ALSFRS-R Score Prediction for Amyotrophic Lateral Sclerosis

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

Guido Barducci[1], Flavio Sartori[1], Giovanni Birolo[1], Tiziana Sanavia[1] and Piero Fariselli[1]

[1]*Computational Biomedicine Unit, Dept. of Medical Sciences, University of Turin, Turin, Italy*

Abstract

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disorder that results in the gradual deterioration of motor abilities, leading to challenges in breathing, speaking, swallowing, and ultimately death, typically occurring within a few years. The symptoms of ALS can vary significantly from one individual to another, affecting various bodily functions and areas. To assess this wide range of symptoms, the Amyotrophic Lateral Sclerosis Functional Rating Scale - Revised (ALSFRS-R) is utilized. Predicting the ALSFRS-R score is clinically relevant for personalizing patient monitoring. To address this need, the Intelligent Disease Progression Prediction challenge was organized, tasking participants with developing novel methods to predict these scores using non-invasive sensor data that monitor some individual characteristics. The competition included two tasks that differed only in the way the ALSFRS-R questionnaires were completed: either by medical staff (task 1) or by the patient (task 2). Given the limited number of patients on the training set, it was decided to use a relatively simple model, Random Forest, and to preselect sensor features by retaining those most correlated with the outcome to be predicted. We selected the model with the lowest MAE estimated by cross-validation on the challenge training set. The competition results demonstrate that our method attained on the test set an average Mean Absolute Error (MAE) of 0.234 and 0.311, along with a Root Mean Square Error (RMSE) of 0.519 and 0.601 for tasks 1 and 2, respectively. Although the error may appear very low, this is because questionnaire values tend to remain constant from one visit to another, thus facilitating prediction.

Keywords

Machine Learning, ALS, ALSFRS-R

1. Introduction

Amyotrophic lateral sclerosis (ALS), also known as neuropathy, is a rapidly progressive and ultimately fatal neurological disease that affects the neurons controlling voluntary muscles in the arms, legs, and face. The yearly incidence of ALS is around 1 to 2.6 cases per $100,000$ individuals, while the prevalence is approximately 6 cases per $100,000$. ALS belongs to a group of motor neuron disorders and typically results in death. Previous studies report approximately 48% and 24% survival rates at 3 and 5 years respectively, with around 4% surviving beyond 10 years. However, population-based studies show lower 5-year survival rates, ranging from 4% to 30% [1] [2].

The symptoms of this disease can vary greatly from case to case and can affect different functions and areas of the body. To describe this wide range of symptoms, the Amyotrophic Lateral Sclerosis Functional Rating Scale - Revised (ALSFRS-R) is employed. It consists of a 12-item inventory, with each item rated on a $0–4$ scale by patients and/or caregivers, resulting in a maximum score of 48 points. ALSFRS-R assesses patients' levels of self-sufficiency in areas including feeding, grooming, ambulation, and communication [1] [3] .

Given the variability of this disease, monitoring checks should vary depending on its characteristics, such as the progression rate. Currently, there is no system capable of predicting the course of the disease, making it very challenging to personalize patient visits based on disease progression. The goal

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

guido.barducci@unito.it (G. Barducci); flavio.sartori@unito.it (F. Sartori); giovanni.birolo@unito.it (G. Birolo); tiziana.sanavia@unito.it (T. Sanavia); piero.fariselli@unito.it (P. Fariselli)

0009-0005-1052-8495 (G. Barducci); 0009-0004-3833-6551 (F. Sartori); 0000-0003-0160-9312 (G. Birolo); 0000-0003-3288-0631 (T. Sanavia); 0000-0003-1811-4762 (P. Fariselli)

of this paper is to utilize information collected through the sensors of a commercial fitness smartwatch, past ALSFRS-R scores, and static features (such as age, sex, etc.) to predict future ALSFRS-R scores. The questionnaire data available has two different sources: they can be filled out by a doctor or by the patient through the use of a dedicated smartphone application. Therefore, the challenge has been divided into two tasks with the same goal but using data from different sources characterized by different frequencies of intervals between one questionnaire and the next, as well as differing medical or personal opinions, which may lead to different scoring choices despite similar symptoms. To solve these tasks, classical machine learning models were used instead of deep learning given the small number of patients in the training dataset. For more details we refer the reader to the challenge overview papers [4, 5].

