# Predicting ALS survival and functional decline via metaheuristic feature optimisation

Thibault Anani [1] JF. Pradat-Peyre [1,2] F. Delbot [1,2] C. Desnuelle [3] AS. Rolland [4] D. Devos [4] PF. Pradat [5,6]

[1]LIP6, CNRS, Sorbonne Université, Paris (France)
[2]UFR SEGMI, Université Paris Nanterre (France)
[3]Université Côte d'Azur, Nice (France)
[4]Université de Lille, INSERM, CHU Lille, ACT4ALS (France)
[5]Hôpital de la Pitié–Salpêtrière — AP-HP, Paris (France)
[6]LIB, Sorbonne Université, Paris (France)

## Background

**Amyotrophic Lateral Sclerosis (ALS)**, a progressive **neurodegenerative disease with no curative treatment** and affecting motor neurons leads to motor weakness, atrophy, spasticity and difficulties with speech, swallowing and breathing.

Due to the **heterogeneity of the disease**, clinicians find it especially **challenging to predict the rapidity of disease progression** in patients.

**Machine Learning** methods based on **large patients datasets** have **successfully identified underlying correlations** in patient data.

However, due to **too many irrelevant features**, learning models achieve **limited quality and interpretability** in predicting disease progression.

## Objectives

1. Conceptualise a machine learning model to **predict patient survival over 1 year and functional decline** based on their characteristics.
2. **Optimise the quality of the model** in terms of performance and explainability **by selecting the most relevant subset of features**.
3. **Identify groups of patients with similar characteristics** and survival rates based on the results of the models.

## Feature Selection

Feature selection consists in **choosing a part of the features considered as the most relevant** in the construction of the model to **improve the performance** of the learning algorithm used. Several potential benefits :

- **Performance** : Improved predictive quality.
- **Interpretability** : More parsimonious and readable models.
- **Efficiency** : Reduced prediction and maintenance time.
- **Robustness** : Reduced risk of overfitting.

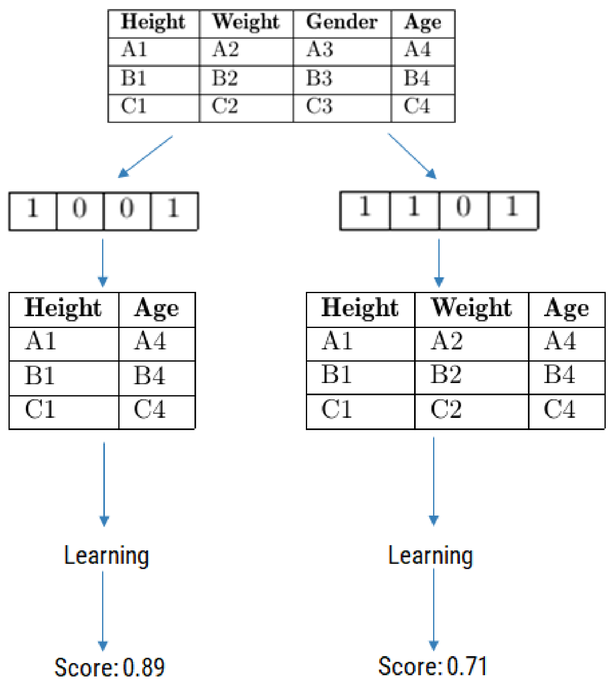


Figure 1. Illustrative example of feature subsets. The first model is better than the second.

**Testing all combinations of features** is time-consuming and **leads to a combinatorial explosion** due to the excessive number of features.

## Metaheuristic

Metaheuristics are **computational methods for solving complex optimisation problems** and finding a **solution close to the optimal one**.

A **metaheuristic optimised for this problem was developed** to find the best feature subset [1].

## Data Description

### Training and validation data

The **training data consist of 3782 patients** from two different clinical trials : Pro-Act and Exonhit therapeutics (1-year Survival Rate : 79%) [2].

### Testing data

Two cohorts are used to test the model :

1. **1261 patients** from Pro-Act and Exonhit (1-year Survival Rate : 79%).
2. **158 patients** from, another independent cohort, PULSE (1-year Survival Rate : 74%).

### Predictors and features

Over **43 explanatory features** in total :

- Age
- Gender
- Weight
- Size
- Symptom onset
- Symptom duration
- Forced vital capacity
- Pulse
- Diastolic blood pressure
- Systolic blood pressure
- **ALSFRS**
- etc.

## Results : Model Performance

### Predictive Quality

| Data | Methods | Balanced Accuracy | Number of Features |
|---|---|---|---|
| Validation | No selection | 74.12 | 43 |
| | Metaheuristic | **76.05** | **19** |
| Test | No selection | 75.25 | 43 |
| | Metaheuristic | **76.23** | **19** |
| Test PULSE | No selection | 75.61 | 43 |
| | Metaheuristic | **76.33** | **19** |

Table 1. Performance of the prediction model with and without feature selection.

Feature selection significantly reduces the number of features required to design the model while increasing its predictive efficiency.

### Model Interpretability : Feature importance

How can one determine the importance of a feature ?

