CEUR-WS.org/Vol-3740/paper-125.pdf

# Using Wearable and Environmental Data to Improve the Prediction of Amyotrophic Lateral Sclerosis and Multiple Sclerosis Progression: an Explorative Study

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2024

Elena Marinello[1], Alessandro Guazzo[1], Enrico Longato[1], Erica Tavazzi[1], Isotta Trescato[1], Martina Vettoretti[1] and Barbara Di Camillo[1,2,*]

[1]*Department of Information Engineering, University of Padova, Padova, Italy*
[2]*Department of Comparative Biomedicine and Food Science, University of Padova, Padova, Italy*

## Abstract

Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) are chronic diseases with a severe impact on patients' lives. Both diseases create significant psychological and economic burdens due to alternating acute phases requiring hospital and home care. One possible solution could be the employment of sensor data to develop predictive models that can assist clinicians in making treatment and therapeutic decisions. In the context of the iDPP@CLEF 2024 challenge, this work aims to develop and compare different machine-learning approaches for predicting the Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised (ALSFRS-R) scores in ALS patients, and relapses in MS patients, using wearable and environmental data, respectively. Specifically, the analysis focuses on the impact of these data and seeks to determine whether their incorporation enhances predictive performance. The results showed that there is indeed an improvement in the models' performance when sensor data are considered, in both the disease. In particular, in the case of ALS the Root Mean Square Error (RMSE) range, over the predicted twelve ALSFRS-R score, improved from [0.463-0.733] to [0.286-0.582] when incorporating the wearable data, as well as in the case of MS, where the inclusion of environmental data has improved the prediction of relapse, with the RMSE decreasing from 72.992 to 69.564.

## Keywords

Amyotrophic Lateral Sclerosis, Multiple Sclerosis, Logistic Regression, Ridge Regression, Random Forest, Wearable Data, Environmental Data

## 1. Introduction

Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) are chronic neurodegenerative diseases. ALS affects the motor neurons, causing progressive degeneration of nerve cells in the spinal cord and brain, leading to an average life expectancy of three to five years [1]. ALS symptoms usually are primarily related to weakness in the upper and lower limbs, or slurred speech and difficulty in swallowing [2]. On the other hand, MS affects the myelinated axons in the central nervous system, causing damage to both the myelin and the axons to varying degrees. The progression of MS is highly variable and unpredictable, with the most common phenotype being relapsing-remitting: a progression pattern characterized by periods of exacerbations of the symptoms, called relapse, alternated with more stable periods [3].

Given the heterogeneous and unpredictable nature of these diseases, patients end up alternating periods in the hospital and at home, while dealing with the uncertainty of how long each acute or stable phase will last [4]. This can represent a psychological and economic burden for both patients and

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

[*] Corresponding author.

elena.marinello@unipd.it (E. Marinello); guazzoales@dei.unipd.it (A. Guazzo); enrico.longato@unipd.it (E. Longato); erica.tavazzi@unipd.it (E. Tavazzi); isotta.trescato@phd.unipd.it (I. Trescato); martina.vettoretti@unipd.it (M. Vettoretti); barbara.dicamillo@unipd.it (B. D. Camillo)

0009-0007-2445-5762 (E. Marinello); 0000-0001-5155-2567 (A. Guazzo); 0000-0001-5940-645X (E. Longato); 0000-0001-6188-6413 (E. Tavazzi); 0000-0003-0625-993X (I. Trescato); 0000-0002-5020-1818 (M. Vettoretti); 0000-0001-8415-4688 (B. D. Camillo)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

caregivers. Clinicians, on their part, would welcome tools that can assist them throughout all stages of patient treatment by offering personalized therapeutic recommendations and identifying when urgent interventions are necessary. Predictive tools can indeed be powerful in predicting the progression of ALS disability and the occurrence of relapses in MS.

In the context of the iDPP@CLEF 2024 challenge, participants were asked to predict the progression of the ALS patients' disability status using prospective data, and predict the occurrence of relapses for MS patients by exploiting environmental and MS-specific retrospective data [5, 6]. The Challenge consisted of three tasks, described in the following sections: Section 1.1 and 1.2 refer to Task 1 and Task 2, respectively, while Section 1.3 refers to Task 3.

### 1.1. Task 1: ALS Disability Score from Wearable Data

Task 1 focused on using data collected through wearable devices to predict the patient's disability status measured by the twelve scores of the revised ALS functional rating scale (ALSFRS-R) [7]. These ALSFRS-R scores were assigned by medical doctors during routine visits scheduled every three months. The goal of this task was to determine whether the ALSFRS-R scores assigned by clinical experts could be reliably predicted from wearable data.

### 1.2. Task 2: ALS Patient Self-assessment Score from Wearable Data

Similarly to Task 1, Task 2 consisted of the use of data collected through wearable devices, to predict the patient's disability status, measured by the ALSFRS-R scores. In this case, the scores were self-assessed by patients via an auto-evaluation questionnaire delivered through an app once a month. The goal was to determine whether the ALSFRS-R scores obtained from self-assessment questionnaires could be reliably predicted from wearable data.

### 1.3. Task 3: Relapse from EDDS Sub-scores and Environmental Data

Task 3 considered the prediction of an MS relapse using environmental data and Expanded Disability Status Scale subscores (EDSS) [8]. The goal of this task was to explore whether exposure to different pollutants can be considered a useful variable in predicting the occurrence of relapses in MS patients.