The paper is divided into the following sections: 2 Related Work, which reports some papers addressing topics similar to this work; 3 Methodology, where the entire procedure that led to the predictions for the two tasks is outlined; 4 Experimental Setup, which details the procedures used; 5 Results, where the obtained results along with performance metrics are presented; and 6 Conclusions and Future Work, which reviews the essential steps of the paper and proposes alternative methodologies that could be useful for improving predictions.

2. Related Work

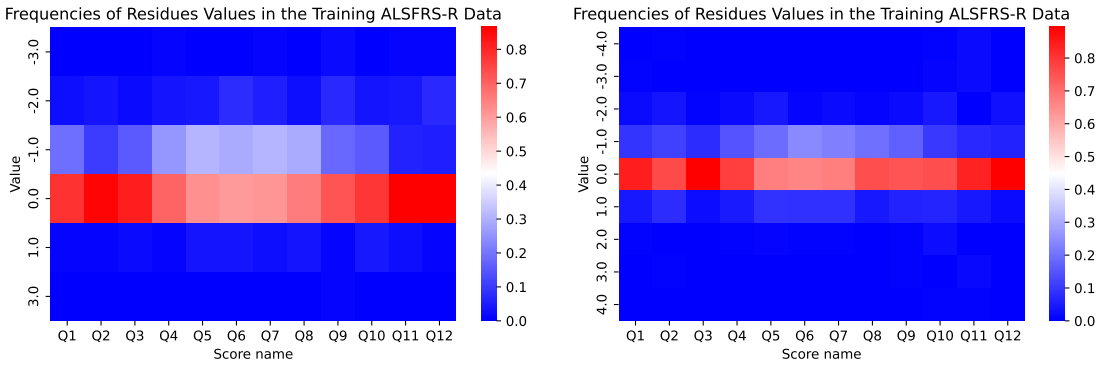
The quest to identify prognostic factors and build predictive models for amyotrophic lateral sclerosis (ALS) progression has been a longstanding challenge, but one of paramount importance. ALS exhibits significant variability in its progression and outcomes, posing obstacles to making accurate predictions. Many methodologies have undergone rigorous testing using data from the PRO-ACT database. While this repository may not perfectly capture the full spectrum of ALS patients in the population, it stands as the largest publicly accessible dataset amalgamating ALS clinical trials [6] [7] [8].

Wearable devices have been effectively used to study individuals with ALS, demonstrating a link between ALS progression and behavior and function patterns in people with amyotrophic lateral sclerosis, as measured by digital wearables [9] [10] [11] [12]. These measurements include total activity volume, active versus sedentary time, and time spent at home. Additionally, wearable devices are increasingly utilized to investigate physical activity in populations with cardiovascular disease, multiple sclerosis, arthritis, and other conditions. Although studies have been conducted to predict characteristics related to the ALSFRS score, such as its score and slope, to date, there are no predictors leveraging data from smartwatches to predict the ALSFRS score [13] [8]. Hence, it is imperative to investigate such data types extensively to ascertain if they can enhance diagnostic predictions.

3. Methodology

Three different types of data were used to predict ALSFRS-R scores: sensor data from Garmin VivoActive 4 smartwatches, static feature data, and ALSFRS-R questionnaire data. Regarding sensor data, these consist of 90 different features per day and are characterized by a large number of missing values (in many days, no features were recorded, rendering the sensor vector absent for those days) due to both the data collection device and patient behavior. The static data are baseline characteristics recorded at a specific time, they include: sex, diagnostic delay, age at diagnosis, forced vital capacity (FVC), weight, and body mass index (BMI). Finally, the ALSFRS-R data are of the same type as those to be predicted but collected at a previous time.

