Brainteaser

D9.9 Evaluation Challenge: Report on the analysis of the experimental results, proceedings, and integration with EOSC (48)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No GA101017598

Project Title	BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis
Grant Agreement No	GA101017598
Contract start date	01/01/2021
Contract duration	54 Months

Document ID	Deliverable_D9.9 Evaluation Challenge: Report on the analysis of the experimental results, proceedings, and integration with EOSC (48)
Deliverable leader	UNIPD
Due date	31/12/2024
Deliverable date	31/12/2024
Dissemination level	PUBLIC (PU)

AUTHORS – CONTRIBUTORS

Name	Organization
Giovanni Birolo	University of Turin, Turin, Italy
Pietro Bosoni	University of Pavia, Pavia, Italy
Guglielmo Faggioli	University of Padua, Padua, Italy
Helena Aidos	University of Lisbon, Lisbon, Portugal
Roberto Bergamaschi	University of Pavia, Pavia, Italy
Paola Cavalla	University of Turin, Turin, Italy “Città della Salute e della Scienza”, Turin, Italy
Adriano Chiò	University of Turin, Turin, Italy
Arianna Dagliati	University of Pavia, Pavia, Italy
Mamede de Carvalho	University of Lisbon, Lisbon, Portugal
Giorgio Maria Di Nunzio	University of Padua, Padua, Italy
Piero Fariselli	University of Turin, Turin, Italy
Jose Manuel García Dominguez	Gregorio Marañon Hospital in Madrid, Madrid, Spain
Sergio González	Universidad Politécnica de Madrid, Spain
Marta Gromicho	University of Lisbon, Lisbon, Portugal
Alessandro Guazzo	University of Padua, Padua, Italy
Enrico Longato	University of Padua, Padua, Italy
Sara C. Madeira	University of Lisbon, Lisbon, Portugal
Umberto Manera	University of Turin, Turin, Italy
Stefano Marchesin	University of Padua, Padua, Italy
Laura Menotti	University of Padua, Padua, Italy
Gianmaria Silvello	University of Padua, Padua, Italy
Eleonora Tavazzi	IRCCS Foundation C. Mondino in Pavia, Pavia, Italy
Erica Tavazzi	University of Padua, Padua, Italy
Isotta Trescato	University of Padua, Padua, Italy
Martina Vettoretti	University of Padua, Padua, Italy
Barbara Di Camillo	University of Padua, Padua, Italy
Nicola Ferro	University of Padua, Padua, Italy

PEER – REVIEWERS

Name	Organization
Giovanni Birolo	UNITO
Piero Fariselli	UNITO

DOCUMENT HISTORY

Version	Date	Author/Organization	Modifications	Status
0.1	20/09/2024	UNIPD	Initial outline	Draft
0.2	30/09/2024	UNIPD	First draft	Draft
0.3	15/10/2024	UNIPD	Second draft	Draft

0.4	30/10/2024	UNIPD	Formatting	Draft
0.5	15/11/2024	UNIPD	Revising	Pre-final
1.0	19/12/2024	UNIPD	Final draft	Final draft
2.0	24/12/2024	Maria F. Cabrera/UPM	Final review and final version	Final

Disclaimer

This deliverable may be subject to final acceptance by the European Commission. The information and views set out in this document are those of the authors and do not necessarily reflect the official opinion of the European Commission. Neither the Commission nor any person acting on the Commission's behalf may hold responsible for the use which may be made of the information contained therein.

Copyright message

Copyright message © BRAINTEASER Consortium, 2021-2024. This document contains original unpublished work or work to which the author/s holds all rights except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Multiple Sclerosis (MS) and Amyotrophic Lateral Sclerosis (ALS) are neurodegenerative diseases characterized by progressive or fluctuating impairments in motor, sensory, visual, and cognitive functions. Patients with these diseases endure significant physical, psychological, and economic burdens due to hospitalizations and home care while grappling with uncertainty about their conditions.

AI tools hold promise for aiding patients and clinicians by identifying the need for intervention and suggesting personalized therapies throughout disease progression.

The objective of iDPP@CLEF is to develop AI-based approaches to describe the progression of these diseases. The ultimate goal is to enable patient stratification and predict disease progression, thereby assisting clinicians in providing timely care.

iDPP@CLEF 2024 continues the work of the previous editions, iDPP@CLEF 2022 and 2023. The 2022 edition focused on predicting ALS progression and utilizing explainable AI. The 2023 edition expanded on this by including environmental data and introduced a new task for predicting MS progression. This edition extends the MS dataset with environmental data and introduces two new ALS tasks aimed at predicting disease progression using data from wearable devices. This marks the first iDPP edition to utilize prospective data directly collected from patients involved in the BRAINTEASER project.

1 INTRODUCTION

Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) are two severe and impactful diseases that cause progressive neurological impairment. The progression of these diseases is typically heterogeneous, resulting in significant variability in aspects such as treatment, outcomes, quality of life, and overall patient needs. This variability presents challenges not only for patients but also for clinicians and caregivers.

For example, patients with ALS often need specific treatments like Non-Invasive Ventilation (NIV) or Percutaneous Endoscopic Gastrostomy (PEG) at certain stages of their disease progression. Similarly, MS patients may experience debilitating relapses that severely impact their quality of life. Therefore, it would be highly beneficial to anticipate the needs of individuals affected by these diseases to provide them with the most timely and effective care. However, the heterogeneous nature of these conditions makes it challenging to develop effective prognostic tools that work the same and are effective for every patient.

This underscores the importance of creating automatic tools to assist clinicians in decision-making throughout disease progression, facilitating personalized therapeutic choices. In particular, developing new automatic predictive approaches based on Artificial Intelligence (AI) requires a proper framework for designing and evaluating different tasks, such as:

- Stratifying patients according to their phenotype throughout disease evolution.
- Predicting disease progression in a probabilistic, time-dependent manner.
- Providing a better and more explainable understanding of the mechanisms underlying MS and ALS.

A key aspect is that these approaches should rely on shared resources that enable proper benchmarking, comparable, and reproducible experimentation. In fact, only by properly measuring and comparing the effectiveness of the various developed tools we can understand how to improve them. The Intelligent Disease Progression Prediction at CLEF (iDPP) Lab aims to provide an evaluation infrastructure for developing such AI algorithms. iDPP proposes to go beyond the current state of the art by systematically addressing issues related to applying AI in clinical practice for ALS and MS. In addition to defining risk scores based on the probability of short- or long-term events, iDPP also focuses on providing clinicians with structured and understandable data.

iDPP 2024 is the final iteration of an evaluation cycle begun in 2022, comprising three challenges aimed at fostering reproducible and comparable evaluation of AI-based approaches for predicting the progression of ALS and MS. The first edition, iDPP 2022, focused exclusively on ALS, challenging participants to predict the probability that patients would need specific medical treatments based on their medical history. The second edition, iDPP 2023, not only built upon iDPP 2022 by extending its dataset with environmental data to determine the impact of the environment on patient needs, but it also introduced a new task to predict the risk of disability increases for patients living with MS.

This final edition, iDPP 2024, further extends the 2023 dataset by including environmental data for MS patients to investigate the impact of pollution and external environmental factors on MS progression. Additionally, two new tasks have been introduced: predicting

the progression of ALS, measured by the ALSFRS-R scale, based on the patient's clinical history and data obtained from wearable devices and sensors.

The paper is organized as follows: section 2 presents related challenges; section 3 describes its tasks; section 4 discusses the developed dataset; section 5 explains the setup of the Lab and introduces the participants; section 6 introduces the evaluation measures adopted to score the runs; section 7 analyzes the experimental results for the different tasks; section 8 outlines the actual iDPP@CLEF workshop; finally, section 9 draws some conclusions and outlooks some future work.

This is an extended version of the condensed overview for the iDPP 2024 Lab (Birolo et al., 2024a)

2 RELATED CHALLENGES

There have been no other Labs on this or similar topics within CLEF before the start of iDPP. iDPP 2022 and 2023 were the first two iterations of the Lab and the current, 2024, is the third.

While no major challenges – besides iDPP 2023 – regarding MS have been carried out yet, more interest has been shown toward ALS. In particular, three major challenges were organized on this topic: the DREAM 7 ALS Prediction challenge[1] in 2012 and the DREAM ALS Stratification challenge[2] in 2015 and a Kaggle challenge[3] in 2021. The DREAM 7 ALS Prediction challenge consisted of using 3 months of ALS clinical trial information (months 0–3) to predict the future progression of the disease (months 3–12), expressed as the slope of change in ALSFRS-R (Cedarbaum et al., 1999). Later on, the DREAM ALS Stratification challenge (Küffner et al., 2015) required participants to stratify ALS into subgroups based on their characteristics, to understand patient profiles better and provide personalized ALS treatments. Finally, the Kaggle challenge employed clinical and genomic data to obtain a better understanding of the mechanisms underlying ALS and determine why some people with ALS tend to have a faster progression of the disease compared to others.

At the current time, most of the datasets used to evaluate AI algorithms for MS are based on closed and proprietary datasets. In this sense iDPP paved the way for a reproducible and effectively open science in the research domain of the AI used for predicting the progression of MS.