To address the proposed problems, a broad set of predictive models based on different methodological approaches were trained using different subsets of the variables, provided by the challenge organizers. This study aimed to evaluate whether considering wearable data to predict ALS disability and environmental data to predict MS relapses leads to better performance with respect to models that only consider disease-specific variables collected during routine visits. To ensure consistency, all models were trained using a common framework including feature selection (via backward elimination), and hyperparameter optimization (via random search). The results suggest that collecting data from wearable devices can improve the prediction of ALS disability status. However, patients must be properly trained to use the sensors correctly. Similarly, environmental data can be beneficial for predicting the progression of MS by identifying the occurrence of relapses, focusing mainly on sensor data recorded a few days before the relapse.

The paper is organized as follows: Section 2 introduces related works and the main methodological approaches implemented until now to address ALS and MS progression prediction. Section 3 describes the methodologies employed in this study in terms of data processing and the machine-learning techniques used. Section 4 discusses the obtained results and, finally, Section 5 summarizes the key take home messages of this work.

## 2. Related Work

Different approaches have been proposed in the literature to predict the prognosis of ALS and MS patients. For both of the diseases, prediction tasks frequently employ a variety of machine-learning methodologies, with classification and regression being the most common approaches. The choice between these methods typically depends on the specific research question and the chosen outcome [9].

Regarding ALS prognosis, most studies aimed to estimate changes in the ALSFRS-R over time [10, 11, 12, 13, 14]. Different studies classified patients by disease progression rates (e.g., Slow/Fast, Low/High) [15, 16, 17], while others have developed a model to predict when a patient will need Non-Invasive Ventilation (NIV) support within a given time window [18, 19, 20]. Relevant biomarkers for prediction include BMI, Forced Vital Capacity (FVC), age at onset, and disease duration, as well as longitudinal data (e.g., slope, minimum, maximum, mean, standard deviation) [10]. Magnetic resonance imaging (MRI) has also shown a significant impact on prediction, alongside these clinical variables [21]. Regression models include Random Forest (RF) regressor and generalized boosting models [18, 22]. Recently, also graphical modeling techniques such as Dynamic Bayesian Networks (DBN) have been employed to model ALS disease progression [23]. Classification models included Support Vector Machine (SVM) [16], and RF classifier [15].

On the other hand, most of the models related to MS prognosis considered as outcomes the occurrence of relapses [24, 25] and the evolution over time of the EDSS [26, 27, 28, 29]. The models most commonly used for classification were Logistic Regression (LR) and SVM [30], while for regression, the most popular technique was Linear Regression [31]. Demographic (including age and sex), clinical, MRI (such as T2 lesion volume or number and brain atrophy), cerebrospinal fluid, and electrophysiology variables were retained as predictors in the models studied in the literature [31].

In general, for both ALS and MS, the inclusion of wearable and environmental data, respectively, in literature models is limited [32, 33]. Typically, studies focus on defining a baseline, where data are collected, and then developing a model based on this baseline to provide predictions for future outcomes [34, 35]. The main limitation of this approach is that it does not thoroughly exploit the dynamic aspect of the disease described by the full temporal evolution of data sequences, conversely to what is extensively investigated within the scope of the Challenge.

# 3. Methodology

A common data processing was performed for Task 1 and Task 2 involving ALS data, instead, the data processing for Task 3, which considered MS data, was slightly different. Then, a single model-training framework was considered for all methodological approaches across the three tasks. The following sections describe: the data processing steps needed to obtain the final set of input variables for Tasks 1, 2 (Section 3.1.1), and 3 (Section 3.1.2); the training framework used to develop the models (Section 3.2); and the description of the submitted runs (Section 3.4).

## 3.1. Data Processing

### 3.1.1. ALS Data Processing (Tasks 1 and 2)

The structure of the datasets provided for Task 1 and 2 was identical. The main difference between the data provided for these tasks lay in how ALSFRS-R scores were collected. In fact, for Task 1 ALSFRS-R scores were assigned by clinicians during routine visits performed more or less every three months. Instead, for Task 2, ALSFRS-R scores were self-assigned by the patients via a questionnaire delivered periodically ($\sim$ once a month) through the BRAINTEASER app. Hence, the same processing pipeline was adopted for these two tasks.

Six static variables evaluated at the first visit were available, namely: sex, diagnostic delay, age at diagnosis, FVC, weight, and BMI. The only processing performed on these static variables concerned the sex variable which was mapped to a boolean variable equal to 0 for male patients and 1 for female patients.

All ALSFRS-R measurements collected for each patient were made available to participants despite the Task 1 and 2 goals being only the prediction of the ALSFRS-R subscores following the first visit (Task 1) or of the self-assessment score (Task 2). Hence, all available information was fully exploited to obtain a more rich and robust dataset. Specifically, each pair of consecutive ALSFRS-R subscores was considered as an independent entry characterized by the same static information of the patient they belonged to. The first set of ALSFRS-R subscores of each pair was used as input variables named start_Q*, where * represents the ALSFRS-R question number and ranges from 1 to 12. Instead, the second set of ALSFRS-R subscores of each pair were used as the target variables named end_Q*. The final sample size of data used to train models for the first task was of 131 entries (from 52 unique patients) and the one of data used to train models for the second task was of 163 entries (from 52 unique patients).