To leverage these challenging data, two different approaches have been explored: the Mono Window approach and the Double Window approach. The Mono Window approach is the simpler of the two: for each prediction, only the sensors recorded within 7 days prior to the questionnaire to be predicted are used (these can be utilized in various ways, such as averaging or taking the median). The second approach involves considering two sensor data windows instead of one: the first window adjacent to the questionnaire to be predicted, and the second adjacent to the previous available questionnaire. The idea behind this second method is to provide the model with more information about the changes



(a) *Task 1* (b) *Task 2*

Figure 1: Frequencies of residue values in the training ALSFRS-R data for tasks 1 and 2. As you can see, the most frequent value is 0 for each task. Consequently, the majority of questionnaire values remain unchanged compared to the previous ones.

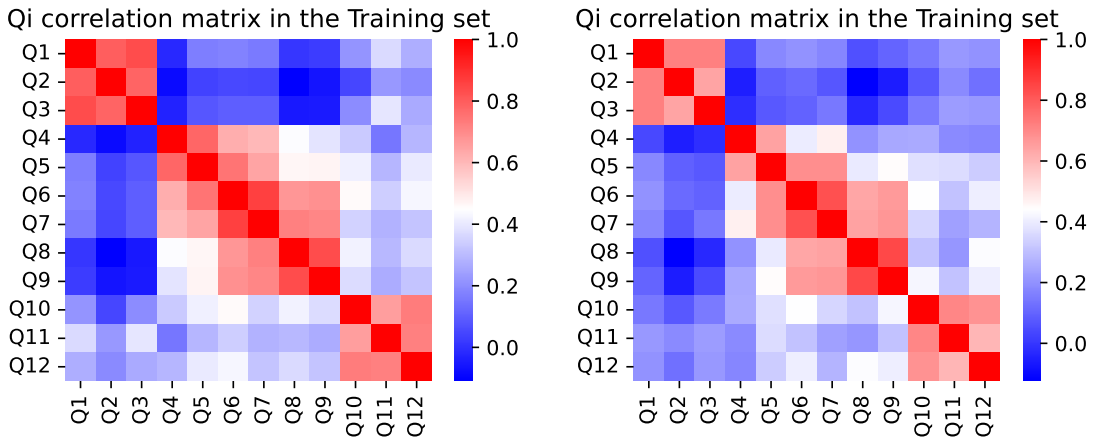
in recorded parameters over time. Despite the large amount of sensor data (13.946 feature vectors) and the fact that the second method seems more natural for handling this type of data, it is heavily penalized by the irregularity of the sensors. Indeed, two 7-day windows with at least 3 days of sensor data were not available in 20 out of 54 patients. For this reason, the choice to use the first approach was mandatory. In this process, for each patient, all sensor data outside the temporal window of the 7 days preceding the ALSFRS-R values to be predicted was disregarded. The sensor data within each time window were averaged along the time to obtain one feature vector of length 90 for each window that is less affected by daily variability. Once the feature vectors representing the sensor data were obtained, they were concatenated with the vectors of static features and with the vectors of previous ALSFRS-R data before the questionnaires to be predicted; by doing this the the final feature vectors were obtained.

Regarding the outcomes to predict, these are the values of ALSFRS-R questionnaires after the last ones available for the training. To solve this task, it has been observed that the questionnaire values tend to remain constant between visits (see Figure 1); therefore, it was decided to use the previous time's questionnaire score as the prediction baseline and to fit the model on the residuals. To obtain the final prediction value, it was sufficient to add the predicted residual to the value of the previous questionnaire relative to the one to be predicted[1]. The Random Forest Classifier was chosen as the model; unlike deep learning models, it does not require large datasets for training, making it suitable for this task. Before being used to fit the model, the data were preprocessed by scaling and performing feature selection, as explained in the section 4.

4. Experimental Setup

The training dataset for this challenge comprises data from 52 different patients for both tasks. These data can be categorized into three types: static data, ALSFRS-R data, and data from a smartwatch sensor. By analyzing the correlation matrices, two important facts can be observed: as for the ALSFRS-R data, it can be observed that they can be divided into several correlated groups depending on the area affected by the disease (see Figure 2). Meanwhile, regarding the sensor data, there are strongly correlated features, as shown in the figure 3. As shown in the image, the correlated features are grouped into distinct blocks. For the cardiac features, two main blocks can be identified: the first, smaller block pertains to Heart Rate Variability (HRV), while the second block encompasses other cardiac characteristics related to the RR interval. For the respiratory features, two blocks are associated with respiration and blood oxygenation. Lastly, another block of correlated features pertains to the patient's steps. The grouping of features

[1]The samples in the training set corresponding to a residual value occurring fewer than 8 times have been discarded; indeed, they are too few to be recognized by the model.