2.1 iDPP 2022

iDPP 2022[4] (Guazzo et al., 2022a, 2022b) was the first edition of the Lab and concerned exclusively the ALS disease progression prediction. Being the pilot Lab, a large share of effort was devoted to understanding the challenges and limitations linked to the shared evaluation campaigns, when it comes to AI applied in the medical domain. iDPP 2022 was organized into 3 tasks:

- Pilot Task 1 - Ranking Risk of Impairment: The focus of the first task of iDPP 2022 was on ranking patients based on the risk of impairment, defined as the need for specific medical treatments, such as NIV, PEG, or death. Participants were given information on the motor functioning of the patients, measured according to the ALSFRS-R scale (Cedarbaum et al., 1999), in time and were asked to rank patients based on the time-to-event risk of experiencing impairment in each specific domain.
- Pilot Task 2 - Predicting Time of Impairment: it refined Task 1 by asking participants to predict when specific impairments will occur (i.e. in the correct time window). In this regard, the task focused on assessing model calibration in terms of the ability

[1] https://dreamchallenges.org/dream-7-phil-bowen-als-prediction-prize4life/

[2] https://dx.doi.org/10.7303/syn2873386

[3] https://www.kaggle.com/alsgroup/end-als

[4] https://brainteaser.health/open-evaluation-challenges/idpp-2022/

of the proposed algorithms to estimate the probability of an event close to the true probability within a specified time window.

- Position Paper Task 3 - Explainability of Artificial Intelligence algorithms: The task focused on the evaluation and discussion of AI-based explainable frameworks for intelligent disease progression prediction able to explain the multivariate nature of the data and the model predictions.

One of the major outputs of iDPP 2022 was the 3 datasets released. In particular, the datasets contain data for the prediction of specific events related to ALS. Such datasets are fully anonymized retrospective details about 2250 real patients. The patients were recruited from two medical institutions in Turin, Italy, and Lisbon, Portugal. The datasets contain static data about patients (e.g. age, onset date, gender) and event data (i.e. 18,512 ALSFRS-R questionnaires and 4,015 spirometries). 6 groups participated in iDPP 2022 and submitted a total of 120 runs.

2.2 iDPP 2023

Similarly to iDPP 2022, iDPP 2023[5] (Faggioli et al., 2023a, 2023b) was also organized into three tasks, focusing on either ALS or MS. More in detail, Tasks 1 and 2 of iDPP 2023 concerned MS, while Task 3 built upon iDPP 2022 and extended the ALS tasks of the previous iteration of the Lab. To summarize iDPP 2023 tasks:

- Task 1: Predicting Risk of Disease Worsening (MS) This task focused on predicting the probability that, given the history of the patient, they would undergo a worsening, according to two different definitions of worsening.
- Task 2: Predicting Cumulative Probability of Worsening (MS) The second task had a similar objective to task 1, with the major difference that, instead of predicting the risk at an absolute level, participants were required to predict the cumulative probability of worsening over 10 years.
- Task 3: Position Papers on the Impact of Exposition to Pollutants (ALS) The third task extended the first task of iDPP 2022 and concerned the ranking of the patients based on the risk of impairment. The major difference to iDPP 2022 was that participants were given environmental data to determine if such data was a good predictor of the risk of impairment.

iDPP 2023 extended the iDPP 2022 datasets with three 2 datasets for MS. In particular, such datasets contained static data about patients, MS-related details (e.g., the EDSS score, results of MRIs, evoked potentials measures), and a label indicating if the patient underwent a worsening, based on the worsening definitions of Task 1 and 2. 10 teams submitted a total of 163 runs at the end of iDPP 2023.

[5] https://brainteaser.dei.unipd.it/challenges/idpp2023/

3 TASKS

In the remainder of this section, we describe each task in more detail.

3.1 Task 1: Predicting ALSFRS-R Score from Sensor Data (ALS)

Task 1 focuses on predicting the twelve scores of the ALSFRS-R (ALS Functional Rating Scale - Revised), assigned by medical doctors roughly every three months, from the sensor data collected via the app. The ALSFRS-R is a somehow “subjective” evaluation usually performed by a medical doctor and this task will help in answering a currently open question in the research community, i.e. whether it could be derived from objective factors.

Participants were given the ALSFRS-R questionnaire at the first visit with the scores for each question together with the time (number of days from diagnosis) at which the questionnaire was taken. Participants will be given the time of the second visit (number of days from diagnosis) together with all the sensor data up to the time of the second visit.

Participants had to predict the values of the ALSFRS-R sub-scores at the second visit.

3.2 Task 2: Predicting Patient Self-assessment Score from Sensor Data (ALS)

The second task concerning ALS focuses on predicting the self-assessment score assigned by patients from the sensor data collected via the app. Self-assessment scores correspond to each of the ALSFRS-R scores but, while the latter ones are assigned by medical doctors during visits, these scores are assigned via auto-evaluation by patients themselves using the provided app.

If the self-assessment performed by patients, more frequently than the assessment performed by medical doctors every three months or so, can be reliably predicted by sensor and app data, we can imagine a proactive application which, monitoring the sensor data, alerts the patient if an assessment is needed.

Participants were given the first set of self-assessed scores together with the time (number of days from diagnosis) at which the questionnaire was taken. Participants were also given the time of the second auto-evaluation (number of days from diagnosis) together with all the sensor data up to the time of the second auto-evaluation. Participants had to predict the values of the self-assessed scores at the second auto-evaluation, happening one or two months after the first one.

3.3 Task 3: Predicting Relapses from EDSS Sub-scores and Environmental Data (MS)

The third task focuses on predicting a relapse using environmental data and EDSS (Expanded Disability Status Scale) sub-scores. This task allows us to assess if exposure to different pollutants is a useful variable in predicting a relapse.

Participants were asked to predict the week of the first relapse after the baseline considering environmental data based on a weekly granularity, given the status of the patient at the baseline, which is the first visit available in the considered time span (after January 1, 2013). For each patient, the date of the baseline will be week 0 and all the other weeks will be relative to it.

Participants were given all the environmental data about a patient, i.e. also observations which may happen after the relapse to be predicted. All the patients are guaranteed to experience, at least, one relapse after the baseline.

4 DATASET

For iDPP 2024 we release three datasets: two completely new datasets for ALS and an extension of the iDPP 2023 dataset concerning MS. More in detail, the two new ALS datasets comprise a common training part with 52 training patients, whose ALSFRS-R scores were both annotated by the clinicians and self-assessed. Concerning the test sets, 21 and 11 patients were included in them for Task 1 and Task 2, respectively. Regarding MS, the part of the dataset concerning static variables and MS-related information is the same as the one used for iDPP 2023. The major improvement regards environmental data that have been added to the dataset.

4.1 Tasks 1 and 2: ASL Dataset with Clinical or self-assessed ALSFRS-R

The datasets for Task 1 and Task 2 were collected from ALS-diagnosed patients recruited during the BRAINTEASER project from three centers in Lisbon, Madrid, and Turin. At recruitment, patients were given a commercial fitness tracker (the Garmin VivoActive 4 smartwatch), and data from its sensors was collected during a follow-up period with a median duration of 270 days. Patients were encouraged to wear the watch as much as they were comfortable with, ideally all the time, both while awake and sleeping. Each day of data for each patient was summarized into a vector of 90 statistics related to heart rate and beat-to-beat interval, respiration rate, and nocturnal pulse oximetry. Sensor data was not available every day for each patient.

During the same period, disease progression was assessed by their clinician using the ALSFRS-R questionnaire (roughly every three months, following standard clinical practice). Patients also used the same questionnaire to self-assess their progression through a smartphone app developed specifically by the BRAINTEASER project, the Patient App. They were prompted for the assessment once per month, though the actual frequency varied and depended on patient compliance.

4.1.1 Creation of the datasets

Patients with insufficient data were excluded from the challenge dataset. Specifically, this included those with less than three months of follow-up data, those with more than 50% of sensor data missing, and those without at least two clinical or self-assessed ALSFRS-R evaluations. After applying these criteria, a dataset of 83 patients was obtained, with a median of 254 days of sensor data per patient. These patients and their data were then divided into a training group (common to both Tasks 1 and 2) and two task-specific testing groups.

4.1.2 Split into training and test

The patients were split into three groups:

- training: patients with at least two clinical and two self-assessed ALSFRS-R evaluations;
- test-ct: patients with at least two clinical but without two self-assessed ALSFRS-R evaluations;

- test-app: patients with at least two self-assessed but without two clinical ALSFRS-R evaluations.

The training set thus included 52 patients with a median of 3.5 clinical and 5 self-assessed ALSFRS-R evaluations (189 and 301 in total, respectively). The test-ct set (the test set for Task 1) included 21 patients, whose first clinical ALSFRS-R evaluations were included as features and the second evaluations were the prediction target. The test-app set (the test set for Task 2) included 11 patients and was built in the same way using the self-assessed ALSFRS-R evaluations. The full available sensor data for all patients was included in both the training and test datasets, while only the clinical (resp. self-assessed) ALSFRS-R evaluations were included for Task 1 (resp. Task 2). A comparative description of the datasets is shown in Table 1.

Table 1. Comparison between training and test populations for Task 1 and 2. Continuous variables are presented as median (interquartile range); categorical variables as count (percentage on available data), for each level. "Sensor adherence" is the ratio of days with available sensor data during the whole sensor follow-up.

