Predicting the Functional Rating Scale and Self-Assessment Status of ALS Patients with Sensor Data

Notebook for the iDPP@CLEF Lab at CLEF 2024

Andreia S. Martins[†], Daniela M. Amaral[†], Eduardo N. Castanho[†], Diogo F. Soares, Ruben Branco[*], Sara C. Madeira and Helena Aidos

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease causing progressive loss of cognitive and motor functions. Due to limited understanding of its mechanisms, there is no cure. Prognosis is still crucial for the effective planning of symptom treatment, however, the heterogeneity in patient progression drives the need for precision medicine research. iDPP CLEF 2024 aims to develop novel methodologies for predicting ALS disease progression, enabling the community to combine efforts and improve current prognostic methods. This report discusses our participation in tasks 1 and 2, evaluating the impact of sensor data on improving the prediction of ALSFRS-R scores. The proposed methodology combines temporal summarization techniques (extracting relevant statistics from the sensors), feature selection and extraction methods, and state-of-the-art classifiers for each ALSFRS-R question independently. Results show that random forest models yield the best overall performance, and selecting the k-best features and biclustering were the best overall feature selection and extraction strategies for tasks 1 and 2, respectively.

Keywords

Amyotrophic Lateral Sclerosis, Prognostic Prediction, Time Series Data, Biclustering, Multi-Class Classification

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a devastating neurodegenerative disease characterized by the progressive degeneration of motor neurons, leading to muscle weakness, atrophy, and eventual paralysis [1]. The progression of ALS varies significantly among patients, with some experiencing rapid deterioration while others decline more slowly [2]. This variability complicates the ability to predict disease trajectory, making it challenging for clinicians to offer accurate prognoses and for patients to make informed decisions about their future care [3].

Traditionally, clinical assessments of ALS progression rely on periodic evaluations using scales like the ALS Functional Rating Scale-Revised (ALSFRS-R) [4]. Although essential, these assessments provide only snapshots of a patient's condition at discrete time points and can miss subtle but critical changes between visits. This intermittent data collection limits the ability to detect early signs of disease worsening and delays the implementation of necessary interventions.

Recent advancements in sensor technology present a promising solution to these limitations. Sensors can generate a rich, real-time dataset by continuously monitoring physiological parameters such as muscle activity, respiratory function, and movement patterns [5]. This continuous data capture offers a detailed and dynamic view of a patient's condition, potentially revealing early indicators of disease progression that would otherwise go unnoticed between clinical visits [6].

However, to fully understand and predict ALS progression, it is essential to complement sensor data with patients' self-assessment data [7]. Self-assessments provide critical insights into subjective symptoms such as pain, fatigue, and emotional well-being, which are not easily quantifiable through sensors

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*]Corresponding author.

[†]These authors contributed equally.

asmartins@ciencias.ulisboa.pt (A. S. Martins); daniela.amaral@tecnico.ulisboa.pt (D. M. Amaral); ejcastanho@ciencias.ulisboa.pt (E. N. Castanho); dfsoares@ciencias.ulisboa.pt (D. F. Soares); rmbranco@ciencias.ulisboa.pt (R. Branco); sacmadeira@ciencias.ulisboa.pt (S. C. Madeira); haidos@ciencias.ulisboa.pt (H. Aidos)

alone. Integrating objective sensor data with subjective self-assessment data creates a comprehensive, multidimensional dataset encompassing measurable physical changes and the patient's lived disease experience [8].

In this context, within the iDPP CLEF 2024 challenge[1] framework, we tackled Tasks 1 and 2, which target predicting the twelve scores of the ALSFRS-R from sensor data. Task 1 aims to predict the score assigned by the clinician at the second visit, while Task 2 targets the second patient's self-assessment score. This paper reports the work done to overcome this challenge. We approach this challenge as a multi-label, multi-class classification approach with high-dimensional data. To handle the longitudinal datasets, we consider a double-step approach that transforms the time series sensor data using statistics computed from a time window period. Additionally, we test two feature selection strategies (K-Best features in all sensors and K-Best features in each sensor) and one feature extraction strategy (Biclustering-based features). To classify the ALSFRS-R scores, we train several state-of-the-art classifiers for each question independently.

2. Related Work

Sensor technology has gained significant traction in recent years for monitoring ALS patients. Wearable sensors, such as accelerometers and gyroscopes, have continuously monitored motor function, gait, and other physical activities [5, 9, 10, 11]. Accelerometer studies demonstrated their effectiveness in capturing detailed movement patterns, providing valuable data for assessing motor decline in ALS patients [6, 10, 9]. Vieira et al. [12] developed a model targeting ALS progression prediction based on voice samples and accelerometer measurements from a four-year longitudinal dataset. This model was used to predict bulbar-related and limb-related ALSFRS-R scores. Straczkiewicz et al. [11] used wrist wearables and ALSFRS-R self-entries data to propose new measures to quantify the count and duration of upper limb movements.

In addition to sensor data, integrating patients' self-assessment data has proven beneficial in understanding ALS progression. Studies have shown that self-reported pain, fatigue, and quality of life measures can provide critical insights that complement objective sensor data [13, 7].

Machine learning techniques have been increasingly applied to predict disease progression in ALS [14]. Predicting the progression of the functional domains (twelve questions) assessed by the well-known functional scale, the ALSFRS-R was also investigated by Gordon and Lerner [15]. They modeled a multiclass classifier using demographic, respiratory assessments, genetic data, and other dynamic data to predict the values of each ALSFRS-R question at the time of the last patient visit.

Subspace techniques, such as pattern mining, biclustering, and triclustering, discover local patterns with non-constant coherencies with potential for predictive tasks. Martins et al. [16] recently proposed combining itemset mining with sequential pattern mining to uncover disease presentation and progression patterns in ALS patients and utilize these patterns to forecast the need for NIV. In a similar approach with the same prognostic target, Matos et al. [17] suggested a classifier based on biclustering. Biclustering [18, 19] was used to locate groups of patients with similar values in subsets of clinical features (biclusters), which were then combined with static data as features. Although promising, none of these methods considered the temporal relationship of features. Soares et al. [20] proposed BicTric, a classifier capable of learning predictive models from both static and temporal data using discriminative patterns obtained through biclustering and triclustering [21, 22, 23]. Recently, Soares et al. [24] enhanced BicTric with TCtriCluster, a triclustering algorithm incorporating temporal contiguity constraints. These approaches utilized temporal preprocessing with snapshots and the time windows method proposed by Carreiro et al. [25] to learn predictive models for various clinically relevant ALS endpoints.