(a) *Task 1* (b) *Task 2*

Figure 2: Correlation matrix of ALSFRS-R values in the training data for both task 1 and task 2. They reveal that the scores can be clustered into distinct groups. This indicates significant correlations among certain ALSFRS-R items, suggesting underlying patterns or relationships within the data.

into these blocks is expected, as they represent different statistics describing the same physiological processes.

The two tasks differ only in terms of the subject and the frequency with which the questionnaires are filled out. This makes the two tasks slightly different primarily for two reasons: the data from the first task should reasonably be more objective, as it is a clinician who fills out the questionnaires rather than the individual patient. Additionally, the data from the second task are compiled through an app and have therefore a higher and more irregular frequency, as depicted in Figure 4.

Regarding data preprocessing, the features were initially scaled using Min-Max Scaler and imputed with the mean or mode depending on their continuous or categorical nature. Concerning the sensor data, an additional process was added: the correlation between each of these features and the questionnaire to be predicted was calculated, and only those features with a correlation above a certain threshold were retained. This approach was chosen due to the low number of samples available to train the model compared to the total number of features. After the data preprocessing, they were used to train a Random Forest Classifier. To determine the optimal hyperparameters of the model[2], along with the correlation threshold utilized to select the sensor features[3], the following cross-validation strategy was employed. The training set of the challenge was divided into an inner training set and a validation set $(80\text{-}20\%)$. Hyperparameter optimization was conducted via cross-validation on the inner training set using a grid search method[4], and the chosen hyperparameters for each the tasks are displayed in Table 1 and Table 2.

5. Results

The challenge was divided into two tasks, Task 1 and Task 2, each involving the construction of a model to predict the values of ALSFRS-R questionnaires completed by either a clinician or the patient using a dedicated smartphone application. Despite testing two different macro types of methods: Mono Window and Double Window, the low number of patients in the training set resulted in significantly lower performance for the second type. Therefore, we only submitted results from the first type. These

[2]The Random Forest hyperparameters tested were maxing deep and max features; they have been tested respectively in ranges $[2, 9]$ and [sqrt, 'log2', None]. The number of trees was fixed at 300.
[3]The thresholds tested has been $[0, 0.05, 0.1, 0.15, 0.2, 0.25]$.
[4]During the fold creation process, care was taken to obtain stratified folds for outcomes to ensure partitions with similar percentages for each outcome category.