- A **logistic regression model** consists of **constructing a linear function** where each feature is associated with a coefficient $\beta$ [3].
- Learning consists of **identifying the coefficients $\beta$ that minimise the error**.
- The **higher the coefficient**, the **greater the impact the feature will have** on the prediction.

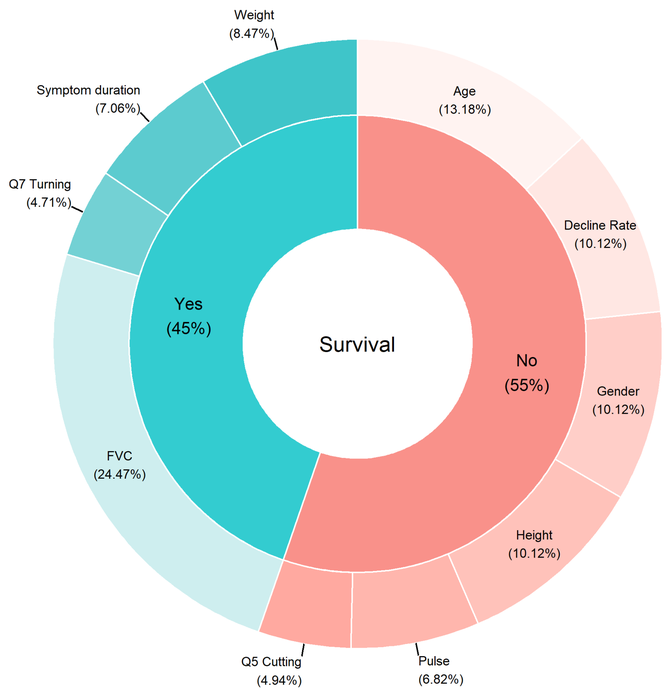


Figure 2. Relative importance of the 10 most important features selected based on their positive/negative effect on survival.

## Results : Patients Profile

**Clustering of patients** according to the **survival rates** estimated by the model :

| Predicted survival (%) | Number of patients |
|---|---|
| Less than 20 | 162 |
| Between 20 and 40 | 244 |
| Between 40 and 60 | 294 |
| Between 60 and 80 | 386 |
| More than 80 | 373 |

- **Clean separation of curves** : Proper discrimination (C-index $= 84\%$).
- **Overall performance is satisfactory** : (Accuracy $= 76\%$, AUC $= 84\%$).
- The model **effectively identifies patients based on their risk level.**

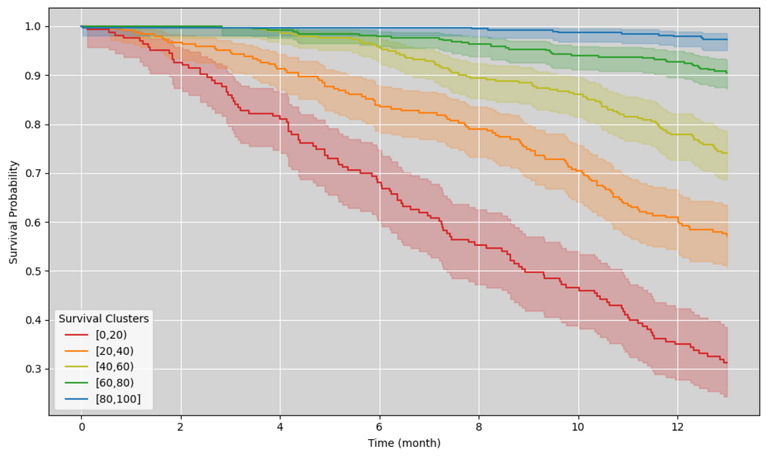


Figure 3. Survival curves based on test data. The coloured areas represent confidence intervals. Example: Red curve→ Low survival and FVC; high decline, age and pulse rate.

## Discussion

Our study highlights the value of ML in predicting survival and disease progression in ALS. **Feature selection, often overlooked in the literature, significantly improves model quality** (Full study : [4]).

Notably, since patients exhibit varying stages of disease progression at baseline, our **models are specifically calibrated to capture progression relative to everyone's initial state**.

While feature selection using this approach slows model construction, **applying the model to new patients requires minimal computation time**, even on low-capacity devices.

This **approach's flexibility extends beyond ALS**, applying to any disease or prediction problem with structured data. Its generalisability makes it a reproducible, adaptable tool for broad medical use.

## Try out the models in real time

Figure 4. QR Code that redirects to our web application (ALSML) designed to assist clinicians.

## References

[1] Thibault Anani et al. An optimised version of differential evolution heuristic for feature selection. In *Communications in Computer and Information Science*, Communications in computer and information science, pages 197–213. Springer Nature Switzerland, Cham, 2024.

[2] Nazem Atassi et al. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, November 2014.

[3] Fabian Pedregosa et al. Scikit-learn: Machine learning in python, 2012.

[4] Thibault Anani et al. Feature selection using metaheuristics to predict annual amyotrophic lateral sclerosis progression. *Amyotroph. Lateral Scler. Frontotemporal Degener.*, pages 1–16, July 2025.