Variable	Level	Task 1/2 Train	Task 1 test	Task 2 test
Sex	Female	11 (21.15%)	9 (42.86%)	4 (36.36%)
	Male	41 (78.85%)	12 (57.14%)	7 (63.64%)
1-5 Diagnostic delay (months)	median (IQR)	0.8 (0.4-1.3)	0.9 (0.4-1.8)	1.0 (0.4-1.6)
1-5 Age at diagnosis	median (IQR)	56 (49-64)	62 (57-66)	60 (52-66)
1-5 FVC	median (IQR)	85 (79-95)	84 (79-98)	92 (79-113)
1-5 Weight	median (IQR)	75 (64-81)	67 (60-71)	65 (60-70)
1-5 BMI	median (IQR)	25 (23-27)	24 (22-26)	22 (21-25)
1-5 ALSFR-R CT (count)	median (IQR)	3.5 (2.0-5.0)	-	-
1-5 ALSFR-R APP (count)	median (IQR)	5.0 (3.0-8.0)	-	-
1-5 Sensor follow-up (months)	median (IQR)	9.8 (5.2-13.6)	8.9 (5.3-14.2)	5.9 (5.5-8.3)
1-5 Sensor adherence	median (IQR)	98% (89%-100%)	98% (85%-100%)	100% (99%-100%)

4.2 Task 3: MS Dataset

The dataset used for Task 3 in iDPP@CLEF 2024 is structured similarly to those from iDPP@CLEF 2023, though some features (e.g., evoked potentials, MRIs) were not included, and certain records have been filtered based on the purpose of the task.

4.2.1 Updates over IDPP@CLEF 2023

In the 2024 dataset, EDSS data before January 1, 2013 (aligned with the start of environmental data collection) were filtered, and patients without EDSS follow-ups were removed. Additionally, patients who did not experience a relapse after their first non-filtered EDSS follow-up (i.e., the baseline for each patient) were excluded.

The dataset has been expanded to incorporate environmental data, which includes information on patients' exposure to various air pollutants identified as significant public health risks in the latest World Health Organization (WHO) global air quality guidelines (World Health Organization, 2021), such as particulate matter (PM) - encompassing both $PM_{2.5}$ (particles with an aerodynamic diameter of 2.5 micrometers or less) and PM_{10} (particles with an aerodynamic diameter of 10 micrometers or less) - as well as ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and several weather factors (including wind speed, relative humidity, sea level pressure, global radiation, precipitation, and average, minimum, and maximum temperatures).

Air pollutant data from public monitoring stations were collected daily from the European Air Quality Portal using the DiscoMap tool[6]. The geographical coordinates (longitude and latitude) of each monitoring station were matched to specific postcodes, identifying the nearest station to each patient's residence postcode. Instead, weather data were gathered daily from the European Climate Assessment and Dataset station network, which provides access to the E-OBS dataset, a daily gridded land-only observational dataset over Europe[7]. Each grid was matched with the nearest monitoring station using Euclidean distance based on geographical coordinates. This approach ensured that air pollution and weather data were aligned with the same spatial and temporal granularity. Daily environmental measurements were aggregated into weekly averages from each patient's baseline. As additional features, the number of days per week spent over the respective WHO recommended air quality guideline levels for short-term (24 hours) exposure was computed for each air pollutant (World Health Organization, 2021).

Finally, a subset of 380 MS patients from the Turin and Pavia research centers was selected for Task 3 in iDPP@CLEF 2024, compared to 550 patients for Task 1 and 638 for Task 2 in iDPP@CLEF 2023. The resulting MS dataset[8] includes static variables with demographic and clinical information, EDSS scores with corresponding Functional System (FS) sub-scores, environmental measurements, and the outcome time, representing the week of the first relapse occurrence after the baseline for each patient. EDSS follow-ups are reported between the baseline and the outcome time, while environmental measurements span from January 1, 2013, to December 30, 2023. It is important to note that environmental data may have gaps due to availability. When

[6] https://discomap.eea.europa.eu/Index

[7] https://www.ecad.eu/download/ensembles/download.php

[8] https://brainteaser.dei.unipd.it/challenges/idpp2024/assets/other/ms/ms-variables-description.txt

considering only environmental data preceding the outcome time, the median number of weeks available for each patient is 59, with an interquartile range of 103.25 weeks. The distributions of air pollutant concentrations (measured in micrograms per cubic meter), averaged across patients over these weeks, are depicted in the boxplots of Figure 1, where the red stars indicate the WHO recommended air quality guideline levels for 24-hour exposure (World Health Organization, 2021).

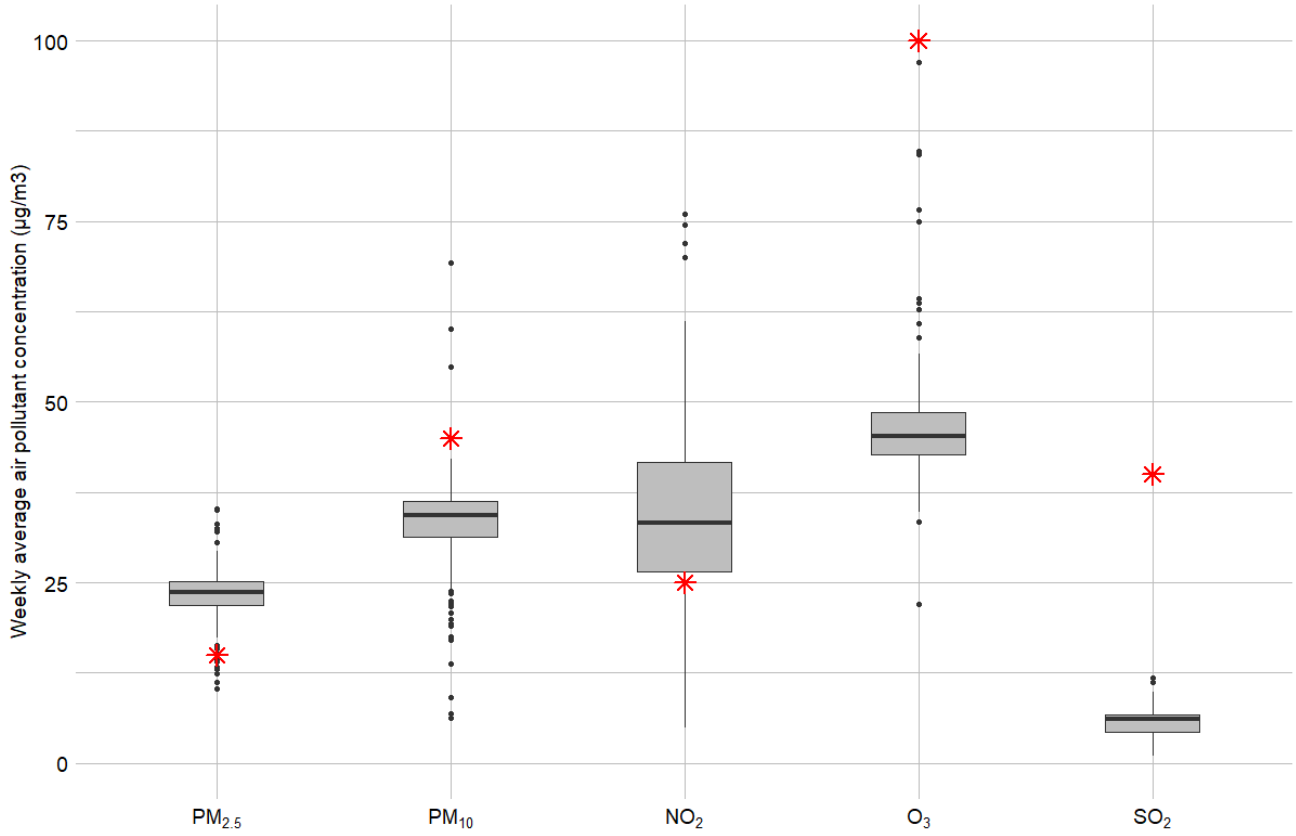

Figure 1. Boxplots of weekly average air pollutant concentrations across patients. Red stars represent the World Health Organization (WHO) recommended air quality guideline levels for 24-hour exposure.

4.2.2 Split into training and test

The dataset was split into a training set (70%) and a test set (30%), with subjects stratified by outcome time to ensure an even distribution across both sets. The distribution of static data, including demographic and clinical information, and EDSS were verified to be similar in both training and test sets. Additionally, since environmental exposure is considered, the distribution of patients from the two clinical centres and their residence classification (cities, rural areas, and towns) was checked to be balanced.

Statistical tests, including the Kruskal-Wallis test for continuous variables and the Chi-squared test for categorical and ordinal variables, were performed to assess the appropriateness of the stratification. Special attention was given to sparsely observed levels in categorical variables to ensure rare levels appeared only in the training set if at all. Table 2 provides a comparison of variable distributions between the training and test sets, confirming that the split meets the best-practice quality standards.

Table 2. Comparison between training and test populations for MS task. Continuous variables are presented as median (interquartile range); categorical variables as count (percentage on available data), for each level.