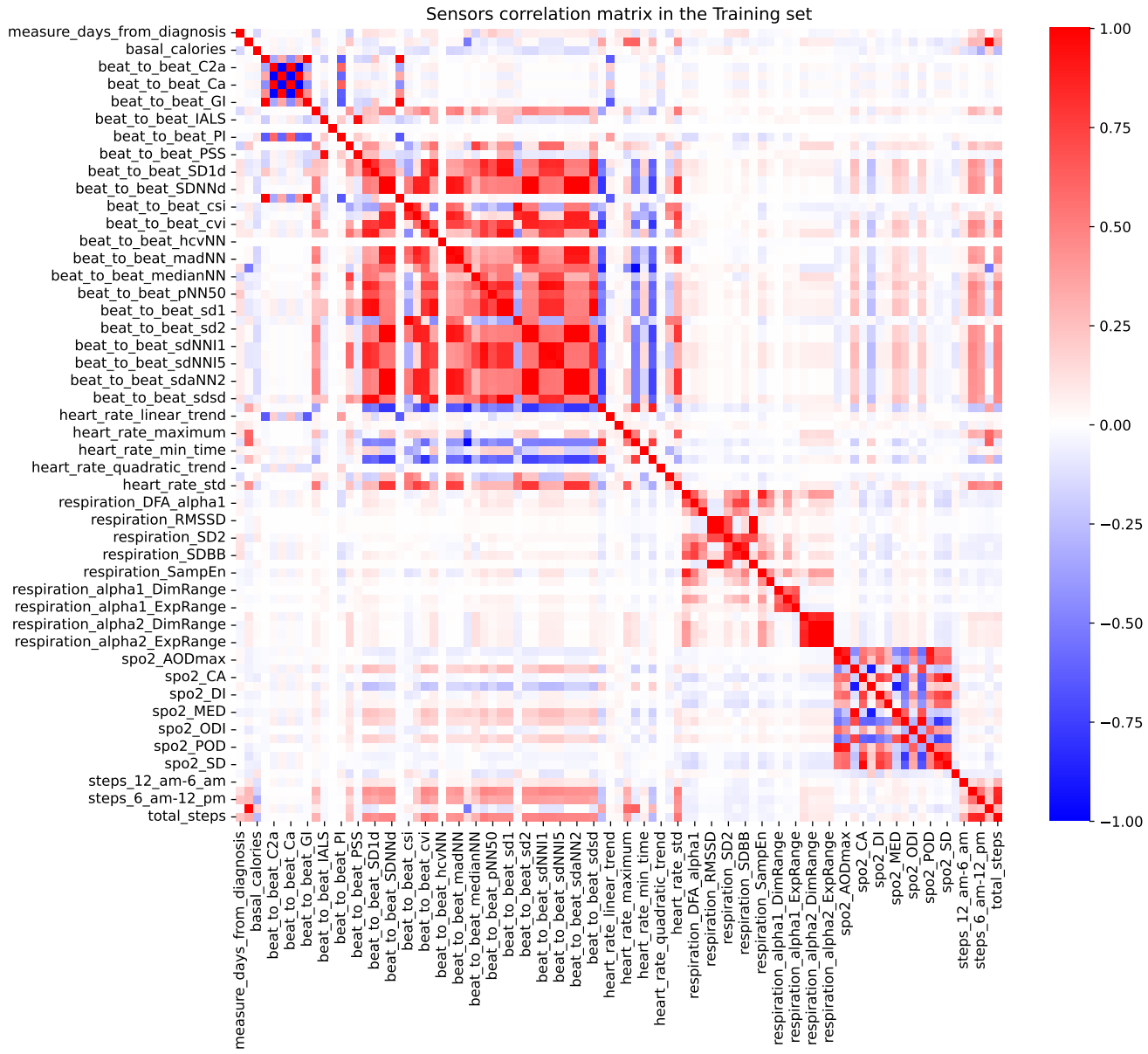
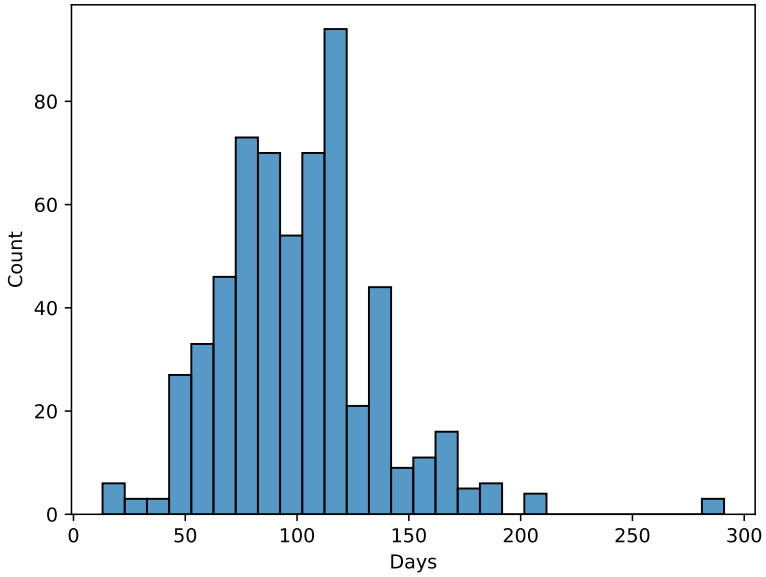


Figure 3: Sensors Correlation Matrix.