Integrating multi-modal data sources, including sensor data, self-assessments, and traditional clinical metrics, has shown potential in providing a more comprehensive understanding of ALS progression. Johnson et al. [8] conducted a study combining wearable sensor data with patient-reported outcomes

[1] http://brainteaser.dei.unipd.it/challenges/idpp2024/

and clinical assessments, demonstrating that multi-modal data fusion could enhance predictive accuracy and offer deeper insights into disease dynamics.

3. Methodology

The objective of Tasks 1 and 2 of the iDPP CLEF 2024 challenge is to predict the values of the ALSFRS-R sub-scores of a second evaluation, given the values of the first evaluation. This would imply a reduced set of training instances (52 patients, in total), so we decided to generalize the challenge to predict the ALSFRS-R sub-scores of any evaluation given a previous evaluation, resulting in 121 training instances for Task 1 and 220 instances for Task 2.

The dataset made available [26, 27] with this challenge contains information on ALS patients comprising the following data: static (including demographic and clinical information), all the ALSFRS-R evaluations (comprising the scores of the 12 questions for each patient), and sensor data (collected from the sensors of a fitness smartwatch). Figure 1 illustrates the processing of the dataset.

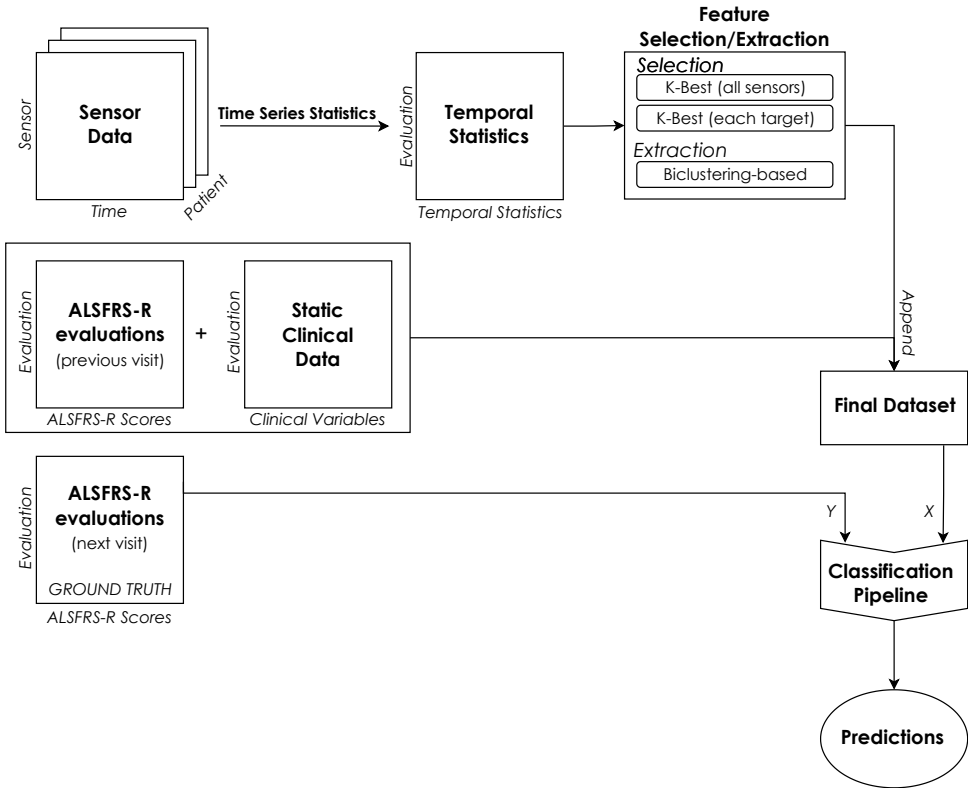

Figure 1: Data processing pipeline. Addressing the challenge implies handling data from three sources: static clinical variables, ALSFRS-R scores, and sensor data. To handle the highly dimensional sensor time series data, we computed statistics for each sensor and then applied feature selection or extraction strategies to reduce the dimensionality of the sensor dataset. The final dataset (that feeds the classifiers) aggregates these data sources.

Tasks 1 and 2 face a significant hurdle due to the sensor dataset's high dimensionality, stemming from a large number of sensor features (90 in total) and the numerous time points (approximately 268 sensor records per patient). To address this issue, we used a two-step processing of the dataset: first, we *extracted temporal statistics* from the longitudinal datasets. Second, we used *feature selection or extraction techniques* to obtain a representation of the sensor statistics with smaller dimensionality.

3.1. Time Series Statistics

We derived new features from the longitudinal sensor data for each evaluation using summarization techniques, consisting of statistical metrics such as mean, standard deviation, minimum and maximum

Table 1
Number of excluded sensor features, by category. The sensor data can be grouped into 6 distinct categories (Category). For each original sensor feature within these categories, 6 statistical metrics - mean, standard deviation, minimum value, maximum value, first value, and last value - were computed (#Computed Features). Features exhibiting zero or near-zero variance (#Low Variance) and those highly correlated with other features within the same category (#High Correlation) were removed from the dataset.

		Task 1		Task 2	
Category	**#Computed Features**	**#Low Variance**	**#High Correlation**	**#Low Variance**	**#High Correlation**
calories	18	0	10	0	9
steps	24	0	3	0	3
beat_to_beat	240	13	116	10	108
heart_rate	60	10	1	9	2
respiration	108	0	27	0	31
SpO_2	90	0	20	0	16

values, and the first and last values of each feature (as in Branco et al. [28]). To avoid the bias introduced by considering the entire sensor data history, these metrics were computed within fixed time intervals, specifically considering the interval $[t - \delta, t]$, where t represents the day of the target appointment and δ is the number of days within the interval (set to 15 days for Task 1 and 7 days for Task 2). This computation resulted in 540 new sensor features (90 original sensor features $\times$ 6 statistical metrics).

Another issue encountered with the dataset was missing values, even after the aforementioned computations. To address this, various interpolation and imputation techniques were explored, with polynomial interpolation of degree 5 proving to be the most effective in minimizing variance decrease across the feature sets.

After the interpolation step, sensor features exhibiting zero or near-zero variance (less than 10^{-5}) were deemed uninformative and consequently removed. Furthermore, highly correlated sensor features within the same category (calories, steps, beat_to_beat, heart_rate, respiration, and SpO_2) were also eliminated to mitigate redundancy. The selection of features for removal was based on Pearson correlation, with a correlation threshold set at 0.95 (see Table 1).