Variable	Level	Levels Training	Levels Test
Sex	Female	148 (74.37%)	54 (66.67%)
	Male	51 (25.63%)	27 (33.33%)
Ethnicity	Caucasian	181 (90.96%)	77 (95.06%)
	Hispanic	2 (1.00%)	-
	Black African	2 (1.00%)	-
	NA	14 (7.04%)	4 (4.94%)
Residence classification	Cities	53 (26.63%)	20 (24.69%)
	Rural Area	52 (26.13%)	22 (27.16%)
	Towns	94 (47.24%)	39 (48.15%)
Centre	Pavia	129 (64.82%)	58 (71.61%)
	Turin	70 (35.18%)	23 (28.39%)
Occurrence of MS in pediatric age	FALSE	176 (88.44%)	77 (95.06%)
	TRUE	23 (11.56%)	4 (4.94%)
Age at onset	median (IQR)	28 (22-36)	30 (24-34)
Age at baseline	median (IQR)	38 (31-47)	38 (33-47)
Diagnostic delay	median (IQR)	12 (4-47)	12 (3-28)
Spinal cord symptom	FALSE	143 (71.86%)	54 (66.67%)
	TRUE	56 (28.14%)	27 (33.33%)
Brainstem symptom	FALSE	146 (73.37%)	57 (70.37%)
	TRUE	53 (26.63%)	24 (29.63%)
Eye symptom	FALSE	148 (74.37%)	59 (72.84%)
	TRUE	51 (25.63%)	22 (27.16%)
Supratentorial symptom	FALSE	140 (70.35%)	50 (61.73%)
	TRUE	59 (29.65%)	31 (38.27%)
Other symptoms	FALSE	197 (99.00%)	80 (98.77%)
	Sensory	1 (0.50%)	1 (1.23%)

Variable	Level	Levels Training	Levels Test
	Epilepsy	1 (0.50%)	-
EDSS	median (IQR)	2.0 (1.5-3.0)	2.0 (1.5-3.5)
	NA	3 (0.36%)	0 (0.00%)
Outcome time	median (IQR)	59 (24-122)	53 (25-130)

5 LAB SETUP AND PARTICIPATION

In the remainder of this section, we detail the guidelines the participants had to comply with to submit their runs and the submissions received by iDPP.

5.1 Guidelines

Participating teams should satisfy the following guidelines:

- The runs should be submitted in the textual format described below;
- Each group can submit a maximum of 30 runs for each of Task 1 and Task 2 and Task 3.

5.1.1 Task 1 Run Format

Runs should be submitted as a text file (.txt) with the following format:

```
10061925618906738677 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10160033396142711519 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10287479530859953248 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
12398828804459792214 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10038199677222038201 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
...
```

where:

- Columns are separated by a white space;
- The first column is the patient ID, a hashed version of the original patient ID (should be considered just as a string);
- Columns from 2 to 13 represent the predicted ALSFRS-R sub-score. Each column corresponds to an ALSFRS-R question, e.g. column 2 to Q1, column 3 to Q2, and so on). Each value is expected to be integer in the range [0, 4];
- The last column is the run identifier, according to the format described below. It must uniquely identify the participating team and the submitted run.

It is important to include all the columns and have a white space delimiter between the columns. No specific ordering is expected among patients (rows) in the submission file.

5.1.2 Task 2 Run Format

Runs should be submitted as a text file (.txt) with the following format:

```
10061925618906738677 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10160033396142711519 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10287479530859953248 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
12398828804459792214 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
10038199677222038201 1 2 3 4 1 2 3 4 1 2 3 4 upd_T1_myDesc
...
```

where:

- Columns are separated by a white space;
- The first column is the patient ID, a hashed version of the original patient ID (should be considered just as a string);
- Columns from 2 to 13 represent the predicted self-assessed sub-score. Each column corresponds to an ALSFRS-R question, e.g. column 2 to Q1, column 3 to Q2, and so on). Each value is expected to be integer in the range [0, 4];
- The last column is the run identifier, according to the format described below. It must uniquely identify the participating team and the submitted run.

It is important to include all the columns and have a white space delimiter between the columns. No specific ordering is expected among patients (rows) in the submission file.

5.1.3 Task 3 Run Format

Runs should be submitted as a text file (.txt) with the following format:

```
10061925618906738677    10    upd_T3_myDesc
10160033396142711519    47    upd_T3_myDesc
10287479530859953248    13    upd_T3_myDesc
12398828804459792214    1     upd_T3_myDesc
10038199677222038201    9     upd_T3_myDesc
...
```

where:

- Columns are separated by a white space;
- The first column is the patient ID, a hashed version of the original patient ID (should be considered just as a string);
- The second column is the predicted week at which the first relapse after the baseline happens. The value is expected to be an integer starting from 1;
- The third column is the run identifier, according to the format described below. It must uniquely identify the participating team and the submitted run.

It is important to include all the columns and have a white space delimiter between the columns. No specific ordering is expected among patients (rows) in the submission file.

5.1.4 Submission Upload

Runs should be uploaded to the repository provided by the organizers. Following the repository structure discussed above, for example, a run submitted for the first task should be included in submission/task1.

Runs should be uploaded using the following name convention for their identifiers: <teamname>_T<1|2|3>_<freefield>, where:

- teamname is the name of the participating team;
- T<1|2|3> is the identifier of the task the run is submitted to, e.g. T1 for Task 1;
- freefield is a free field that participants can use as they prefer to further distinguish among their runs. Please, keep it short and informative.

For example, a complete run identifier may look like upd_T1_myDesc, where:

- upd is the University of Padua team;
- T1 means that the run is submitted for Task 1;
- myDesc suggests an appropriate description for the run.

The name of the text file containing the run must be the identifier of the run followed by the txt extension. In the above example upd_T1_myDesc.txt

5.2 Participants

A total of 28 teams registered to iDPP 2024, out of which eight teams were able to submit one run in at least one task. Table 3 reports the details about teams that managed to submit at least one run. Furthermore, Table 3 outlines in which tasks each team participated in and how many runs they were able to submit. In total, 97 runs were submitted to iDPP 2024. The most participated task was Task 1 with 59 runs and 6 teams participating. Subsequently, Task 2 had 31 runs submitted by six different teams. Finally, only two teams participated in Task 3, with a total of 7 runs submitted. The most prolific participant was UNIPD, with a total of 20 runs.

Table 3. Number of runs submitted by each participant team in iDPP 2024.

	Task 1 (ALS)	Task 2 (ALS)	Task 3 (MS)	Total
BIT.UA	7	7	—	14
CompBiomedUniTO	1	1	—	2
FCOOL	9	9	—	18
iDPPExplorers	15	—	—	15
Mandatory	19	—	—	19
Stefagroup	—	—	3	3
UBCS	—	6	—	6
UNIPD	8	8	4	20
Total	59	31	7	97

6 EVALUATION MEASURES

In both Tasks 1 and 2, the prediction targets were the future scores of the ALSFRS-R evaluation, which are integers in the [0-4] range. Since the scores are discrete, we could have framed the predictive task as a classification problem. However, we opted for a regression problem to be able to penalize larger errors more (e.g., with a target score of 3, predicting 1 should be worse than predicting 2). Task 3, where the target was the week of the relapse, was also framed quite naturally as a regression task for similar reasons. Thus, we evaluated all tasks using the same two state-of-the-art evaluation measures to assess the performance of regression models: the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). The formulas for RMSE and MAE are shown in the Equations below respectively, where n represents the number of observations, y_i is the actual value of the dependent variable for the i-th observation, and $\hat{y}_i$ is the predicted value of the dependent variable for the i-th observation.

Both metrics can explain the performance of a model in an interpretable manner since their units are the same as the target variable (e.g., weeks); together, they can provide a comprehensive evaluation of the three prediction tasks, with smaller values indicating better results. The RMSE measures how much, on average, the model's predictions deviate from the actual values. By squaring the errors before averaging them, RMSE gives higher weight to large errors. MAE represents the average absolute difference between actual and predicted values. Unlike RMSE, MAE grows linearly with the error magnitude. Therefore, it provides a clear representation of the average error, is less sensitive to outliers, and does not emphasize large errors as much as RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Both metrics can explain the performance of a model in an interpretable manner since their units are the same as the target variable (e.g., weeks); together, they can provide a comprehensive evaluation of the three prediction tasks.

The RMSE measures how much, on average, the model's predictions deviate from the actual values. This statistical index ranges from 0 to ∞, with smaller values indicating better simulation results. By squaring the errors before averaging them, RMSE gives higher weight to large errors. MAE represents the average absolute difference between actual and predicted values. Unlike RMSE, MAE treats all errors equally, regardless of their magnitude. Therefore, it provides a clear representation of the average error, is less sensitive to outliers, but does not emphasize large errors as much as RMSE.

7 RESULTS

For each task, we report the analysis of the performance of the runs submitted by the Lab's participants according to the measures described in Section 6

7.1 Task 1: Predicting ALSFRS-R Score from Sensor Data (ALS)

Clinicians monitor ALS progression through frequent visits, typically every two to three months, to promptly detect any worsening of symptoms. Consequently, ALSFRS-R scores usually remain fairly stable between these appointments, making the most recent score a reliable predictor for the next assessment. While some deterioration in at least one score is not uncommon, using the last observed value as a predictive measure is both simple and effective, as most scores will not change. This approach, which we will call "naive" since it does not use sensor data, is particularly useful for bulbar and respiratory scores, which show more stability in the challenge dataset, and where sensor data might not be as effective in detecting eventual changes. The distribution of ALSFRS-R scores and the amount of worsening between consecutive visits in the training set is shown in Figure 2 and Figure 3.

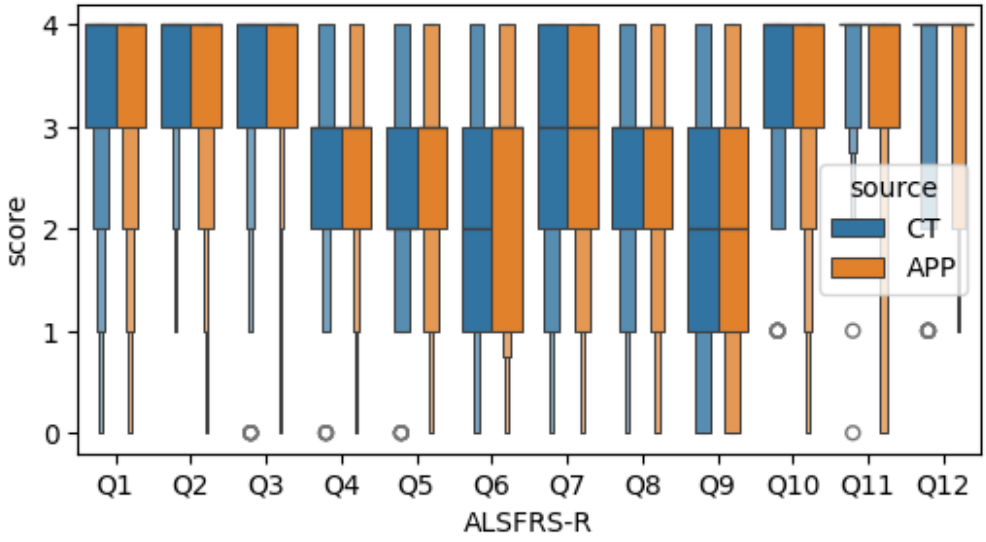

Figure 2. Distribution of worsening between two consecutive ALSFRS-R evaluations in the training set, for each score.