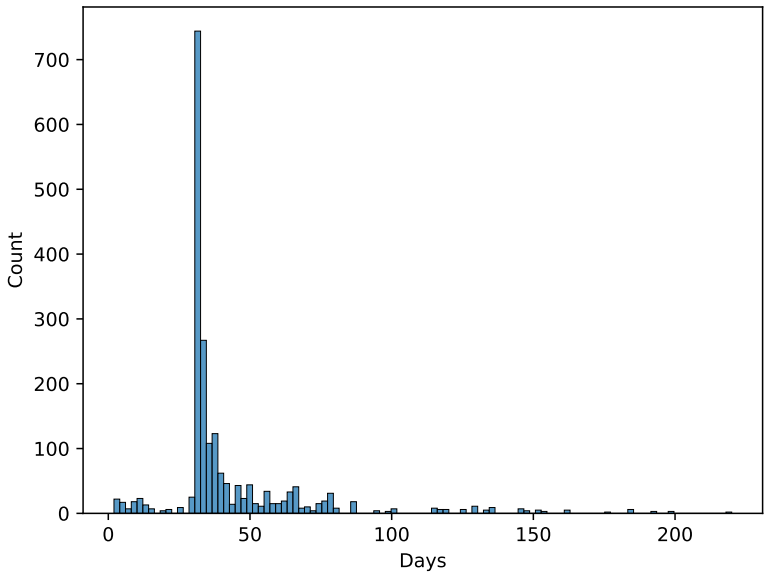
were obtained using a Random Forest Classifier, which was trained after determining the optimal hyperparameters through five-fold cross-validation; the average metrics for each ALSFRS-R question over the 5 validation folds are reported in Table 3 for the first task and in Table 4 for the second one.

6. Conclusions and Future Work

The models attempted to predict the ALSFRS-R questionnaire values were constrained by the small size of the training dataset. Despite experimenting with models utilizing various time intervals, the only ones proving useful for prediction were those solely relying on a window of sensor data adjacent to the questionnaire to be predicted, without leveraging information from more distant times. This limitation stems from the challenging nature of the data, which contains a large number of missing values. Among the models tested, the one demonstrating the best performance and subsequently used for submission was based on Random Forest, preceded by a feature selection step to reduce the number of sensor features.



(a) *Task 1*



(b) *Task 2*

Figure 4: Here's the histogram showing the difference in days between consecutive questionnaires for task 1 and task 2. As you can see, for task 2, the questionnaires are filled out at more irregular intervals and with greater frequency compared to task 1

The performance obtained shows significant variability depending on the questionnaire number; The Mean Absolute Error (MAE) calculated on the test set is 0.23 and 0.31 respectively for Task 1 and Task 2, while the Root Mean Squared Error (RMSE) is 0.52 and 0.60 respectively. The lower error is observed in the first task, which could be attributed to the fact that the questionnaire compilation by clinical staff tends to be more reliable and objective compared to the subjective opinion from the patient. These seemingly promising results are unfortunately attributed to the ALSFRS-R questionnaires mostly remaining constant from one visit to another, making it very easy to achieve high prediction performance.

To address this issue, one potential approach to improve is using data augmentation to increase the number of questionnaires in the training set. To improve predictions, methods of deep learning could be tested, leveraging much longer sequences of sensor data (such as recurrent neural networks). However,

Table 1
This table pertains to Task 1. It presents the hyperparameters associated with Random Forest along with the correlation threshold used for feature selection. This process involved removing sensor features with a correlation to the outcome lower than the specified threshold.

Q	depth	max features	correlation threshold
Q1	3	sqrt	0.2
Q2	2	sqrt	0.0
Q3	3	log2	0.0
Q4	8	sqrt	025
Q5	7	log2	0.15
Q6	6	sqrt	0.2
Q7	8	log2	0.1
Q8	9	sqrt	0.15
Q9	6	sqrt	0.15
Q10	2	log2	0.0
Q11	2	log2	0.0
Q12	3	log2	0.25

Table 2
This table pertains to Task 2. It presents the hyperparameters associated with Random Forest along with the correlation threshold used for feature selection. This process involved removing sensor features with a correlation to the outcome lower than the specified threshold.

Q	depth	max features	correlation threshold
Q1	4	sqrt	0.25
Q2	5	log2	0.0
Q3	3	sqrt	0.0
Q4	4	log2	0.0
Q5	7	sqrt	0.05
Q6	9	log2	0.0
Q7	7	sqrt	0.0
Q8	9	log2	0.25
Q9	8	sqrt	0.25
Q10	8	log2	0.25
Q11	3	sqrt	0.0
Q12	4	sqrt	0.15

these models require significantly more data for training, which represent the biggest obstacle for this task.