3.2. Feature Selection and Extraction Techniques

The sensor statistics obtained from the previously discussed step are still high dimensional, as there are 340 features for Task 1 and 352 features for Task 2. Subsequently, we applied three techniques (two feature selection (i) and (ii), and one feature extraction (iii)) to reduce the dataset dimensionality:

(i) K-Best features in **all sensors**;
(ii) K-Best features in **each target**;
(iii) **Biclustering**-based features.

The first two feature selection techniques are based on a k-best selection strategy. First, we selected the top 5 features for predicting each target question based on ANOVA F-value between labels and features. Predictions were then made using the set of highest-ranked sensor statistical features across all questions (All Sensors). Alternatively, a specialized prediction approach was also adopted wherein the top 5 features were selected independently for each ALSFRS-R question based on mutual information (Each Target) (see Table 2). These selections were made using the `SelectKBest` class of the `sklearn.feature_selection` Python module.

As an alternative to these aforementioned feature selection strategies, we used a feature extraction strategy based on biclustering to reduce the dataset dimensionality. Biclustering, the simultaneous clustering of rows and columns of a data matrix, has shown its ability to discover local patterns with non-constant coherencies in both descriptive and predictive learning tasks [21, 18]. Our approach, illustrated in Figure 2, applies biclustering to the *Patient*$\times$*Sensor Feature* training matrix to obtain the

Table 2
Number of selected top-ranked features, by category. Predictions were made using the pairs of strategy-models of highest-ranked computed sensor features, based on the ANOVA F-value, across all questions (All Sensors). Additionally, a specialized prediction method was employed, wherein the top 5 features were independently selected for each ALSFRS-R question based on mutual information (Each Target).

		Task 1						Task 2					
		calories n = 8	steps n = 21	beat_to_beat n = 111	heart_rate n = 49	respiration n = 81	SpO_2 n = 70	calories n = 9	steps n = 21	beat_to_beat n = 122	heart_rate n = 49	respiration n = 77	SpO_2 n = 74
All Sensors		6	6	13	5	3	3	6	5	19	1	5	5
Each Target	**Q1**	3	0	1	1	0	0	1	0	0	0	1	3
	Q2	0	0	0	3	0	2	4	0	1	0	0	0
	Q3	5	0	0	0	0	0	5	0	0	0	0	0
	Q4	0	0	4	0	1	0	0	0	3	0	0	2
	Q5	0	0	5	0	0	0	0	0	5	0	0	0
	Q6	0	3	2	0	0	0	0	0	5	0	0	0
	Q7	0	4	1	0	0	0	0	0	5	0	0	0
	Q8	0	5	0	0	0	0	0	4	1	0	0	0
	Q9	0	5	0	0	0	0	0	4	1	0	0	0
	Q10	0	0	2	1	2	0	1	0	3	0	1	0
	Q11	1	0	3	0	0	1	0	0	4	1	0	0
	Q12	0	0	5	0	0	0	0	0	0	0	5	0

biclusters, with the row pattern of each bicluster being computed as the mean value of each column. Then, the Euclidean distance between each training (and test) sample and the row pattern of each bicluster is computed to obtain a reduced representation of the training (and test) set.

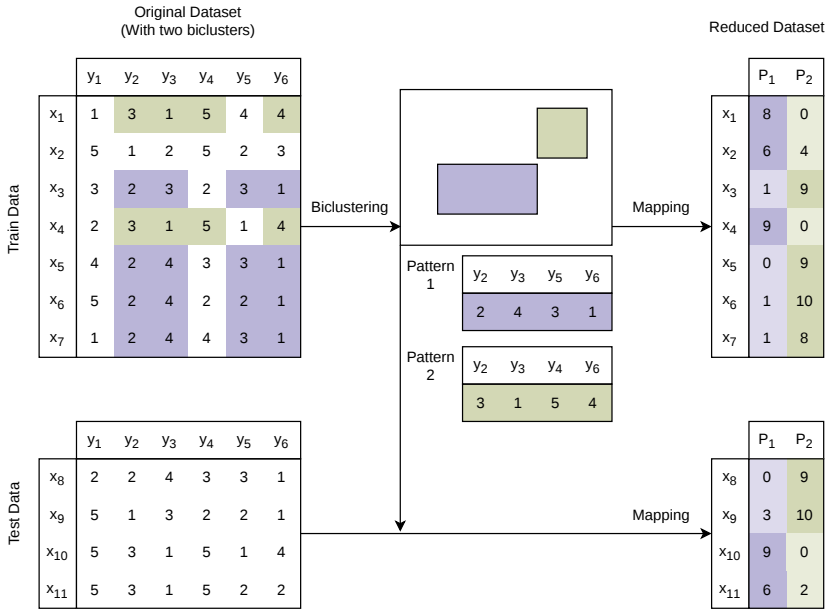

Figure 2: We used an approach based on biclustering-computed features. First, we apply a biclustering algorithm to obtain a set of biclusters (sub-matrices) from the dataset. Second, we compute its *row pattern* for each bicluster. Finally, we compute the distance between each row of the dataset and each bicluster to obtain the new reduced dataset. To simplify the representation of this methodology, we illustrate the pattern of a bicluster by the mode of each column (instead of the mean value) and use the Manhattan distance between each row and bicluster instead of the Euclidean distance.

We considered *Spectral Biclustering* to mine the biclusters as implemented in scikit-learn [29, 30]. The number of biclusters influences the number of features in the reduced dataset. In our approach, we tested values for the number of biclusters and selected the value that maximizes the number of non-trivial biclusters (biclusters with more than 2 rows and columns).

3.3. Modeling and Hyperparameter Optimization

In this section, we discuss our classification methodology, as illustrated in Figure 3.