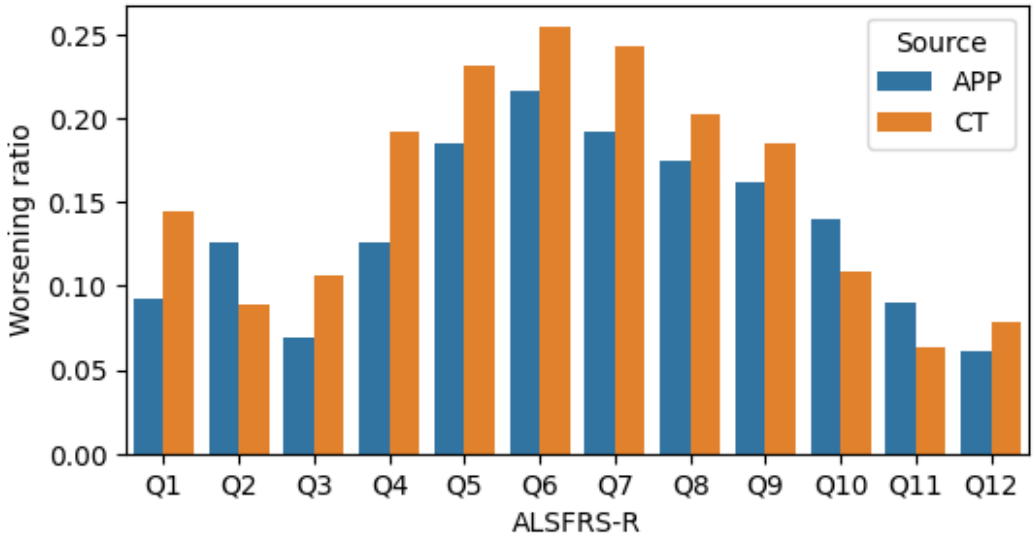

Figure 3. Average ratio of worsening between two consecutive ALSFRS-R evaluations in the training set, for each score.

Four teams—iDPPExplorers, Mandatory, FCOOL, and UNIPD—employed this strategy in one of their runs for Task 1, achieving the lowest errors with both metrics (0.20 MAE and 0.49 RMSE) and securing joint first place. The full error scores and rankings for all submitted runs are reported in Table 4.

Note that other runs, which also utilize sensor data, demonstrate performance very close to the first place. Due to the small size of the test set, error estimates exhibit large standard deviations, making it impossible to assert significant differences in the top scores.

The rankings are obtained considering the average of the performance for all twelve ALSFRS-R scores and show how the naive predictors that propagate the last observed score are globally optimal. However, this is not the case for each single ALSFRS-R score, where other runs often have lower errors, as can be seen in Figure 3. Again, given the small size of the test sets, these differences in performance are not statistically significant. However, it is also reasonable that the data collected by the sensors can be more helpful in prediction for some scores than others: this is especially evident for Q9 and Q11 in Task 1 and for Q4 and Q12 in Task 2.

Minimum MAE reached by naive runs versus non-naive runs for each score in Task 1 and 2. Naive runs are those that use the last observed values as their prediction.

7.2 Task 2: Predicting Patient Self-assessment Score from Sensor Data (ALS)

Task 2 is very similar to Task 1, with several teams employing the same methods as they did for Task 1. However, in Task 2, the ALSFRS-R assessments by patients are less regular in timing and less consistent in scoring compared to assessments by clinicians, although they are generally more closely spaced.

The predict-the-last-scores approach remains the top performer, albeit with slightly higher errors (0.29 MAE and 0.58 RMSE), placing the UNIPD and FCOOL teams in joint first place again. Full results are reported in Table 5.

Table 4. MAE and RMSE for Task 1 runs. For both MAE and RMSE, results are reported as the average error across all twelve ALSFRS-R scores, the average standard deviation (computed by bootstrapping the test set one thousand times) and their respective ranking.

	metric	MAE	RMSE
team	run		
fcool	locf	0.20±0.20 (#1)	0.49±0.20 (#1)
idppexplorers	naive	0.20±0.22 (#1)	0.49±0.22 (#1)
unipd	hold	0.20±0.21 (#1)	0.49±0.21 (#1)
mandatory	d1	0.20±0.19 (#1)	0.49±0.19 (#1)
idppexplorers	EN	0.22±0.17 (#2)	0.50±0.17 (#2)
CBMUnito	RF-MonoWindow	0.23±0.19 (#3)	0.52±0.19 (#3)
bitua	ensemble-max	0.25±0.18 (#4)	0.54±0.18 (#4)
	temporalAnalysis	0.29±0.24 (#5)	0.61±0.24 (#6)
unipd	average	0.33±0.18 (#6)	0.60±0.18 (#5)
	logistic-ALSFRS	0.34±0.21 (#7)	0.64±0.21 (#8)
fcool	RFClassifier	0.35±0.22 (#8)	0.68±0.22 (#15)
unipd	rf	0.36±0.22 (#9)	0.65±0.22 (#11)
idppexplorers	voting	0.37±0.15 (#10)	0.65±0.15 (#10)
bitua	moremetrics	0.37±0.23 (#10)	0.68±0.23 (#16)
mandatory	12hist14	0.37±0.19 (#11)	0.65±0.19 (#9)
unipd	rf-reg	0.37±0.19 (#12)	0.64±0.19 (#7)
mandatory	1hist09	0.38±0.31 (#13)	0.72±0.31 (#30)
bitua	median	0.38±0.23 (#14)	0.70±0.23 (#20)
fcool	2nd-best-both-metrics	0.39±0.26 (#15)	0.71±0.26 (#25)
bitua	mean	0.39±0.26 (#15)	0.71±0.26 (#21)
mandatory	1hist05	0.39±0.20 (#16)	0.66±0.20 (#12)
unipd	ridge	0.39±0.20 (#17)	0.69±0.20 (#17)
idppexplorers	gb	0.40±0.18 (#18)	0.69±0.18 (#18)
mandatory	1hist04	0.40±0.26 (#18)	0.66±0.26 (#13)
	12hist10	0.41±0.23 (#19)	0.67±0.23 (#14)
unipd	optrun	0.41±0.19 (#20)	0.71±0.19 (#22)
idppexplorers	svm	0.41±0.23 (#20)	0.75±0.23 (#33)
fcool	best-both-metrics	0.41±0.22 (#20)	0.71±0.22 (#26)
mandatory	12hist13	0.42±0.24 (#21)	0.72±0.24 (#28)
bitua	ensemble-avg	0.42±0.23 (#22)	0.71±0.23 (#24)
idppexplorers	lr	0.42±0.20 (#23)	0.73±0.20 (#32)

	metric	MAE	RMSE
mandatory	1hist03	0.42±0.24 (#24)	0.69±0.24 (#19)
	12hist11	0.43±0.28 (#25)	0.72±0.28 (#27)
fcool	3rd-best-both-metrics	0.43±0.26 (#25)	0.78±0.26 (#39)
mandatory	d0	0.44±0.14 (#26)	0.72±0.14 (#29)
	1hist08	0.44±0.26 (#27)	0.71±0.26 (#23)
idppexplorers	et	0.44±0.24 (#27)	0.78±0.24 (#36)
	dt	0.44±0.22 (#28)	0.72±0.22 (#31)
	knn	0.46±0.19 (#29)	0.77±0.19 (#35)
bitua	ensemble-min	0.47±0.30 (#30)	0.80±0.30 (#40)
idppexplorers	bestModels	0.47±0.24 (#31)	0.81±0.24 (#42)
	lstm	0.48±0.27 (#32)	0.82±0.27 (#43)
mandatory	1hist07	0.48±0.21 (#33)	0.75±0.21 (#34)
	1hist02	0.48±0.32 (#34)	0.78±0.32 (#37)
idppexplorers	nn	0.49±0.24 (#35)	0.80±0.24 (#41)
mandatory	1hist06	0.49±0.29 (#36)	0.78±0.29 (#38)
idppexplorers	rf	0.51±0.29 (#37)	0.86±0.29 (#47)
fcool	LogisticRegression	0.51±0.28 (#38)	0.84±0.28 (#46)
idppexplorers	bagging	0.51±0.35 (#39)	0.89±0.35 (#49)
unipd	logistic	0.51±0.27 (#40)	0.83±0.27 (#45)
fcool	SVC	0.54±0.34 (#41)	0.89±0.34 (#48)
	XGBClassifier	0.57±0.15 (#42)	0.83±0.15 (#44)
	majority-class	0.66±0.52 (#43)	1.09±0.52 (#50)

Table 5. MAE and RMSE for Task 2 runs. For both MAE and RMSE, results are reported as the average error across all twelve ALSFRS-R scores, the average standard deviation (computed by bootstrapping the test set one thousand times) and the respective ranking.