For each patient, 90 variables collected multiple times through wearable sensors were available. The processing of these variables consisted of the extraction of first-order descriptors (such as mean, first and last recorded values, and minimum and maximum values) considering all values recorded within a time window starting from the date of the start ALSFRS-R of the considered entry to the date of the end ALSFRS-R score of the same entry. The window length, expressed in days, was also included in the set of possible predictors. Moreover, the slopes of change of the following variables were also considered: total_calories, total_steps, spo2_av, heart_rate_mean, heart_rate_baseline. The slope of change was obtained as the angular coefficient of a linear fit of all recorded values for each variable within the considered time window.

To build the training set, it was instrumental to consider ALSFRS-R pairs collected after the first visit, in order to obtain a robust and rich set of variables extracted from wearable data. The richness and quality of such data tend to improve over time as the patient learns how to properly use, and becomes more familiar with, the device provided at the first visit.

After this first processing step, 487 variables were available for each entry in the dataset. Specifically, one variable for the unique patient identifier, one variable for the window length expressed in days, 12 variables for the start ALSFRS-R scores, 12 target variables for the ALSFRS-R scores to be used as outcomes, $90 * 5 = 450$ variables for the first-order descriptors of the 90 wearable sensor variables, 5 variables for the considered slopes of change, and the 6 static variables.

From this full set of variables, those with more than $50\%$ missing values and those that were almost constant (auto-correlation coefficient $> 0.9$) were removed. Finally, collinear variables were removed by iteratively excluding those with a correlation coefficient $> 0.9$. After this step, 131 out of 487 variables were considered for Task 1, and 134 out of 487 variables were considered for Task 2.

Then, normalization was performed to avoid introducing bias related to the different dynamic ranges of each variable and to promote consistency between the scale of the coefficients that might be estimated during model training. Specifically, min-max scaling was used and the normalization parameters were derived considering only the whole training set and applied to the test set.

Finally, the imputation of missing values in the processed input variables was performed using the mice R package [36]. Also for the imputation, parameters were estimated on the whole training set and applied to the test set.

### 3.1.2. MS Data Processing (Task 3)

The processing concept for Task 3 was similar to the one proposed for Task 1 and 2 but had to be adapted considering the different structure of data available for this task.

Fifteen static variables evaluated at the first visit were available. Five variables were related to demographic information, five variables were related to MS diagnosis, and five variables were related to symptoms. The sex variable was mapped to a boolean variable equal to 0 if the patient was male and 1 if female. The variable centre was mapped to a boolean variable equal to 0 if the patient was followed at the clinic in Pavia and 1 if at the clinic in Turin. The variable residence classification consisted of three possible levels: cities, towns, and rural area. This variable was mapped to two dummy variables: residence_city and residence_rural_area. The variable ethnicity was excluded as almost all patients were caucasian. Two variables related to diagnosis criteria were excluded as almost all patients were diagnosed according to the same criterion. After these steps, 12 static variables remained.

Multiple EDSS recordings were also available from the baseline date to the date of the first relapse.

Hence, first-order descriptors were extracted also for the EDSS value considering all measurements within this time window.

For each patient, a set of 20 environmental measurements related to pollutant levels and meteorological indicators were available. Such measurements were available both before and after the baseline. Hence, similarly to what was done for wearable sensor data in Tasks 1 and 2, a set of first-order descriptors (such as mean, first and last recorded values, and minimum and maximum values) was extracted for each variable considering all values recorded within two time windows. The first time window started at the date of the first available environmental measure and ended at the baseline date. The second time window, instead, started at the baseline date and ended at the first recorded relapse date.

After this first processing step, 219 variables were available for each patient in the dataset. Specifically, one variable for the unique patient identifier, one target variable for the relapse week to be used as the outcome, $20{*}5 = 100$ variables for the first-order descriptors of the 20 environmental variables measured before the baseline, $20{*}5 = 100$ variables for the first-order descriptors of the 20 environmental variables measured after the baseline, 5 variables for the first-order descriptors of the EDSS measurements, and the 12 static variables.

From this full set of variables, those with more than $50\%$ missing values and those that are almost constant (auto-correlation coefficient $> 0.9$) were removed. Finally, collinear variables were removed by iteratively excluding those with a correlation coefficient $> 0.9$. After this step, 69 out of 219 variables were considered for Task 3. Two patients were also excluded as almost all their variables were missing, hence, 197 unique patients were considered to train the MS models.

Following what was done for the previous tasks, normalization was performed via min-max scaling and the imputation of missing values in the processed input variables was performed using the mice R package.

### 3.2. Model Training and Evaluation

In Tasks 1 and 2, the prediction targets were the 12 ALSFRS-R scores evaluated, respectively by the clinician and the patients themselves. Each score must be predicted independently and it was an integer within the range [0-4]. Intuitively, this problem can be cast as a multiclass classification with five classes. However, it can also be framed as a regression problem by modifying the model output by rounding it to the nearest integer. Instead, in Task 3, the goal was to predict the week of the first relapse occurrence after the baseline, and, as the weeks are not within a finite range, this can only be approached as a regression problem. However, as the challenge submission rules require an integer value also for the predicted relapse week, the output of regression models developed for Task 3 was rounded to the nearest integer as well.

