neither 0 nor 1, it was rounded to the nearest integer value. Figure B4 is an example illustrating the generation of a mutant. Mut_1 is the first mutant generated in the population, so $i = 1$ and Ind_1 cannot be among the random individuals selected.

The next step was to cross the individuals. Each mutant was associated with an individual and then a cross was made between them to form a new individual called a child. For each chromosome, a random draw determined whether the child's chromosome would be that of the individual or the mutant, depending on the crossover rate CR, a parameter in the algorithm. To avoid a child being strictly similar to the base individual, a chromosome was selected at random (chr_{rand}) which would inevitably inherit the mutant. An example of crossover is illustrated in Figure B5.

Each child was then evaluated and matched with an individual. For each pair (individual, child), only the one with the best score was retained for the next generation. In Figure B6, the child generated by the crossover operation has a better score than the individual. It therefore replaces it for the next generation. These last three stages (mutation, crossover, selection) were repeated for each generation until a stopping criterion was reached. There are several stopping criteria, such as reaching a certain maximum number of generations. Here, the choice was made in terms of execution time and diversity. When the diversity was too low or the algorithm reached the maximum time limit of a single hour, it stopped.

Fig. A1: The features with the greatest impact on the model according to the test data. The present data was calculated using the Shap (SHapley Additive exPlanation) value proposed by Lundberg et al. [53]. The x-axis represents the impact of the feature on the model output, while the y-axis represents the names of the most important features for the model. Each point represents a Shap value and the thickness represents the density. The colour represents whether a value is high (red) or low (blue) depending on the value interval of a feature e.g. FVC is the feature that has the most impact in calculating the model output.

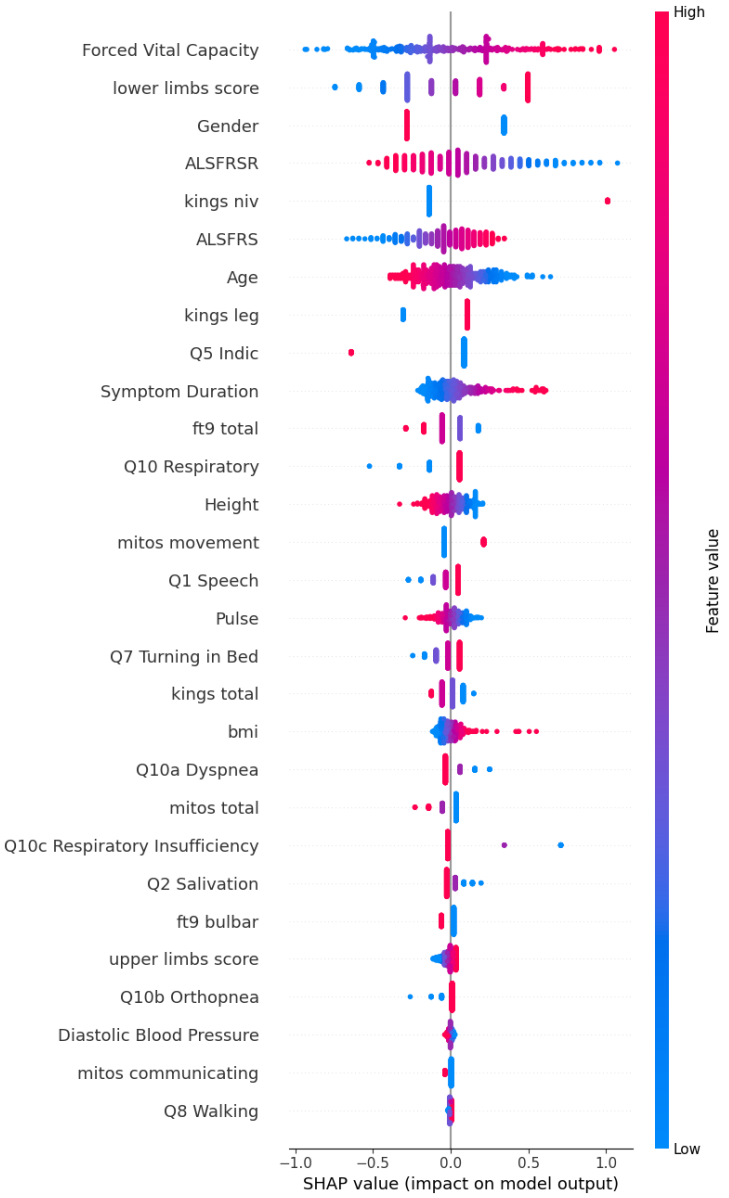


Fig. A2: The calibration curve of the model on the test data for each period. The blue points are the ALSFRS-R scores of the patients placed according to their actual value and that predicted by the model. The red line represents the identity function (Prediction=Reality).