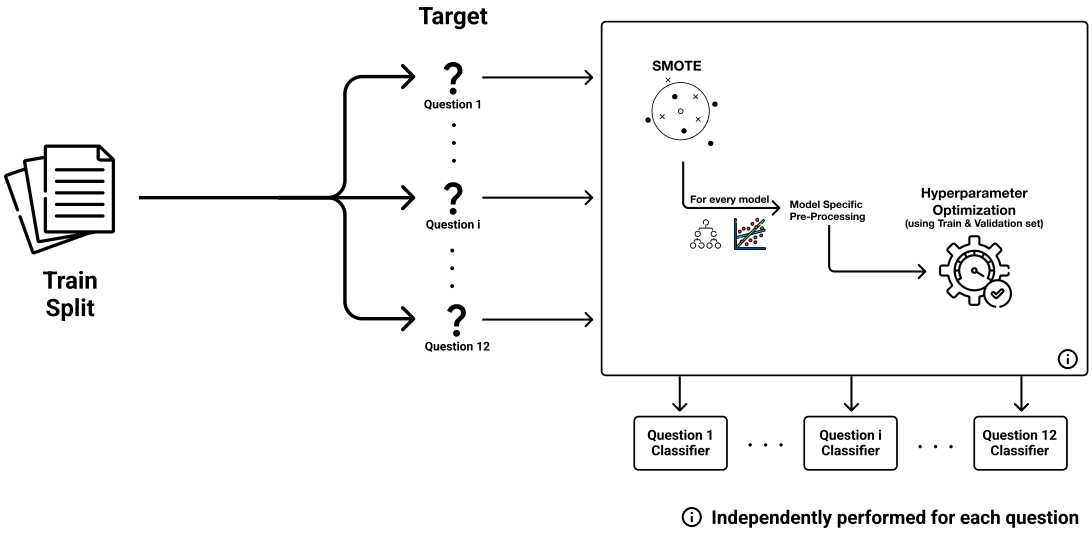

Figure 3: The challenge implies a multi-label, multi-class tasks. To simplify the training, we train classifiers for each question independently. We use SMOTE to compensate for a lack of sufficient representation across each scale value when possible. We train several traditional classifiers for each question, optimized considering the mean absolute error.

The tasks at hand are multi-label, multi-class tasks, which add complexity to the standard modeling techniques. Furthering the difficulty, the labels, which are the ALSFRS-R questions, are not completely independent, as the sub-scores are correlated within the different domains (bulbar, fine motor / upper limb, gross motor /lower limb, and respiratory).

Despite this intricacy, we decided to simplify the task by separating them into independent multi-class problems, where a given patient ALSFRS-R evaluation and their sensor data are used to predict each sub-score individually. Despite not modeling the correlation between questions, we assume that the models could still connect a patient's condition in time with their ability to perform a single function. We train 12 models and combine their predictions to predict the full set of sub-scores.

We consider a set of well-known classifiers covering a diverse range of model types, using scikit-learn [29]: Logistic Regression (LR), Random Forest (RF), XGBoost, and Support Vector Machines (SVC). Each model undergoes a model-appropriate pre-processing if required, and the optimal hyperparameters are searched for, as will be described later on.

For questions that have a sufficient representation across each of the scale values (0 to 4),[2] we employ imblearn [31]'s implementation of SMOTE [32], to alleviate the issue of small training sample size.

It is common to scale the input data for linear models to avoid widely different magnitudes across features that can hurt learning and performance. We use a standard scaler for Logistic Regression and Support Vector Machines to scale the input data.

We optimize the models using the Mean Absolute Error metric, both as a loss function for the model optimization and as a hyperparameter optimization objective, which searches for the best hyperparameter optimization that yields better performance on the validation set. We use Optuna [33] for hyperparameter optimization, with the Tree-Structured Parzen Estimator algorithm (as a sampler), avoiding a grid search brute-force approach to more efficiently sweep the hyperparameter space (see Table 3 for hyperparameter range of each model). The best-performing model is then used for the submissions in the challenge.

To assess the generalization of trained models and to optimize hyperparameters, we split the provided dataset into two sets: a train set and a validation set. As the dataset is multi-label multi-class, regular

[2]Two questions in each task did not qualify, which were questions 11 and 12 for Task 1, and 3 and 11 for Task 2.

Table 3
The hyperparameter space for each model. Int and Float Distributions describe a search space between two integers or floating values, whereas CategoricalDistribution specifies a set of discrete values.

Model	Hyperparameter	Distribution Space
XGBoost Classifier	n_estimators	IntDistribution(100, 1000)
	max_depth	IntDistribution(1, 20)
	learning_rate	FloatDistribution(0.01, 1)
Random Forest	n_estimators	IntDistribution(10, 1000)
	max_depth	IntDistribution(1, 20)
Logistic Regression(max_iter=100000)	C	FloatDistribution(0.01, 10)
SVC(max_iter=100000, cache_size=1000)	C	FloatDistribution(0.01, 10)
	gamma	FloatDistribution(0.01, 10)
	kernel	CategoricalDistribution(["linear", "rbf", "poly", "sigmoid"])

stratified train test splits do not guarantee a representative proportion of each scale value for each question for both splits. We resort to a variant termed iterative stratified train test splitting [34, 35], implemented in the scikit-multilearn package [36]. This method works by iteratively populating both splits and assigning data points at each step to the split that requires them the most to maintain balance. Ultimately, we ensure each split is as similar to the overall dataset as possible. We split the provided training set following a 70/30 ratio, with 70% becoming the training set and 30% the validation set.

All the experiments were run on a Desktop Computer with an AMD Ryzen 9 7950X 16-Core with 64GB of RAM and Ubuntu 22.04.2. The code was run using Python 3.10.11.

4. Results & Discussion

In this section, we cover the results obtained in Tasks 1 and 2 in the challenge, as reported and computed with the private test set made available by the lab organizers.

To examine the impact of our design choices on feature selection or extraction, we define an experimental space beyond the basic analysis of the challenge results. First, for each question, we select the best pair feature selection or extraction strategy and classification model with the top-k (we consider $k = \{1, 2, 3\}$) highest validation metric values for both Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (see section 4.1). Next, to determine which feature selection or extraction performs best, we consider the mean RMSE and MAE across the four classifiers for each question (see section 4.2). Lastly, we will assess whether there is a significant advantage in using one classifier over another. Given that the classifiers are all different types, identifying specific model properties suited for this particular task could lead to improvements for each question (see section 4.3).

4.1. Selecting the best combination of feature strategy and classification model

We conducted experiments to predict the ALSFRS-R questions of a subsequent assessment by combining the best models for each target question based on their validation set performance. Specifically, we submitted the three best-performing pairs for both Tasks (see Table 4).

Table 5 presents the results of the models trained in each feature selection or extraction strategy for predicting each target question, along with the global results (average RMSE and MAE values across all questions). For both Tasks 1 and 2, the best-performing combination of feature selection or extraction strategy and classification model (strategy-model pair) in the test set was the second-best strategy-model pair in predicting the ALSFRS-R questions in the validation set. This suggests that the training and validation sets used for optimizing and validating the classifiers were unsuitable for

Table 4
Results on the validation set for each combination of feature selection or extraction strategy and classification model. RF stands for Random Forest, SVC for Support Vector Machine Classifier, and LR for Logistic Regression.