The core of the model training framework involved the Backward Feature Selection technique [37] and the model's performances were evaluated through the Root Mean Squared Error score [38]. The process started with all the features and iteratively they were removed one by one. At each iteration, for every feature combination, hyperparameter tuning was performed via random search over a given hyperparameter grid [39], using a 5-fold cross-validation (CV). The subset of features that resulted in the lowest RMSE score, was then chosen to train a final model. Its hyperparameters were optimized again using 5-fold CV and random search within the same hyperparameter space. Ultimately, this optimized model was tested on an independent test set, and the results were submitted to the challenge organizers for performance evaluation.

This model training framework was designed to be flexibile, allowing its application across the three different tasks with a variety of methodological approaches. The approaches considered in this study included both linear models (LR and ridge regression), as well as non-linear models (RF). For each of these models, different sets of hyperparameters were tested. For the LR, a single hyperparameter needs optimization: the strength of the regularization applied to the model, C. Similarly, for the ridge regression, the only hyperparameter that needs optimization is the strength of the L2 regularisation, $\alpha$. Both C and $\alpha$ were randomly sampled from 250 values in a log-uniform distribution with support [$10^{-4}$ - $10^{4}$]. Finally, the RF's hyperparameter space consisted of two hyperparameters: the number of trees in each RF, which was uniformly sampled in the interval [50 - 500], and the maximum depth of each tree, which was uniformly sampled in the interval [1 - 100]). By default, the square root of the total number of features was evaluated at each node for splitting.

## 3.3. Considered Subsets of Input Variables

To evaluate whether considering wearable data to predict ALS disability and environmental data to predict MS relapses led to better performance with respect to models that only consider disease-specific variables collected during routine visits, different sets of variables were considered as input for the predictive models. Hence, for Tasks 1 and 2, the target ALSFRS-R value (e.g., end_Q1, see Section 3.1.1) was first predicted by simply holding the corresponding initial ALSFRS-R evaluation (e.g., start_Q1 see Section 3.1.1). The idea behind this approach is to provide a baseline reference point that does not involve any particular prediction model. Then, to provide a slightly more complex benchmark approach, a LR model was trained using only the 12 initial ALSFRS-R scores (e.g., all start_Q*) as possible input variables. The idea behind this second set of considered features is to assess whether considering scores from other ALSFRS-R questions leads to a more accurate prediction of the target ALSFRS-R score with respect to the one obtained by simply holding its initial value. Finally, different models were trained using all available variables (i.e., static, ALSFRS-R, and wearable data) to evaluate whether models developed including also data collected through wearable devices led to better performance with respect to the one developed using only the initial ALSFRS-R scores. The models included LR, ridge regression, RF regressor, and RF classifier.

Similarly, for Task 3, first, a ridge model was trained considering as possible input only static and EDSS variables. Then, a ridge and an RF regressor were trained after including environmental-derived variables in the pool of possible predictors. The idea behind this approach was to check whether including environmental data could improve the first-relapse-week prediction with respect to models that only consider data collected at the first visit and EDSS evaluations.

## 3.4. Description of Submitted Runs

The following runs were submitted for Tasks 1 and 2:

- **Logistic regression (logistic):** LR model with multiclass outcome. All available variables were considered in the pool of possible predictors. Each question was predicted with its independent model trained specifically for that question.
- **Logistic regression considering only ALSFRS-R scores (logistic_ALSFRS):** LR model with multiclass outcome. Only start_Q* variables were considered in the pool of possible predictors. Each question was predicted with its independent model trained specifically for that question.
- **Random Forest classifier (rf):** RF classifier with multiclass outcome. All available variables were considered in the pool of possible predictors. Each question was predicted with its independent model trained specifically for that question.
- **Ridge regression (ridge):** Ridge regression model. All available variables were considered in the pool of possible predictors. Each question was predicted with its independent model trained specifically for that question.
- **Random Forest regressor (rf_reg):** RF regressor model. All available variables were considered in the pool of possible predictors. Each question was predicted with its independent model trained specifically for that question.
- **hold:** Each question was predicted by holding its starting value (i.e., considering the start_Q* variables as predicted end_Q* targets)
- **average:** Each predicted score was obtained as the average output of the LR, RF classifier, ridge, and RF regressor models rounded to the nearest integer (i.e., column-wise average rounded to the nearest integer of logistic, rf, ridge, and rf_reg runs).
- **optrun:** Each question was predicted with the best-performing model for that question (i.e., the one highlighted in bold in Table 1 and 3).

The following runs were submitted for Task 3:

- **Ridge regression (ridge):** Ridge regression model. All available variables were considered in the pool of possible predictors.
- **Ridge regression without considering environmental data (ridge_noenv):** Ridge regression model. Environmental variables were excluded from the pool of possible predictors.
- **Random Forest regressor (rf_reg):** RF regressor model. All available variables were considered in the pool of possible predictors.
- **average:** Each predicted first relapse week was obtained as the average output of the ridge and RF regressor models rounded to the nearest integer (i.e., column-wise average rounded to the nearest integer of ridge and rf_reg runs).

## 4. Results

The results for the three tasks are reported in the sections below. For Tasks 1 and 2, the results for ALSFRS-R scores prediction are reported in Section 4.1 and Section 4.2, respectively. For Task 3, the results for the week of the occurrence of the first relapse are reported in Section 4.3.