	metric	MAE	RMSE
team	run		
fcool	locf	0.29±0.15 (#1)	0.58±0.15 (#1)
1-4 unipd	hold	0.29±0.15 (#1)	0.58±0.15 (#1)
1-4 CBMUnito	RF-MonoWindow	0.31±0.16 (#2)	0.60±0.16 (#2)
1-4	ensemble-max	0.33±0.14 (#3)	0.61±0.14 (#3)
	moremetrics	0.37±0.17 (#4)	0.65±0.17 (#4)
	mean	0.39±0.18 (#5)	0.71±0.18 (#8)
	median	0.40±0.21 (#6)	0.69±0.21 (#5)
1-4 fcool	2nd-best-both-metrics	0.41±0.15 (#7)	0.71±0.15 (#6)
1-4	ensemble-avg	0.42±0.22 (#8)	0.71±0.22 (#7)
	idpp2024-bitua	0.43±0.24 (#9)	0.72±0.24 (#9)
1-4 unipd	average	0.49±0.20 (#10)	0.78±0.20 (#12)

	metric	MAE	RMSE
1-4 fcool	3rd-best-both-metrics	0.50±0.13 (#11)	0.78±0.13 (#10)
1-4 unipd	logistic-ALSFRS	0.50±0.19 (#11)	0.85±0.19 (#18)
1-4 bitua	ensemble-min	0.50±0.24 (#12)	0.82±0.24 (#14)
1-4	rf	0.52±0.20 (#13)	0.78±0.20 (#11)
	rf-reg	0.52±0.12 (#14)	0.82±0.12 (#13)
1-4	best-both-metrics	0.53±0.20 (#15)	0.84±0.20 (#15)
	RFClassifier	0.53±0.24 (#16)	0.85±0.24 (#17)
1-4 unipd	ridge	0.55±0.27 (#17)	0.85±0.27 (#16)
1-4	LogisticRegression	0.57±0.21 (#18)	0.89±0.21 (#19)
	XGBClassifier	0.59±0.17 (#19)	0.93±0.17 (#20)
1-4	optrun	0.61±0.27 (#20)	0.96±0.27 (#21)
	logistic	0.66±0.29 (#21)	0.99±0.29 (#22)
1-4 fcool	SVC	0.67±0.19 (#22)	1.01±0.19 (#23)
1-4	features100	0.82±0.43 (#23)	1.20±0.43 (#26)
	featuresall	0.89±0.41 (#24)	1.25±0.41 (#27)
	features10	0.94±0.49 (#25)	1.33±0.49 (#28)
	features25	0.96±0.21 (#26)	1.14±0.21 (#24)
	features20	1.02±0.24 (#27)	1.18±0.24 (#25)
1-4 fcool	majority-class	1.03±0.44 (#28)	1.47±0.44 (#29)
1-4 ubcs	features50	1.11±0.51 (#29)	1.51±0.51 (#30)

7.3 Task 3: Predicting Relapses from EDSS Sub-scores and Environmental Data (MS)

Table 6 displays the RMSE and MAE scores for all submissions made for Task 3, with consistent scoring positions across both metrics. Additionally, the scatter plot in Figure 4 offers a visual representation of the performance of all submitted runs, where the x-axis denotes actual values and the y-axis represents predicted values. Ideally, perfect predictions would result in points aligning along a straight line with a slope of 1.

The top-performing strategy is associated with the UNIPV UNIPV_t3_rf run (Bosoni et al., 2024) which employs a Random Forest (RF) model after thorough preprocessing stages. Regarding the adoption of environmental features, it is notable that all submissions from the UNIPV (Stefagroup) team incorporate environmental variables for relapse predictions. In contrast, the UNIPD team offers both methods with and without the inclusion of environmental variables, achieving their best results with the UNIPD UNIPD_t3_ridge_noenv run, which excludes environmental variables (Marinello et al., 2024).

Table 6. MAE and RMSE results (with the respective rankings) for all the submitted runs for Task 3.

Team	Run	MAE	RMSE	
Stefagroup	UNIPV_t3_rf	22.49 (#1)	41.52 (#1)	
	UNIPV_t3_lmer_first	28.05 (#2)	48.07 (#2)	
	UNIPV_t3_lmer_last	47.74 (#3)	72.51 (#3)	
UNIPD	UNIPD_t3_ridge_noenv	61.37 (#4)	78.62 (#4)	
	UNIPD_t3_average	65.80 (#5)	79.26 (#5)	
	UNIPD_t3_rf_reg	66.63 (#6)	79.74 (#6)	
	UNIPD_t3_ridge	68.59 (#7)	89.84 (#7)	

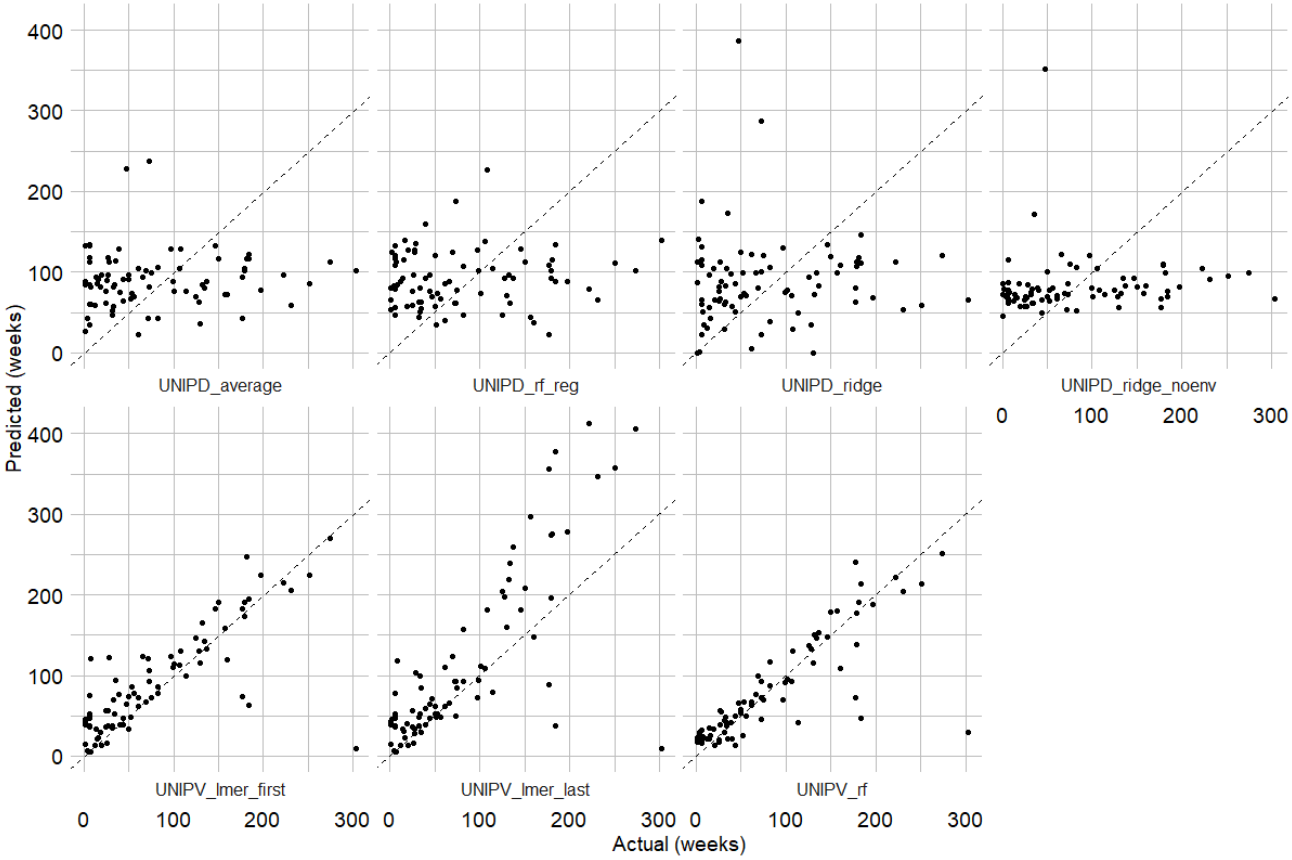

Figure 4. Actual versus Predicted values for each run submitted for Task 3.

7.4 Approaches

In this section, we provide a short summary of the approaches adopted by participants in iDPP. There are two separate sub-sections, one for Task 1 and 2 – focused on ALS progression prediction – and one for Task 3 – which concerns the MS relapse prediction, using environmental data.

7.4.1 Tasks 1 and 2

Silva & Oliveira (2024) (Team BIT.UA) focus on Tasks 1 and 2. Their proposed approaches employ machine learning techniques that rely on RF ensembles. They observed that the most effective solutions are based on temporal analysis, with the maximization strategy being the top-performing approach. Additionally, they emphasize the importance of proper handling of missing data. The authors noted inconsistent performance across the two tasks. Specifically, their approaches tended to be more effective on Task 1, while performance on Task 2 was less satisfactory. Silva and Oliveira (2024) attribute this behavior to the variability of the underlying data: Task 1 data, produced by clinicians, was more stable, whereas Task 2 data, produced directly by patients, appeared to be less stable.