		Best pair		2nd best pair		3rd best pair	
	Question	**Strategy**	**Model**	**Strategy**	**Model**	**Strategy**	**Model**
Task 1	**Q1**	All Sensors	XGBoost	Each Target	XGBoost	Each Target	RF
	Q2	Each Target	RF	Biclustering	RF	All Sensors	RF
	Q3	Biclustering	RF	Biclustering	XGBoost	All Sensors	RF
	Q4	Each Target	SVC	All Sensors	RF	All Sensors	SVC
	Q5	Each Target	RF	All Sensors	XGBoost	Each Target	SVC
	Q6	Biclustering	XGBoost	Biclustering	RF	All Sensors	XGBoost
	Q7	Each Target	SVC	Each Target	RF	All Sensors	XGBoost
	Q8	All Sensors	SVC	All Sensors	LR	All Sensors	XGBoost
	Q9	Biclustering	XGBoost	Biclustering	RF	All Sensors	RF
	Q10	All Sensors	SVC	Biclustering	SVC	Each Target	LR
	Q11	Biclustering	XGBoost	Each Target	XGBoost	Biclustering	RF
	Q12	All Sensors	RF	Biclustering	RF	All Sensors	LR
Task 2	**Q1**	All Sensors	SVC	Biclustering	RF	Each Target	RF
	Q2	Each Target	XGBoost	Biclustering	SVC	Each Target	LR
	Q3	Biclustering	XGBoost	Biclustering	SVC	All Sensors	XGBoost
	Q4	Each Target	XGBoost	All Sensors	XGBoost	All Sensors	RF
	Q5	All Sensors	RF	Biclustering	RF	Each Target	RF
	Q6	All Sensors	XGBoost	Each Target	RF	All Sensors	RF
	Q7	All Sensors	RF	Biclustering	RF	Each Target	XGBoost
	Q8	Biclustering	RF	Each Target	XGBoost	Each Target	RF
	Q9	Biclustering	XGBoost	Biclustering	RF	Biclustering	SVC
	Q10	All Sensors	SVC	Each Target	RF	Each Target	XGBoost
	Q11	All Sensors	XGBoost	Biclustering	RF	Each Target	RF
	Q12	Biclustering	RF	All Sensors	SVC	Each Target	RF

predicting the ALSFRS-R questions in the second evaluation. These sets included all evaluations made available for the challenge, leading the models to be trained for predicting the next evaluation rather than specifically the second evaluation.

In Task 1, two questions related to the bulbar domain, Q1 and Q2, and one respiratory question, Q11, were the easiest to predict (RMSE 0.309, MAE 0.095). Specifically, Q1 and Q11 were best predicted using the XGBoost classifier with the All Sensors (Best strategy-model pair) and Each Target (2nd best strategy-model) feature selection strategies, respectively. Question Q2 was best predicted using the RF classifier with the All Sensors strategy (3rd best strategy-model pair). In contrast, motor-related questions, Q7 (trunk domain) and Q9 (lower limb domain) had the highest prediction errors (RMSE 0.873, MAE 0.476).

For Task 2, questions Q11 and Q12 were correctly classified for all the evaluations (RMSE 0.000 and MAE 0.000). Both the questions used the RF classifier and the Biclustering strategy (2nd best strategy-model and Best strategy-model pair, respectively). Question Q11 was also correctly classified for all evaluations using the Each Target strategy (3rd best strategy-model pair). Conversely, Q4 had the most misclassified evaluations (RMSE 1.044, MAE 0.545).

Table 5
Results of the submitted strategy-model pairs. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics for the three best strategy-model pairs presented in Table 4. The performance metrics are provided for each target question and averaged across all the 12 questions (Global).

	Model		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Global
Task 1	**Best strategy-model pair**	RMSE	**0.309**	0.577	**0.436**	0.900	**0.655**	0.900	1.113	**0.655**	1.291	0.756	0.378	**0.577**	0.712
		MAE	**0.095**	0.238	**0.190**	0.524	**0.429**	0.619	0.857	**0.429**	0.810	0.381	0.143	**0.238**	0.413
	2nd best strategy-model pair	RMSE	0.787	0.690	0.655	**0.816**	0.756	0.900	0.900	0.873	**0.873**	**0.378**	**0.309**	**0.577**	**0.709**
		MAE	0.429	0.286	0.333	**0.476**	0.476	0.619	0.619	0.572	**0.381**	**0.143**	**0.095**	**0.238**	**0.389**
	3rd best strategy-model pair	RMSE	0.787	**0.309**	0.756	0.976	1.000	**0.690**	**0.873**	0.787	1.069	0.926	0.378	0.845	0.783
		MAE	0.429	**0.095**	0.381	0.571	0.619	**0.476**	**0.476**	0.429	0.667	0.476	0.143	0.429	0.433
Task 2	**Best strategy-model pair**	RMSE	**0.905**	**0.739**	**0.522**	1.206	1.279	1.000	0.674	0.798	**0.302**	1.414	1.206	**0.000**	0.837
		MAE	0.636	**0.364**	**0.273**	0.727	0.909	0.818	0.455	0.636	**0.091**	1.091	0.364	**0.000**	0.530
	2nd strategy-model pair	RMSE	1.000	0.798	0.739	**1.044**	0.953	**0.739**	**0.522**	0.798	**0.302**	**1.000**	**0.000**	0.603	**0.708**
		MAE	0.636	0.455	0.364	**0.545**	**0.545**	**0.545**	**0.273**	**0.636**	**0.091**	**0.636**	**0.000**	0.182	**0.409**
	3rd best strategy-model pair	RMSE	0.953	0.798	**0.522**	**1.044**	**0.853**	0.905	0.905	**0.739**	0.603	1.679	**0.000**	0.302	0.775
		MAE	**0.545**	0.455	**0.273**	**0.545**	**0.545**	0.818	0.636	**0.545**	0.364	1.182	**0.000**	0.909	0.500

4.2. Feature Selection and Extraction Comparison

As previously mentioned, one feature extraction and two feature selection strategies were assessed: biclustering and K-Best selection, both globally for all questions (All Sensors) and individually for each question (Each Target).

Table 6 presents the average model performance in the test set for each ALSFRS-R question and feature selection or extraction method. Overall, no strategy clearly outperformed the others, with the metrics typically not differing much between models with the same target question. However, the preferred strategy does change with the target.