### 4.1. Task 1 Results

Table 1 presents the CV results for Task 1. Each column represents one of the predicted ALSFRS-R scores (Q1 - Q12), while the rows indicate the considered models. Each cell displays the average CV RMSE. RMSE values highlighted in bold represent the lowest value of each column, thus indicating the best-performing model for each predicted question.

The ridge model was the best-performing one for six out of twelve scores (Q1, Q4, Q6, Q7, Q11, Q12), with RMSE values ranging between 0.228 for Q1 and 0.570 for Q7. The LR model, when considering all available variables, also showed reliable performance, achieving the best prediction for four out of twelve scores (Q2, Q3, Q9, Q10). Its RMSE values ranged from 0.286 for Q2 to 0.582 for Q9. Conversely, the RF regressor yielded the best predictions for Q5 and Q8, with RMSE scores of 0.508 and 0.479, respectively. Finally, the hold approach and RF classifier were the worst-performing among all the models. Additionally, the LR model using only the ALSFRS-R score did not perform well, suggesting that performance improved when wearable data was added. In general, it is possible to observe that adding first all the ASLFRS-R scores, and consequentially all the other sensor variables, increased the performance in the cross-validation phase, leading to lower RMSE values.

Table 2 shows the results of Task 1 submitted runs as evaluated by the challenge organizers. The name of the submitted run is reported in the first column of Table 2. Then, columns two and three of Table 2 show the two metrics used by the organizers to evaluate participants' submitted runs on the independent test set: RMSE and Mean Absolute Error (MAE), respectively.

Results observed in CV were not confirmed on the test set, with the best-performing model being the hold method (RMSE = 0.491, MAE = 0.202) and the LR using all available variables yielding the worst result (RMSE = 0.830, MAE = 0.511). One possible explanation could be that the training set is more robust compared to the test set, since it includes data from later visits, while the test set only contains data from the initial visits. Therefore, these results are likely due to insufficient data collection during the initial visits when patients either have not started using the wearable devices or are still becoming familiar with how to use them.

**Table 1**
CV RMSE values for methods considered in Task 1. Each column represents an ALSFRS-R question. The minimum values of each column are highlighted in bold.

| Model | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LR | 0.395 | **0.286** | **0.288** | 0.526 | 0.522 | 0.606 | 0.604 | 0.568 | **0.582** | **0.243** | 0.393 | 0.504 |
| LR ALSFRS only | 0.463 | 0.456 | 0.438 | 0.568 | 0.583 | 0.700 | 0.733 | 0.640 | 0.805 | 0.507 | 0.470 | 0.548 |
| Ridge | **0.228** | 0.351 | 0.302 | **0.495** | 0.533 | **0.493** | **0.570** | 0.532 | 0.676 | 0.318 | **0.389** | **0.439** |
| RF classifier | 0.416 | 0.379 | 0.386 | 0.591 | 0.564 | 0.580 | 0.636 | 0.589 | 0.619 | 0.461 | 0.471 | 0.557 |
| RF regressor | 0.443 | 0.370 | 0.397 | 0.512 | **0.508** | 0.561 | 0.621 | **0.479** | 0.648 | 0.455 | 0.457 | 0.511 |
| Hold | 0.531 | 0.479 | 0.471 | 0.665 | 0.710 | 0.796 | 0.762 | 0.654 | 0.856 | 0.553 | 0.546 | 0.630 |

**Table 2**
Runs RMSE and MAE values for methods considered in Task 1. The minimum values are highlighted in bold.

| Runs | RMSE | MAE |
|---|---|---|
| logistic | 0.830 | 0.511 |
| logistic_ALSFRS | 0.636 | 0.341 |
| ridge | 0.687 | 0.392 |
| rf | 0.650 | 0.361 |
| rf_reg | 0.636 | 0.373 |
| **hold** | **0.491** | **0.202** |
| average | 0.596 | 0.333 |
| optrun | 0.707 | 0.412 |

## 4.2. Task 2 Results

Table 3 presents the CV results for Task 2. Each column represents one of the predicted ALSFRS-R scores (Q1 - Q12), while the rows indicate the considered models. Each cell displays the average CV RMSE. RMSE values highlighted in bold represent the lowest value of each column, thus indicating the best-performing model for each predicted question.

In this task, the LR model, when considering all available variables, achieved the best results for seven out of twelve scores (Q2, Q3, Q5, Q7, Q10, Q11, Q12), with RMSE values ranging between 0.139 for Q3 and 0.595 for Q10. The ridge regression, also showed good performance compared to other models, achieving the best prediction for four out of twelve scores (Q1, Q4, Q6, Q9). Its RMSE values ranged from 0.292 for Q1 to 0.449 for Q6. Conversely, the RF regressor yielded the best prediction only for Q8 with an RMSE of 0.372. Finally, the hold and RF classifier performed the worst among all the models. In general, it is possible to observe that adding first all the ASLFRS-R scores, and then all the other sensor variables, led to a performance increase in the CV phase, resulting in lower RMSE values.

Table 2 shows the results of Task 2 submitted runs as evaluated by the challenge organizers. The name of the submitted run is reported in the first column of Table 2. Then, columns two and three of Table 2 show the two metrics used by the organizers to evaluate participants' submitted runs on the independent test set: RMSE and MAE, respectively.

Results observed in CV were not confirmed on the test set also for this second task, with the best-performing model being once again the hold method (RMSE = 0.577, MAE = 0.287) and the LR with wearable data available yielding the worst results on the test set (RMSE = 0.9930, MAE = 0.659). These results are in line with those observed in Task 1. Additionally, the scores assigned during this period are based on self-evaluation, which may further impact the accuracy of the data.