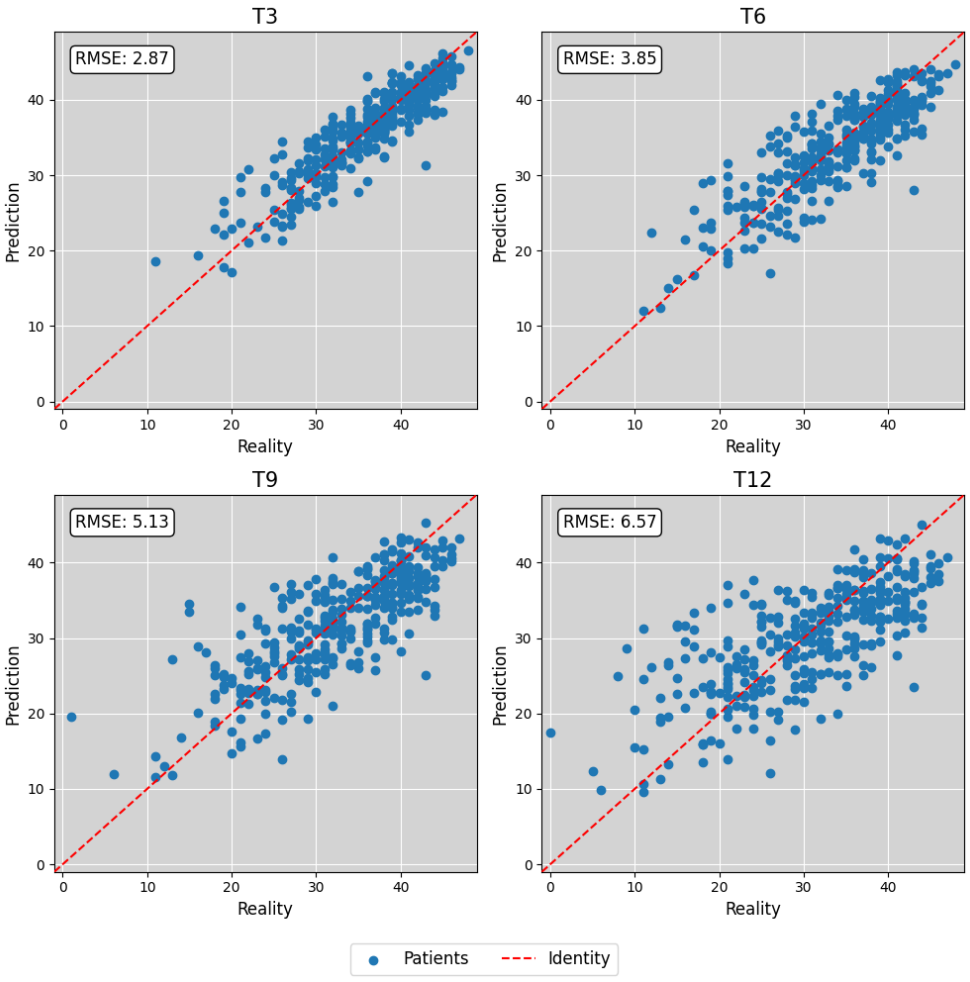


Fig. B3: Example of a population P of 6 individuals Ind for an arbitrary dataset with 5 explanatory features. The individual Ind_5 is composed of 4 explanatory features $\{chr_0, chr_1, chr_3, chr_4\}$ and the one with the highest score in P.

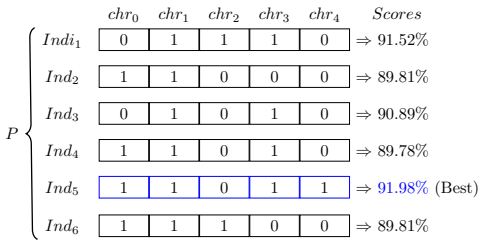


Fig. B4: Example of an individual obtained after mutation with $F = 1$.

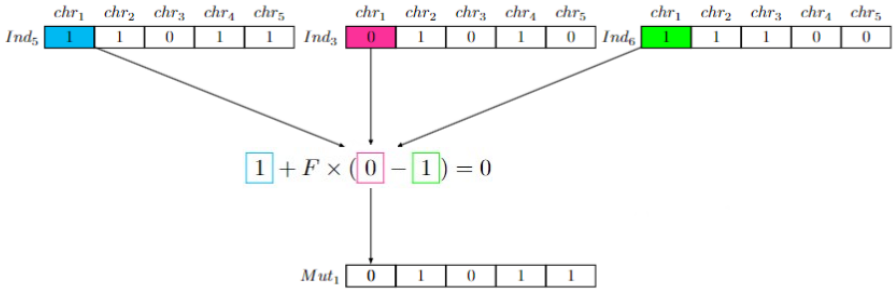


Fig. B5: Example of an individual obtained after crossover with $CR = 0.5$.

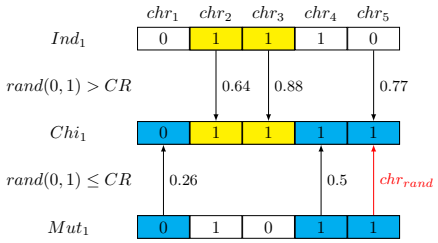


Fig. B6: Example of how the selection stage operates.