In Task 1, the best overall method was individual k-best selection, Each Target (RMSE 0.780, MAE 0.474). It gathered the best average metrics in 6 out of the 12 questions, followed by the biclustering approach (RMSE 0.815, MAE 0.515) with 4 questions. Notably, there may be a preferred strategy by domain: the All Sensors approach performed best in the trunk domain questions (Q6 and Q7), and Each Target yielded the best metrics in the lower limb domain (Q8 and Q9). However, this behavior does not seem to occur for the upper limb domain (Q4 and Q5). For the bulbar (Q1-Q3) and respiratory (Q10-Q12) areas, the Biclustering and Each Target approaches achieved the best performance in two of the three targets. The best average performance was obtained for Q11 (RMSE 0.361, MAE 0.131) and the worst for Q6 (RMSE 0.909, MAE 0.667) and Q9 (RMSE 0.934, MAE 0.560).

For Task 2, the best overall strategy was feature transformation through Biclustering (RMSE 0.805, MAE 0.483), with the best average metrics in 8 out of 12 targets. Compared to Task 1, there is more overlap in the outcome of the three strategies, and as such, the second best method (Each Target; RMSE 0.836, MAE 0.507) had the best average metrics in 5 questions. Also, unlike Task 1, there is no preferred strategy by domain, save for the respiratory questions (Q10-Q12) that are most easily predicted by biclustering-based models. The best average performance was attained in Q12 (RMSE 0.419, MAE 0.318) and the worst in Q10 (RMSE 1.191, MAE 0.818).

4.3. Model Comparison

We conducted experiments to predict the ALSFRS-R questions in the second evaluation using four machine-learning classifiers - Logistic Regression (LR), Random Forests (RF), Support Vector Machine (SVC), and XGBoost (XGB). We optimized their hyperparameters and validated their performance on a validation set derived from the provided training set as described in section 3.3. In addition to these classifiers, we also submitted two naïve approaches: Last Observation Carried Forward (LOCF) and Majority Class.

Table 6
Model results' summary, by feature selection and extraction strategy. Presented Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) report to the average performance of the 4 tested classifiers (LR, RF, SVC, XGBoost), in the test set. The performance metrics are provided for each target question and averaged across all of the strategy's models (Global), with the best outcome in bold.

	Strategy		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Global
Task 1	**Biclustering**	RMSE	**0.730**	**0.775**	0.616	**0.820**	0.909	1.008	1.015	0.843	1.095	**0.825**	0.378	0.761	0.815
		MAE	**0.393**	**0.488**	0.298	**0.440**	0.643	0.702	0.810	0.571	0.667	**0.548**	0.143	0.476	0.515
	All Sensors	RMSE	0.744	0.836	0.883	0.975	0.804	**0.909**	**0.884**	0.733	1.205	0.842	0.477	0.849	0.845
		MAE	0.417	0.571	0.488	0.560	0.536	**0.667**	**0.571**	0.429	0.821	**0.548**	0.190	0.536	0.528
	Each Target	RMSE	0.826	0.813	**0.548**	0.906	**0.765**	0.959	0.948	**0.595**	**0.934**	1.004	**0.361**	**0.703**	**0.780**
		MAE	0.464	0.536	**0.262**	0.488	**0.488**	0.714	0.679	**0.310**	**0.560**	0.631	**0.131**	**0.429**	**0.474**
Task 2	**Biclustering**	RMSE	**0.738**	0.910	0.726	**1.115**	1.028	0.892	**0.574**	0.698	**0.452**	**1.191**	**0.914**	**0.419**	**0.805**
		MAE	**0.432**	0.523	0.364	**0.659**	**0.614**	0.705	**0.341**	0.500	**0.227**	**0.818**	**0.295**	**0.318**	**0.483**
	All Sensors	RMSE	0.820	**0.799**	0.749	1.217	1.310	1.034	0.811	0.797	0.689	1.388	1.383	0.603	0.967
		MAE	0.545	**0.477**	0.432	0.727	0.977	0.864	0.554	0.500	0.386	0.977	0.568	0.364	0.614
	Each Target	RMSE	0.808	0.860	**0.686**	1.147	**0.993**	**0.875**	0.696	**0.665**	0.518	1.232	1.128	0.433	0.836
		MAE	0.455	**0.477**	**0.295**	0.727	0.636	**0.636**	0.455	**0.455**	0.273	0.864	0.500	**0.318**	0.507

Table 7 present the performance results for each model in predicting each target question, along with the overall results (average RMSE and MAE values across all questions). Notably, the LOCF approach performed the best for both tasks, indicating minimal variability between the ALSFRS-R scores of the first and second evaluations. On the other hand, the Majority Class approach was the worst performer, with RMSE values of 1.092 for Task 1 and 1.471 for Task 2. A potential reason for the classifiers' overall poor performance is that they were trained to predict the next score rather than specifically the second score, making the models too general for this particular task.

Regarding Task 1, questions Q3 (bulbar domain) and Q10 (respiratory domain) had the lowest prediction error using the LOCF approach (RMSE 0.218, MAE 0.048). Conversely, question Q9 (lower limb domain) predictions were the poorest, with the best classifier being RF (RMSE 0.873, MAE 0.381).

For Task 2, the conclusions are similar to those of Section 4.1. Questions regarding the respiratory domain, Q11, and Q12, were correctly predicted for all the evaluations. Particularly, the LOCF approach correctly predicted all the scores of question Q11, and the LR and RF classifiers accurately predicted all scores for question Q12 (RMSE 0.000, MAE 0.000). The most misclassified question was Q4 (upper limb domain), with an RMSE of 1.044 and MAE of 0.545 using the best-performing model (LOCF).

5. Conclusion

In a fast-acting and debilitating disease like ALS, the ability to predict how it evolves can be critical for clinical decision-making and life-prolonging therapy administration. Thus, the collection of sensor data can be a valuable resource for improving prognosis prediction, as it provides continuous monitoring of the patient's physiological status. This information can complement the periodic clinical assessments and possibly hint at the imminent occurrence of critical events, such as needing ventilation support. Machine learning techniques allow for meaningful insight to be extracted from these large datasets, which can potentially improve the performance of current prognosis prediction approaches or lead to the development of new ones. In the iDPP CLEF 2024 challenge, the main goal was to predict the ALSFRS-R scores (both clinical and self-assessed) of a patient's second assessment, given the first assessment and the sensor records between evaluations.

Our methodology consisted of independent multi-class models, each predicting an ALSFRS-R question. Four classification models were tested: Logistic Regression, Random Forest, XGBoost, and Support

Table 7
Results of the models. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics of four ML classifiers and two naïve approaches across the 12 target questions. The classifiers include Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), and XGBoost. The naïve approaches are the Last Observation Carried Forward (LOCF) and Majority Class. The performance metrics are provided for each target question and averaged across all the questions (Global).