Overall, in this task, the RMSE values were lower than those obtained in Task 1, especially for the hold method. This improvement may be attributed to the fact that clinicians are able to better assign ALSFRS-R scores during visits, resulting in greater variability which leads to a more challenging prediction task. Instead, patients are typically more conservative and tend to assign similar scores between questionnaires. This leads to less variability, which makes the prediction task slightly easier, especially for the hold method.

**Table 3**
CV RMSE values for methods considered in Task 2. Each column represents an ALSFRS-R question. The minimum values of each column are highlighted in bold.

| Model | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LR | 0.360 | **0.453** | **0.139** | 0.466 | **0.525** | 0.497 | **0.412** | 0.447 | 0.452 | **0.595** | **0.414** | **0.263** |
| LR ALSFRS Only | 0.399 | 0.551 | 0.291 | 0.513 | 0.621 | 0.620 | 0.565 | 0.454 | 0.606 | 0.738 | 0.553 | 0.309 |
| Ridge | **0.292** | 0.500 | 0.219 | **0.437** | 0.535 | **0.449** | 0.437 | 0.378 | **0.381** | 0.701 | 0.702 | 0.358 |
| RF classifier | 0.357 | 0.490 | 0.223 | 0.481 | 0.565 | 0.531 | 0.465 | 0.412 | 0.418 | 0.688 | 0.769 | 0.347 |
| RF regressor | 0.367 | 0.513 | 0.280 | 0.497 | 0.525 | 0.567 | 0.488 | **0.372** | 0.492 | 0.638 | 0.680 | 0.345 |
| Hold | 0.384 | 0.701 | 0.313 | 0.514 | 0.631 | 0.597 | 0.612 | 0.525 | 0.602 | 0.821 | 0.917 | 0.450 |

**Table 4**
Runs RMSE and MAE values for methods considered in Task 2. The minimum values are highlighted in bold.

| Runs | RMSE | MAE |
|---|---|---|
| logistic | 0.993 | 0.659 |
| logistic_ALSFRS | 0.854 | 0.500 |
| ridge | 0.850 | 0.545 |
| rf | 0.778 | 0.515 |
| rf_reg | 0.818 | 0.515 |
| **hold** | **0.577** | **0.287** |
| average | 0.783 | 0.493 |
| optrun | 0.962 | 0.606 |

## 4.3. Task 3 Results

Table 5 reports the CV results for Task 3. Each row shows a considered model and its corresponding CV RMSE. The RMSE value highlighted in bold represents the lowest score, indicating the best-performing approach.

The ridge regression with environmental data performed best among others, with an RMSE equal to 69.564. However, in the independent test set, the best performance was achieved without including environmental variables as evidenced in Table 6.

In Task 3, the RMSE is very high, indicating low precision in predicting the relapse week. During the training phase, incorporating environmental data helped achieve better results. However, in the test phase, the performance was better without the environmental data. This discrepancy is likely due to the presence of significant sequences of missing data that needed to be imputed, as there are long intervals between visits in both the MS training and test sets.

**Table 5**
CV RMSE values for methods considered in Task 3. The minimum value is highlighted in bold.

| Model | RMSE |
|---|---|
| **Ridge** | **69.564** |
| Ridge without environmental data | 72.992 |
| RF regressor | 74.972 |
| Average | 82.702 |

**Table 6**
Runs RMSE and MAE values for methods considered in Task 3. The minimum values are highlighted in bold.

| Runs | RMSE | MAE |
|---|---|---|
| ridge | 89.83 | 68.59 |
| **ridge_no_env** | **78.62** | **61.37** |
| rf_reg | 79.73 | 66.63 |
| average | 79.25 | 65.80 |

## 5. Conclusions and Future Work

This study aimed at addressing the three tasks proposed within the iDPP@CLEF 2024 challenge, while also evaluating whether the inclusion of sensor and environmental data helps in improving prediction of ALS and MS progression.

The challenge consisted of three different tasks. In Task 1 and Task 2, the goal was to predict the ALSFRS-R scores, assigned, respectively, by clinicians and by the patients themselves. Instead, Task 3 consisted of predicting the week of the first relapse for MS patients. A flexible training workflow was developed in order to evaluate different methodological approaches and different subsets of input variables under a common, robust training workflow. For Task 1 and Task 2, both classification and regression approaches were explored, namely: LR, ridge regression, RF regressor and RF classifier. In Task 3, only regression models were considered due to the different nature of this task, namely: ridge regression and RF regressor.

In the first two tasks, classification approaches were able to better capture the ALSFRS-R scores variability among the five classes. Instead, the regression approach tended to frequently predict the mean value within the range [0-4]. Moreover, in these tasks, the best CV results were achieved by the ridge regression and LR when including variables derived from wearable devices. On the contrary, when evaluating the models on the independent test sets, the best results were obtained by the hold method. The robustness of the results during CV can be attributed to the nature of the training set, which includes data from all visits. This results in a richer, more complete, and robust dataset characterized by a more refined wearable data collection process with respect to the test set which included only the first couple of visits when patients are still getting familiar with the data collection process and the BRAINTEASER app. Hence, the test data were more noisy and sparse.