References

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

[2] E. O. Talbott, A. M. Malek, D. Lacomis, The epidemiology of amyotrophic lateral sclerosis, Handbook of clinical neurology 138 (2016) 225–238.

[3] 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.

[4] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, A. Guazzo, E. Longato, S. C. Madeira, U. Manera, S. Marchesin, L. Menotti, G. Silvello, E. Tavazzi, E. Tavazzi,

Table 3
Performance metrics on the task one validation set. They have been calculated by taking the mean across the five folds for each Q.

Q	MAE (std)	RMSE (std)
Q1	0.155 (0.080)	0.385 (0.0971)
Q2	0.091 (0.019)	0.298 (0.0317)
Q3	0.159 (0.028)	0.397 (0.034)
Q4	0.266 (0.106)	0.508 (0.107)
Q5	0.337 (0.112)	0.575 (0.096)
Q6	0.370 (0.045)	0.658 (0.063)
Q7	0.280 (0.054)	0.528 (0.051)
Q8	0.229 (0.071)	0.473 (0.080)
Q9	0.319 (0.076)	0.640 (0.098)
Q10	0.165 (0.024)	0.397 (0.030)
Q11	0.000 (0.000)	0.000 (0.000)
Q12	0.128 (0.045)	0.500 (0.089)

Table 4
Performance metrics on the task two validation set. They have been calculated by taking the mean across the five folds for each Q.

Q	MAE (std)	RMSE (std)
Q1	0.140 (0.036)	0.372 (0.050)
Q2	0.271 (0.050)	0.586 (0.062)
Q3	0.083 (0.016)	0.286 (0.029)
Q4	0.197 (0.045)	0.442 (0.050)
Q5	0.360 (0.059)	0.657 (0.040)
Q6	0.324 (0.070)	0.572 (0.062)
Q7	0.320 (0.056)	0.563 (0.051)
Q8	0.229 (0.066)	0.474(0.071)
Q9	0.226 (0.049)	0.473 (0.055)
Q10	0.250 (0.029)	0.577 (0.049)
Q11	0.105 (0.035)	0.319 (0.051)
Q12	0.059 (0.032)	0.237 (0.063)

I. Trescato, M. Vettoretti, B. Di 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, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[5] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García 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. Di 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.

[6] R. Kueffner, N. Zach, M. Bronfeld, R. Norel, N. Atassi, V. Balagurusamy, B. Di Camillo, A. Chio, M. Cudkowicz, D. Dillenberger, et al., Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach, Scientific reports 9 (2019) 690.

[7] M. Tang, C. Gao, S. A. Goutman, A. Kalinin, B. Mukherjee, Y. Guan, I. D. Dinov, Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering, Neuroinformatics 17 (2019) 407–421.

[8] C. Pancotti, G. Birolo, C. Rollo, T. Sanavia, B. Di Camillo, U. Manera, A. Chiò, P. Fariselli, Deep

learning methods to predict amyotrophic lateral sclerosis disease progression, Scientific reports 12 (2022) 13738.

[9] 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).

[10] L. Garcia-Gancedo, M. L. Kelly, A. Lavrov, J. Parr, R. Hart, R. Marsden, M. R. Turner, K. Talbot, T. Chiwera, C. E. Shaw, et al., Objectively monitoring amyotrophic lateral sclerosis patient symptoms during clinical trials with sensors: observational study, JMIR mHealth and uHealth 7 (2019) e13433.

[11] V. Ezeugwu, R. E. Klaren, E. A. Hubbard, P. T. Manns, R. W. Motl, Mobility disability and the pattern of accelerometer-derived sedentary and physical activity behaviors in people with multiple sclerosis, Preventive medicine reports 2 (2015) 241–246.

[12] 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.

[13] A. G. Karanevich, J. M. Statland, B. J. Gajewski, J. He, Using an onset-anchored bayesian hierarchical model to improve predictions for amyotrophic lateral sclerosis disease progression, BMC medical research methodology 18 (2018) 1–13.