	Model		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Global
Task 1	**LOCF**	RMSE	0.488	**0.309**	**0.218**	**0.690**	**0.535**	**0.577**	**0.488**	**0.535**	0.951	**0.218**	**0.309**	**0.577**	**0.491**
		MAE	0.143	**0.095**	**0.048**	**0.286**	**0.286**	**0.333**	**0.238**	**0.190**	0.429	**0.048**	**0.095**	**0.238**	**0.202**
	Majority Class	RMSE	1.512	0.976	1.512	1.254	1.113	1.34	1.327	1.175	1.690	0.309	0.378	0.724	1.092
		MAE	0.857	0.476	0.762	0.814	0.762	0.905	0.810	0.810	1.238	0.0952	0.143	0.333	0.659
	LR	RMSE	1.000	0.756	0.756	0.900	0.787	1.000	1.024	0.873	0.926	0.816	0.378	0.845	0.838
		MAE	0.619	0.381	0.381	0.524	0.524	0.714	0.762	0.571	0.571	0.476	0.143	0.429	0.508
	RF	RMSE	0.690	0.578	0.436	0.926	0.655	0.900	0.900	0.617	**0.873**	0.577	0.378	**0.577**	0.676
		MAE	0.381	0.238	0.190	0.476	0.429	0.619	0.619	0.286	**0.381**	0.238	0.143	**0.238**	0.353
	SVC	RMSE	0.976	0.787	0.617	0.900	1.000	1.234	1.113	0.655	1.291	0.756	0.378	0.951	0.888
		MAE	0.571	0.429	0.286	0.524	0.619	0.857	0.857	0.429	0.905	0.381	0.143	0.524	0.544
	XGBoost	RMSE	**0.309**	1.134	0.655	0.900	0.617	0.900	0.756	0.787	1.291	1.215	0.378	1.024	0.830
		MAE	**0.095**	1.095	0.333	0.429	0.381	0.619	0.476	0.429	0.810	1.095	0.143	0.952	0.571
Task 2	**LOCF**	RMSE	**0.674**	**0.674**	**0.426**	**1.044**	**0.739**	**0.603**	0.739	0.603	**0.302**	**0.522**	**0.000**	0.603	**0.577**
		MAE	**0.455**	**0.273**	**0.182**	**0.545**	**0.364**	**0.364**	**0.364**	0.364	**0.091**	**0.273**	**0.000**	0.182	**0.288**
	Majority Class	RMSE	1.348	0.905	1.168	1.314	1.477	1.809	1.651	1.044	1.883	1.758	2.089	1.206	1.471
		MAE	0.909	0.455	0.636	0.818	1.091	1.636	1.273	0.909	1.545	1.273	1.091	0.727	1.030
	LR	RMSE	0.798	0.790	0.905	1.168	1.168	0.953	**0.674**	0.798	0.603	1.279	1.537	**0.000**	0.890
		MAE	0.455	0.455	0.455	0.818	0.818	0.727	0.455	0.636	0.364	0.909	0.727	**0.000**	0.568
	RF	RMSE	0.905	0.905	0.739	1.087	1.279	0.905	**0.674**	0.798	**0.302**	1.128	1.508	**0.000**	0.852
		MAE	0.636	0.455	0.364	0.636	0.909	0.818	0.455	0.636	**0.091**	0.727	0.636	**0.000**	0.530
	SVC	RMSE	0.905	1.000	0.739	1.128	1.624	1.279	0.853	0.674	0.603	1.414	1.279	0.674	1.014
		MAE	0.636	0.636	0.364	0.727	1.364	1.091	0.545	0.455	0.364	1.091	0.545	0.273	0.674
	XGBoost	RMSE	**0.674**	0.739	0.522	1.206	1.168	1.000	1.044	**0.522**	**0.302**	1.732	1.206	1.000	0.926
		MAE	**0.455**	0.364	0.273	0.727	0.818	0.818	0.727	**0.273**	**0.091**	1.182	0.364	1.000	0.591

Vector Machine. The sensor data was handled first by deriving static features from the longitudinal ones using summarization techniques, i.e., by calculating summary statistics within an observation window before the target date. Then, the feature set was reduced using three methods: K-Best selection across all questions, K-Best selection by question, and biclustering. These models were also compared to baseline approaches Last Observation Carried Forward (LOCF) and Majority Class.

In both tasks, Random Forest yielded the best overall results but did not outperform LOCF, save for a few individual questions. Additionally, there was no consensus regarding the best feature selection or extraction approach. Independent K-Best selection and Biclustering were the best overall methods in tasks 1 and 2, respectively. However, further research is needed to capture the temporal patterns of sensors to fully understand their potential in tracking disease progression as measured by ALSFRS-R scores.

Acknowledgments

This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT) through project

AIpALS ref. PTDC/CCI-CIF/4613/2020 (https://doi.org/10.54499/PTDC/CCI-CIF/4613/2020), LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020), and PhD Research Scholarships to RB (2022.10727.BD), DFS ref. 2020.05100.BD (https://doi.org/10.54499/2020.05100.BD) and ENC ref. 2021.07810.BD (https://doi.org/10.54499/2021.07810.BD); and by BRAINTEASER project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017598.

References

[1] L. C. Wijesekera, P. Nigel Leigh, Amyotrophic lateral sclerosis, Orphanet journal of rare diseases 4 (2009) 1–22.

[2] J. Morris, Amyotrophic lateral sclerosis (ALS) and related motor neuron diseases: an overview, The Neurodiagnostic Journal 55 (2015) 180–194.

[3] S. R. Pfohl, R. B. Kim, G. S. Coan, C. S. Mitchell, Unraveling the complexity of amyotrophic lateral sclerosis survival prediction, Frontiers in neuroinformatics 12 (2018) 36.

[4] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. complete listing of the BDNF Study Group, et al., The ALSFRS-R: a revised als functional rating scale that incorporates assessments of respiratory function, Journal of the neurological sciences 169 (1999) 13–21.

[5] E. Beswick, T. Fawcett, Z. Hassan, D. Forbes, R. Dakin, J. Newton, S. Abrahams, A. Carson, S. Chandran, D. Perry, et al., A systematic review of digital technology to evaluate motor function and disease progression in motor neuron disease, Journal of Neurology 269 (2022) 6254–6268.

[6] R. P. van Eijk, J. N. Bakers, T. M. Bunte, A. J. de Fockert, M. J. Eijkemans, L. H. van den Berg, Accelerometry for remote monitoring of physical activity in amyotrophic lateral sclerosis: a longitudinal cohort study, Journal of neurology 266 (2019) 2387–2395.