In Task 3, the best CV results were achieved by the ridge model incorporating the environmental data. On the contrary, in the independent test set the best performance was obtained by the ridge model without considering the environmental data. This weak result for this task could be attributed to the not properly optimized variable creation process which was designed for the first two tasks and directly applied also in the third task. One possible solution could be to consider dynamic variables instead of computing first-order descriptors, given the long periods between visits, and consequently employ models that account for these dynamic data.

In conclusion, the developed models performed well within the iDPP@CLEF 2024 challenge, while contributing to raise important considerations that go beyond the competition itself. In fact, Tasks 1 and 2 results suggest that collecting wearable data can be a viable path to follow in order to improve the prediction of ALS disability status. However, a key condition that must be respected in order to benefit from the inclusion of these data, is that patients must be properly informed, trained, and followed in order to obtain rich and high-quality data over long periods of time. Otherwise, it might be more effective to rely on data that are commonly collected during routine visits of ALS patients. On the other hand, regarding MS, since the given environmental data and observations have been measured also after the relapse that needed to be predicted, it would be more effective to focus only on the environmental pollutants measured a few days before the relapse, as also confirmed by the literature [40].

## References

[1] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, M. C. Zoing, Amyotrophic lateral sclerosis, The Lancet 377 (2011) 942–955.

[2] L. P. Rowland, N. A. Shneider, Amyotrophic lateral sclerosis, New England Journal of Medicine 344 (2001) 1688–1700.

[3] M. Goldenberg, Multiple sclerosis review, P & T: a peer-reviewed journal for formulary management 37 (2012) 175–84.

[4] M.-H. Soriani, C. Desnuelle, Care management in amyotrophic lateral sclerosis, Revue Neurologique 173 (2017) 288–299.

[5] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. D. Nunzio, P. Fariselli, J. M. G. Dominguez, M. Gromicho, A. Guazzo, E. Longato, S. C. 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), CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[6] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. D. Nunzio, P. Fariselli, J. M. G. Dominguez, M. Gromicho, A. Guazzo, E. Longato, S. C. 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 9th to 12th, 2024, Lecture Notes in Computer Science, Springer, 2024.

[7] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function, Journal of the Neurological Sciences 169 (1999) 13–21.

[8] S. Twork, S. Wiesmeth, M. Spindler, M. Wirtz, S. Schipper, D. Pöhlau, J. Klewer, J. Kugler, Disability status and quality of life in multiple sclerosis: non-linearity of the expanded disability status scale (edss), Health and Quality of Life Outcomes 8 (2010).

[9] E. Tavazzi, E. Longato, M. Vettoretti, H. Aidos, I. Trescato, C. Roversi, A. S. Martins, E. N. Castanho, R. Branco, D. F. Soares, A. Guazzo, G. Birolo, D. Pala, P. Bosoni, A. Chiò, U. Manera, M. de Carvalho, B. Miranda, M. Gromicho, I. Alves, R. Bellazzi, A. Dagliati, P. Fariselli, S. C. Madeira, B. Di Camillo, Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review, Artificial Intelligence in Medicine 142 (2023) 102588.

[10] F. Papaiz, M. E. T. Dourado, R. A. d. M. Valentim, A. H. F. de Morais, J. P. Arrais, Machine learning solutions applied to amyotrophic lateral sclerosis prognosis: A review, Frontiers in Computer Science 4 (2022).

[11] T. Hothorn, H. H. Jung, Randomforest4life: A random forest for predicting als disease progression, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 15 (2014) 444–452. PMID: 25141076.

[12] K. D. Ko, T. El-Ghazawi, D. Kim, H. Morizono, Predicting the severity of motor neuron disease progression using electronic health record data with a cloud computing big data approach, in: 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, 2014, pp. 1–6.

[13] D. Halbersberg, B. Lerner, Temporal modeling of deterioration patterns and clustering for disease prediction of als patients, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 2019, pp. 62–68.

[14] A. A. Taylor, C. Fournier, M. Polak, L. Wang, N. Zach, M. Keymer, J. D. Glass, D. L. Ennist, T. P. R. O.-A. A. C. T. Consortium, Predicting disease progression in amyotrophic lateral sclerosis, Annals of Clinical and Translational Neurology 3 (2016) 866–875.

[15] R. Kueffner, et Al., Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach, Scientific Reports 9 (2019).

[16] A. Greco, M. R. Chiesa, I. Da Prato, A. M. Romanelli, C. Dolciotti, G. Cavallini, S. M. Masciandaro, E. P. Scilingo, R. Del Carratore, P. Bongioanni, Using blood data for the differential diagnosis and prognosis of motor neuron diseases: a new dataset for machine learning applications, Scientific Reports 11 (2021).

[17] M. F. a. Roberto Gomeni, Amyotrophic lateral sclerosis disease progression model, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 15 (2014) 119–129. PMID: 24070404.

[18] S. Pires, M. Gromicho, S. Pinto, M. Carvalho, S. C. Madeira, Predicting non-invasive ventilation in als patients using stratified disease progression groups, in: 2018 IEEE International Conference on

Data Mining Workshops (ICDMW), 2018, pp. 748–757.

[19] A. S. Martins, M. Gromicho, S. Pinto, M. de Carvalho, S. C. Madeira, Learning prognostic models using disease progression patterns: Predicting the need for non-invasive ventilation in amyotrophic lateral sclerosis, IEEE/ACM Transactions on Computational Biology and Bioinformatics 19 (2022) 2572–2583.