Barducci et al. (2024) (Team CompBiomedUniTO) tested different approaches to preselect the sensor features to be fed to a RF Classifier. The first solution exploits the mono window approach, which keeps only sensor data recorded within seven days before the considered questionnaire. The other approach instead considers two windows: the first window is the same as before, and the second window instead considers sensor data recorded when the previously available questionnaire occurred. The second approach aims to provide the model with more information about the changes over time. However, the irregularity of sensor data penalizes the two-windows approach. Indeed, 20 out of 54 patients did not have two 7-day periods with a minimum of three days of sensor data. As a result, only the model using the mono window approach was submitted. In general, the results vary significantly depending on the questionnaire and showed better performance for the first task. The lower error in Task 1 may be due to the questionnaire being completed by clinical staff, whose responses are typically more reliable and objective compared to the subjective opinions provided by patients. To address the raised issue, data augmentation is proposed as a possible solution to increase the number of questionnaires in the training set. In this way, deep learning models could be tested to improve predictions and leverage longer sensor data sequences.

Martins et al. (2024) (Team FCOOL) proposed a methodology consisting of independent multi-class models, each predicting a distinct ALSFRS-R question. The authors tested four classification models: Logistic Regression, RF, XGBoost, and Support Vector Machine. To manage sensor data, they first derived static features from the longitudinal data via summarization techniques, and then reduced the feature set using three methods: top-k selection across questions, top-k selection by question, and biclustering. In both tasks, RF achieved top performance among the considered models, but failed to outperform the Last Observation Carried Forward (LOCF) baseline, except for a few individual questions. Moreover, no consensus was found about the best feature selection or extraction approach. Instead, top-k selection by question was the best approach in Task 1, while biclustering in Task 2.

Mehta et al. (2024) (Team iDPPExplorers) submitted runs only for Task 1 but analyze the approaches for Task 2 on their working notes paper. Their work focuses on handling the temporal aspect of the sensor data, by studying how to compress it via statistical methods that provide interpretability. Among the set of approaches tested in their work, Mehta et. al. (2024) observe that the optimal performance is achieved by both a naive baseline and ElasticNet regression. Nevertheless, the authors also observe that, despite the similar performance, the ElasticNet model is more robust and allows a better understanding of

the contribution of various features. While they did not take part in Task 2, they observed that the proposed approach is able to achieve better results on self-assessed data provided by the patients. Finally, their conclusive remark hints that, while this preliminary analysis did not highlight any major benefit of using sensor data, a larger dataset with a more diverse set of patients might lead to different conclusions.

In Tasks 1 and 2, Marinello et al. (2024) (Team UNIPD) developed a broad set of predictive models based on different methodological approaches using different subsets of the provided variables. The aim of their study was to evaluate whether considering wearable data to predict ALS disability leads to better performance with respect to models that only consider disease-specific variables collected during routine visits. They observe 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 in order to acquire high-quality data leading to significant datasets. Otherwise, if the quality of the acquired wearable data is poor, predicting the next visit ALSFRS-R score by simply holding the current one seems to be a better approach. This is especially true when predicting scores that are self-assigned by patients (task 2), who seem to be more stable and conservative with respect to their clinician during the disability evaluation process over time.

Okere et al. (2024) (Team UBCS) explores different deep-learning techniques to process data, especially to handle missing values. In particular, the authors exploit auto-encoders and multiple imputation techniques to handle missing values and use a RF algorithm to select relevant features. Subsequently, four deep neural networks, such as Multi-Layer Perceptron (MLP), Feed Forward Neural Network (FFNN), Recurrent Neural Network (RNN), and Long-Short Term Memory (LSTM), were trained to perform the two tasks. Experimental results revealed that ensemble predictive models, such as the XGBoost algorithm, show better performance than deep learning models. The authors link the low performance of the models with the small size of the training data.

7.4.2 Task 3

Bosoni et al. (2024) (Team Stefagroup) used Topological Data Analysis to compute personal exposure patterns and then employed two predictive approaches. The former relied on applying Linear Regression, RF, and XGBoost to the last follow-up data. The latter used Mixed-Effects modeling on longitudinal data from first to last follow-up. The results showed that incorporating environmental variables provides information statistically significant for predicting relapses. This outcome underlined the need for better methods to compute personal pollution exposure patterns, thereby enhancing the precision of MS progression predictions.

In task 3 Marinello et al. (2024) (Team UNIPD) developed a broad set of predictive models based on different methodological approaches using different subsets of the provided variables. The aim of their study was to evaluate whether considering environmental data to predict MS relapses leads to better performance with respect to models that only consider disease-specific variables collected during routine visits. They observe that environmental data can be beneficial for predicting the occurrence of MS relapses, however, better solutions should be explored to refine the data collection and variable extraction process in order to obtain more precise and focused predictions.

8 iDPP@CLEF 2024 WORKSHOP

Tasks

iDPP CLEF 2024 offers two evaluation tasks focused on predicting the progression of Amyotrophic Lateral Sclerosis (ALS) from prospective data and one task on impact of exposure to pollutants on predicting the progression of Multiple Sclerosis (MS) from retrospective data.

For Tasks 1 and 2 on ALS participants are given prospective patient data collected over an average of nine months via a dedicated app developed by the BRAINTEASER project and sensor data collected from the sensors of a fitness smartwatch in the context of clinical trials in Turin, Pavia, Lisbon, and Madrid, fully anonymized.

For Task 3 on MS, participants are given a retrospective dataset containing roughly 1.5 years of visits and environmental data. This dataset comes from two clinical institutions, one in Pavia, Italy, and the other in Turin, Italy, and it contains data about real patients, fully anonymized.

All the datasets are highly curated and they are produced from the BRAINTEASER Ontology (BTO), developed by the BRAINTEASER project, which ensures the consistency of the data represented. Moreover, several checks have been performed to ensure that all the instances are clean, contain proper values in the expected ranges, and do not have contradictions.

Task 1 - Predicting ALSFRS-R Score from Sensor Data (ALS)

It focuses on predicting the twelve scores of the ALSFRS-R (ALS Functional Rating Scale - Revised), assigned by medical doctors roughly every three months, from the sensor data collected via the app. The ALSFRS-R is a somehow "subjective" evaluation usually performed by a medical doctor and this task will help in answering a currently open question in the research community, i.e. whether it could be derived from objective factors.

Participants will be given the ALSFRS-R questionnaire at the first visit with the scores for each question together with the time (number of days from diagnosis) at which the questionnaire was taken.

Participants will have to predict the values of the ALSFRS-R sub-scores at the second visit.

Participants will be given the time of the second visit (number of days from diagnosis) together with all the sensor data up to the time of the second visit.

The training data are available upon registration for the challenge and the test data will be available one week before the run submission deadline. Please see the Datasets and Important Dates sections for more information.

Task 2 - Predicting Patient Self-assessment Score from Sensor Data

It focuses on predicting the self-assessment score assigned by patients from the sensor data collected via the app. Self-assessment scores correspond to each of the ALSFRS-R scores but, while the latter ones are assigned by medical doctors during visits, the former ones are assigned via auto-evaluation by patients themselves using the provided app.

If the self-assessment performed by patients, more frequently than the assessment performed by medical doctors every three months or so, can be reliably predicted by sensor and app data, we can imagine a proactive application which, monitoring the sensor data, alerts the patient if an assessment is needed.

Participants will be given the first set of self-assessed scores together with the time (number of days from diagnosis) at which the questionnaire was taken.

Participants will have to predict the values of the self-assessed scores at the second auto-evaluation, happening one or two months after the first one.

Participants will be given the time of the second auto-evaluation (number of days from diagnosis) together with all the sensor data up to the time of the second auto-evaluation.

The training data are available upon registration for the challenge and the test data will be available one week before the run submission deadline. Please see the Datasets and Important Dates sections for more information.

Task 3 - Predicting Relapses from EDDS Sub-scores and Environmental Data (MS)

It focuses on predicting a relapse using environmental data and EDSS (Expanded Disability Status Scale) sub-scores. This task allows us to assess if exposure to different pollutants is a useful variable in predicting a relapse.

Participants will be asked to predict the week of the first relapse after the baseline considering environmental data based on a weekly

Figure 5. The iDPP@CLEF 2024 Particpation Guidelines.

As for the previous years, participants to the iDPP@CLEF challenge had access to the participation guidelines to the iDPP@CLEF website (reachable here: https://brainteaser.dei.unipd.it/challenges/idpp2024).

The challenge had the following schedule:

- Registration closed: April 22, 2024
- Test data release: April 29, 2024

- Runs submission deadline: May 6, 2024
- Evaluation results out: May 17, 2024
- Participant and position paper submission deadline: May 31, 2024
- Notification of acceptance for participant and position papers: June 24, 2024
- Camera-ready participant papers submission: July 8, 2024
- iDPP@CLEF Workshop: September 9-12, 2024 during the CLEF Conference in Grenoble, France

To foster inclusivity, the iDPP@CLEF workshop was held in dual modality, with both online and physical participants. As for iDPP@CLEF 2023 and 2022, the challenge was first presented during the plenary session (Figure 6), where one of the organizers of the challenge introduced the tasks and the main objectives, to favour additional participation in future sessions, as well as to disseminate the findings to the scientific community (Figure 8).

The code of the models produced by the participants is publicly available at https://zenodo.org/records/14030410 (Birolo et al., 2024b)

Figure 6. The iDPP@CLEF Presentation during the plenary meeting.

Figure 7. The audience at the plenary meeting of CLEF 2024.