[7] A. Maier, T. Holm, P. Wicks, L. Steinfurth, P. Linke, C. Münch, R. Meyer, T. Meyer, Online assessment of als functional rating scale compares well to in-clinic evaluation: a prospective trial, Amyotrophic Lateral Sclerosis 13 (2012) 210–216.

[8] S. A. Johnson, M. Karas, K. M. Burke, M. Straczkiewicz, Z. A. Scheier, A. P. Clark, S. Iwasaki, A. Lahav, A. S. Iyer, J.-P. Onnela, et al., Wearable device and smartphone data quantify als progression and may provide novel outcome measures, NPJ Digital Medicine 6 (2023) 34.

[9] J. W. van Unnik, M. Meyjes, M. R. J. van Mantgem, L. H. van den Berg, R. P. van Eijk, Remote monitoring of amyotrophic lateral sclerosis using wearable sensors detects differences in disease progression and survival: a prospective cohort study, Ebiomedicine 103 (2024).

[10] A. S. Gupta, S. Patel, A. Premasiri, F. Vieira, At-home wearables and machine learning sensitively capture disease progression in amyotrophic lateral sclerosis, Nature Communications 14 (2023) 5080.

[11] M. Straczkiewicz, M. Karas, S. A. Johnson, K. M. Burke, Z. Scheier, T. B. Royse, N. Calcagno, A. Clark, A. Iyer, J. D. Berry, et al., Upper limb movements as digital biomarkers in people with als, EBioMedicine 101 (2024).

[12] F. G. Vieira, S. Venugopalan, A. S. Premasiri, M. McNally, A. Jansen, K. McCloskey, M. P. Brenner, S. Perrin, A machine-learning based objective measure for als disease severity, NPJ digital medicine 5 (2022) 45.

[13] S. B. Rutkove, P. Narayanaswami, V. Berisha, J. Liss, S. Hahn, K. Shelton, K. Qi, S. Pandeya, J. M. Shefner, Improved als clinical trials through frequent at-home self-assessment: a proof of concept study, Annals of Clinical and Translational Neurology 7 (2020) 1148–1157.

[14] E. Tavazzi, E. Longato, M. Vettoretti, H. Aidos, I. Trescato, C. Roversi, A. S. Martins, E. N. Castanho, R. Branco, D. F. Soares, et al., Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review, Artificial Intelligence in Medicine (2023) 102588.

[15] J. Gordon, B. Lerner, Insights into amyotrophic lateral sclerosis from a machine learning perspective, Journal of Clinical Medicine 8 (2019) 1578.
[16] A. S. Martins, M. Gromicho, S. Pinto, M. de Carvalho, S. C. Madeira, Learning prognostic models using diseaseprogression patterns: Predicting the need fornon-invasive ventilation in amyotrophic lateralsclerosis, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021).
[17] J. Matos, S. Pires, H. Aidos, M. Gromicho, S. Pinto, M. de Carvalho, S. C. Madeira, Unravelling disease presentation patterns in als using biclustering for discriminative meta-features discovery, in: International Work-Conference on Bioinformatics and Biomedical Engineering, Springer, 2020, pp. 517–528.
[18] S. C. Madeira, A. L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM transactions on computational biology and bioinformatics 1 (2004) 24–45.
[19] E. N. Castanho, H. Aidos, S. C. Madeira, Biclustering fmri time series: a comparative study, BMC bioinformatics 23 (2022) 192.
[20] D. F. Soares, R. Henriques, M. Gromicho, M. de Carvalho, S. C. Madeira, Learning prognostic models using a mixture of biclustering and triclustering: Predicting the need for non-invasive ventilation in amyotrophic lateral sclerosis, Journal of Biomedical Informatics 134 (2022) 104172.
[21] R. Henriques, S. C. Madeira, Flebic: Learning classifiers from high-dimensional biomedical data using discriminative biclusters with non-constant patterns, Pattern Recognition 115 (2021) 107900.
[22] R. Henriques, S. C. Madeira, Triclustering algorithms for three-dimensional data analysis: a comprehensive survey, ACM Computing Surveys (CSUR) 51 (2018) 1–43.
[23] D. F. Soares, R. Henriques, S. C. Madeira, Comprehensive assessment of triclustering algorithms for three-way temporal data analysis, Pattern Recognition (2024) 110303.
[24] D. F. Soares, R. Henriques, M. Gromicho, M. de Carvalho, S. C. Madeira, Triclustering-based classification of longitudinal data for prognostic prediction: targeting relevant clinical endpoints in amyotrophic lateral sclerosis, Scientific Reports 13 (2023) 6182.
[25] A. V. Carreiro, P. M. Amaral, S. Pinto, P. Tomás, M. de Carvalho, S. C. Madeira, Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in amyotrophic lateral sclerosis, Journal of biomedical informatics 58 (2015) 133–144.
[26] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. Di Nunzio, P. Fariselli, J. García Dominguez, A. G. Marta Gromicho, E. Longato, S. Madeira, U. Manera, S. Marchesin, L. Menotti, G. Silvello, E. Tavazzi, E. Tavazzi, I. Trescato, M. Vettoretti, B. D. Camillo, N. Ferro, Overview of iDPP@CLEF 2024: The Intelligent Disease Progression Prediction Challenge, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024, 2024.
[27] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. Di Nunzio, P. Fariselli, J. García Dominguez, A. G. Marta Gromicho, E. Longato, S. Madeira, U. Manera, S. Marchesin, L. Menotti, G. Silvello, E. Tavazzi, E. Tavazzi, I. Trescato, M. Vettoretti, B. D. Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, 2024.
[28] R. Branco, J. Valente, A. Martins, D. Soares, E. Castanho, S. Madeira, H. Aidos, Survival analysis for multiple sclerosis: predicting risk of disease worsening, in: CLEF, 2023.
[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[30] Y. Kluger, R. Basri, J. T. Chang, M. Gerstein, Spectral biclustering of microarray data: Coclustering genes and conditions, Genome Research 13 (2003) 703–716. doi:10.1101/gr.648603.
[31] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, Journal of Machine Learning Research 18 (2017) 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[33] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[34] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, Machine Learning and Knowledge Discovery in Databases (2011) 145–158.

[35] P. Szymański, T. Kajdanowicz, A network perspective on stratification of multi-label data, in: L. Torgo, B. Krawczyk, P. Branco, N. Moniz (Eds.), Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, volume 74 of *Proceedings of Machine Learning Research*, PMLR, ECML-PKDD, Skopje, Macedonia, 2017, pp. 22–35.

[36] P. Szymański, T. Kajdanowicz, A scikit-based Python environment for performing multi-label classification, ArXiv e-prints (2017). arXiv:1702.01460.