[20] S. Pires, M. Gromicho, S. Pinto, M. de Carvalho, S. C. Madeira, Patient stratification using clinical and patient profiles: Targeting personalized prognostic prediction in als, in: I. Rojas, O. Valenzuela, F. Rojas, L. J. Herrera, F. Ortuño (Eds.), Bioinformatics and Biomedical Engineering, Springer International Publishing, Cham, 2020, pp. 529–541.

[21] H. K. van der Burgh, R. Schmidt, H.-J. Westeneng, M. A. de Reus, L. H. van den Berg, M. P. van den Heuvel, Deep learning predictions of survival based on mri in amyotrophic lateral sclerosis, NeuroImage: Clinical 13 (2017) 361–369.

[22] B. Hadad, B. Lerner, Domain adaptation from clinical trials data to the tertiary care clinic – application to als, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2020.

[23] E. Tavazzi, et Al., Predicting functional impairment trajectories in amyotrophic lateral sclerosis: a probabilistic, multifactorial model of disease progression, Journal of Neurology 269 (2022) 3858–3878.

[24] K. Chalkou, E. Steyerberg, M. Egger, A. Manca, F. Pellegrini, G. Salanti, A two-stage prediction model for heterogeneous effects of treatments, Statistics in Medicine 40 (2021) 4362–4375.

[25] Y. Ahuja, N. Kim, L. Liang, T. Cai, K. Dahal, T. Seyok, C. Lin, S. Finan, K. Liao, G. Savovoa, T. Chitnis, T. Cai, Z. Xia, Leveraging electronic health records data to predict multiple sclerosis disease activity, Annals of Clinical and Translational Neurology 8 (2021) 800–810.

[26] R. Schlaeger, M. D'Souza, C. Schindler, L. Grize, S. Dellas, E. Radue, L. Kappos, P. Fuhr, Prediction of long-term disability in multiple sclerosis, Multiple Sclerosis Journal 18 (2012) 31–38.

[27] M. Filippi, P. Preziosa, M. Copetti, G. Riccitelli, M. A. Horsfield, V. Martinelli, G. Comi, M. A. Rocca, Gray matter damage predicts the accumulation of disability 13 years later in ms, Neurology 81 (2013) 1759–1767.

[28] V. Popescu, F. Agosta, H. E. Hulst, I. C. Sluimer, D. L. Knol, M. P. Sormani, C. Enzinger, S. Ropele, J. Alonso, J. Sastre-Garriga, A. Rovira, X. Montalban, B. Bodini, O. Ciccarelli, Z. Khaleeli, D. T. Chard, L. Matthews, J. Palace, A. Giorgio, N. De Stefano, P. Eisele, A. Gass, C. H. Polman, B. M. J. Uitdehaag, M. J. Messina, G. Comi, M. Filippi, F. Barkhof, H. Vrenken, MAGNIMS Study Group, Brain atrophy and lesion load predict long term disability in multiple sclerosis, J. Neurol. Neurosurg. Psychiatry 84 (2013) 1082–1091.

[29] R. Schlaeger, M. D'Souza, C. Schindler, L. Grize, S. Dellas, E. W. Radue, L. Kappos, P. Fuhr, Prediction of long-term disability in multiple sclerosis, Mult. Scler. 18 (2012) 31–38.

[30] Y. Zhao, B. C. Healy, D. Rotstein, C. R. G. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, T. Chitnis, Exploration of machine learning techniques in predicting multiple sclerosis disease course, PLOS ONE 12 (2017) e0174866.

[31] F. S. Brown, S. A. Glasmacher, P. K. A. Kearns, N. MacDougall, D. Hunt, P. Connick, S. Chandran, Systematic review of prediction models in relapsing remitting multiple sclerosis, PLoS One 15 (2020) e0233575.

[32] 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, J. D. Berry, Wearable device and smartphone data quantify als progression and may provide novel outcome measures, npj Digital Medicine 6 (2023).

[33] V. Fuh-Ngwa, Y. Zhou, J. C. Charlesworth, A.-L. Ponsonby, S. Simpson-Yap, J. Lechner-Scott, B. V. Taylor, A. I. Group, Developing a clinical–environmental–genotypic prognostic index for relapsing-onset multiple sclerosis and clinically isolated syndrome, Brain Communications 3 (2021) fcab288.

[34] I. Trescato, A. Guazzo, E. Longato, E. Hazizaj, C. Roversi, E. Tavazzi, M. Vettoretti, B. Di Camillo, Baseline machine learning approaches to predict amyotrophic lateral sclerosis disease progression notebook for the idpp lab on intelligent disease progression prediction at clef 2022, 2022.

[35] A. Guazzo, I. Trescato, E. Longato, E. Tavazzi, M. Vettoretti, B. Camillo, Baseline machine learning approaches to predict multiple sclerosis disease progression, in: CLEF, 2023.

[36] S. Van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, Journal of statistical software 45 (2011) 1–67.

[37] I. M. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[38] T. O. Hodson, Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not, Geoscientific Model Development 15 (2022) 5481–5487.

[39] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[40] J. Roux, D. Bard, E. Le Pabic, C. Segala, J. Reis, J. C. Ongagna, J. ze, E. Leray, Air pollution by particulate matter PM10 may trigger multiple sclerosis relapses., Environ Res 156 (2017) 404–410.