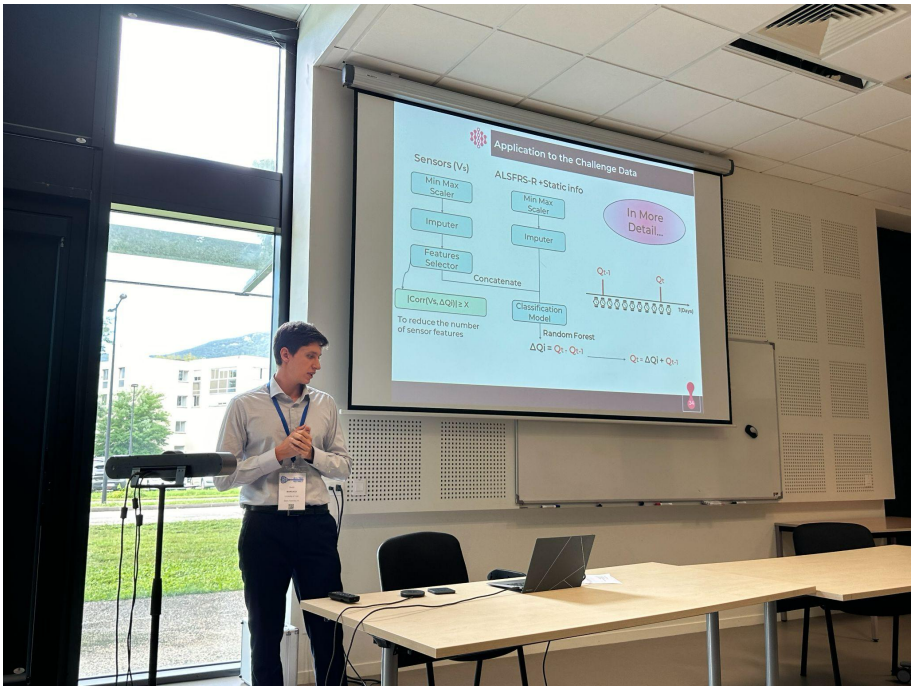

Figure 8. A presentation of one of the participants in iDPP@CLEF 2024.

Figure 9. The audience at the iDPP@CLEF 2024 session.

9 CONCLUSIONS AND FUTURE WORK

iDPP 2024 is the third and last iteration of the iDPP evaluation campaign. The focus of this evaluation campaign was on developing AI models capable of preemptively estimating the risks that patients affected by ALS and MS will need medical support and to describe the progression of their disease, to foster patient stratification and aid clinicians in providing the due care in the most effective and rapid way.

iDPP 2024 operated in continuation with iDPP 2022 and iDPP 2023, expanding previously proposed tasks, but also identifying novel tasks. In particular, iDPP was organized into 3 tasks. The first two tasks focused on predicting the ALSFRS-R for patients affected by ALS, using data collected via environmental sensors and wearable devices. This makes iDPP 2024 the first edition of making use of data collected on patients currently involved in the BRAINTEASER project. The third task of iDPP 2024 built upon the results of iDPP 2023, by focusing on the prediction of the disease progression of patients affected by MS. More in detail, this task focused on predicting when an MS patient will experience a relapse. As an addition over the previous iDPP edition, this year participants were also provided with environmental data that could be used to improve the AI models.

In terms of participation, 28 teams registered in the Lab, suggesting overall interest in the topic from the research community, and 8 teams were able to submit their results for a total of 97 submitted runs. The task that received the most interest was the first, with 59 submissions alone.

While this cycle concludes the evaluation campaign of iDPP, we envision several possible research paths for which iDPP paved the way. First of all, novel and more effective AI approaches can be developed in the future, by using iDPP data as training and evaluation sets. Secondly, iDPP has identified several guidelines and good practices that can be adapted to devise novel shared tasks and evaluation campaigns in the future, either concerning ALS and MS, other neurological diseases, or the medical domain at large.

9.1 Acknowledgments

The work reported in this paper has been partially supported by the BRAINTEASER[9] project (contract n. GA101017598), as a part of the European Union's Horizon 2020 research and innovation programme.

[9] https://brainteaser.health/

10 REFERENCES

Barducci, G., Sartori, F., Birolo, G., Sanavia, T., & Fariselli, P. (2024). ALSFRS-R Score Prediction for Amyotrophic Lateral Sclerosis. In G. Faggioli, N. Ferro, P. Galuscáková, & A. G. S. de Herrera (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://ceur-ws.org/Vol-3740/paper-123.pdf

Birolo, G., Bosoni, P., Faggioli, G., Aidos, H., Bergamaschi, R., Cavalla, P., Chiò, A., Dagliati, A., de Carvalho, M., Di Nunzio, G. M., Fariselli, P., García Dominguez, J. M., Gromicho, M., Guazzo, A., Longato, E., Madeira, S. C., Manera, U., Marchesin, S., Menotti, L., … Ferro, N. (2024a). Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, 118–139. https://doi.org/10.1007/978-3-031-71908-0_6

Birolo, G., Bosoni, P., Faggioli, G., Aidos, H., Bergamaschi, R., Cavalla, P., Chiò, A., Dagliati, A., de Carvalho, M., Di Nunzio, G. M., Fariselli, P., García Dominguez, J. M., Gromicho, M., Guazzo, A., Longato, E., Madeira, S. C., Manera, U., Marchesin, S., Menotti, L., … Ferro, N. (2024b). *iDPP@CLEF 2024—Participants' repositories for the Intelligent Disease Prediction Progression Challenge* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.14030410

Bosoni, P., Vazifehdan, M., Pala, D., Tavazzi, E., Bergamaschi, R., Bellazzi, R., & Dagliati, A. (2024). Predicting Multiple Sclerosis Relapses Using Patient Exposure Trajectories. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://api.semanticscholar.org/CorpusID:271866136

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). *Journal of the Neurological Sciences*, *169*(1–2), 13–21. https://doi.org/10.1016/s0022-510x(99)00210-5

Faggioli, G., Guazzo, A., Marchesin, S., Menotti, L., Trescato, I., Aidos, H., Bergamaschi, R., Birolo, G., Cavalla, P., Chiò, A., Dagliati, A., de Carvalho, M., Di Nunzio, G. M., Fariselli, P., García Dominguez, J. M., Gromicho, M., Longato, E., Madeira, S. C., Manera, U., … Ferro, N. (2023a). Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 343–369). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42448-9_24

Faggioli, G., Guazzo, A., Marchesin, S., Menotti, L., Trescato, I., Aidos, H., Bergamaschi, R., Birolo, G., Cavalla, P., Chiò, A., Dagliati, A., de Carvalho, M., Di Nunzio, G. M., Fariselli, P., García Dominguez, J. M., Gromicho, M., Longato, E., Madeira, S. C., Manera, U., … Ferro, N. (2023b). Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*. https://ceur-ws.org/Vol-3497/paper-095.pdf

Guazzo, A., Trescato, I., Longato, E., Hazizaj, E., Dosso, D., Faggioli, G., Di Nunzio, G. M., Silvello, G., Vettoretti, M., Tavazzi, E., Roversi, C., Fariselli, P., Madeira, S. C., de Carvalho, M., Gromicho, M., Chiò, A., Manera, U., Dagliati, A., Birolo, G., … Ferro, N. (2022a). Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 395–422). Springer International Publishing. https://doi.org/10.1007/978-3-031-13643-6_25

Guazzo, A., Trescato, I., Longato, E., Hazizaj, E., Dosso, D., Faggioli, G., Di Nunzio, G. M., Silvello, G., Vettoretti, M., Tavazzi, E., Roversi, C., Fariselli, P., Madeira, S. C., de Carvalho, M., Gromicho, M., Chiò, A., Manera, U., Dagliati, A., Birolo, G., … Ferro, N. (2022b). Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge. *CLEF 2022: Conference and Labs of the Evaluation Forum*. https://ceur-ws.org/Vol-3180/paper-88.pdf

Küffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L., Li, G., Fang, L., Mackey, L., Hardiman, O., Cudkowicz, M., Sherman, A., Ertaylan, G., Grosse-Wentrup, M., Hothorn, T., van Ligtenberg, J., Macke, J. H., Meyer, T., Schölkopf, B., … Leitner, M. L. (2015). Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nature Biotechnology*, *33*(1), 51–57. https://doi.org/10.1038/nbt.3051

Marinello, E., Guazzo, A., Longato, E., Tavazzi, E., Trescato, I., Vettoretti, M., & Camillo, B. D. (2024). Using Wearable and Environmental Data to Improve the Prediction of Amyotrophic Lateral Sclerosis and Multiple Sclerosis Progression: An Explorative Study. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://ceur-ws.org/Vol-3740/paper-125.pdf

Martins, A. S., Amaral, D. M., Castanho, E. N., Soares, D. F., Branco, R., Madeira, S. C., & Aidos, H. (2024). Predicting the Functional Rating Scale and Self-Assessment Status of ALS Patients with Sensor Data. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://ceur-ws.org/Vol-3740/paper-126.pdf

Mehta, R., Pramov, A., & Verma, S. (2024, July 10). Machine Learning for ALSFRS-R Score Prediction: Making Sense of the Sensor Data. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://doi.org/10.48550/arXiv.2407.08003

Okere, C. C., Thuma, E., & Mosweunyane, G. (2024). UBCS at IDPP: Predicting Patient Self-Assessment Score from Sensor Data using Machine Learning Algorithms. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://ceur-ws.org/Vol-3740/paper-128.pdf

Silva, J. M., & Oliveira, J. L. (2024). BIT.UA at iDPP: Predictive Analytics on ALS Disease Progression Using Sensor Data with Machine Learning. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. https://ceur-ws.org/Vol-3740/paper-129.pdf

World Health Organization. (2021). *WHO Global Air Quality Guidelines: Particulate Matter (PM2. 5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide* (1st ed). World Health Organization.