Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024

Giovanni Birolo[1,*], Pietro Bosoni[2,*], Guglielmo Faggioli[3,*], Helena Aidos[4], Roberto Bergamaschi[2], Paola Cavalla[1,5], Adriano Chiò[1], Arianna Dagliati[2], Mamede de Carvalho[4], Giorgio Maria Di Nunzio[3], Piero Fariselli[1], Jose Manuel García Dominguez[6], Marta Gromicho[4], Alessandro Guazzo[3], Enrico Longato[3], Sara C. Madeira[4], Umberto Manera[1], Stefano Marchesin[3], Laura Menotti[3], Gianmaria Silvello[3], Eleonora Tavazzi[7], Erica Tavazzi[3], Isotta Trescato[3], Martina Vettoretti[3], Barbara Di Camillo[3], and Nicola Ferro[3]

[1] University of Turin, Italy
{adriano.chio, giovanni.birolo, piero.fariselli, umberto.manera}@unito.it
[2] University of Pavia, Italy
pietro.bosoni@unipv.it, roberto.bergamaschi@mondino.it, arianna.dagliati@unipv.it
[3] University of Padua, Italy
{barbara.dicamillo, giorgiomaria.dinunzio, guglielmo.faggioli, nicola.ferro, enrico.longato, stefano.marchesin, laura.menotti, gianmaria.silvello, erica.tavazzi, martina.vettoretti}@unipd.it, {alessandro.guazzo, isotta.trescato}@phd.unipd.it
[4] University of Lisbon, Portugal
mamedemg@mail.telepac.pt, mgromichosilva@medicina.ulisboa.pt, {haidos, sacmadeira}@fc.ul.pt
[5] "Città della Salute e della Scienza", Turin, Italy
paola.cavalla@unito.it
[6] Gregorio Marañon Hospital in Madrid, Spain
jgarciadominguez@salud.madrid.org
[7] IRCCS Foundation C. Mondino in Pavia, Italy
eleonoratavazzi@gmail.com

Abstract. Multiple Sclerosis (MS) and Amyotrophic Lateral Sclerosis (ALS) are two neurodegenerative diseases that cause progressive or alternating neurological impairments in motor, sensory, visual, and cognitive functions. Patients affected by these diseases undergo the physical, psychological, and economic burdens of hospital stays and home care while facing uncertainty. A possible aid to patients and clinicians might come from AI tools that can preemptively identify the need for intervention and suggest personalized therapies during the progression of these diseases.

The objective of iDPP@CLEF is to develop automatic approaches based on AI that can be used to describe the progression of these two neurodegenerative diseases, with the final goal of allowing patient stratification as well as the prediction of the disease progression, to help clinicians in assisting patients in the most timely manner.

[*] These authors contributed equally.

iDPP@CLEF 2024 follows the two prior editions, iDPP@CLEF 2022 and 2023. iDPP@CLEF 2022 focused on ALS progression prediction and approaches of explainable AI for the task. iDPP@CLEF 2023 built upon iDPP@CLEF 2022 by extending the datasets provided during the previous edition with environmental data. Additionally, the 2023 edition of iDPP@CLEF introduced a new task focused on the progression prediction of MS. In this edition, we extended the MS dataset of iDPP@CLEF 2023 with environmental data. Furthermore, we introduced two new ALS tasks, focused on predicting the progression of the disease using data obtained from wearable devices, making it the first iDPP edition that uses prospective data collected directly from the patients involved in the BRAINTEASER project.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) and *Multiple Sclerosis (MS)* are two severe and extremely impactful diseases that cause progressive neurological impairment in people living with them. Typically, for both diseases, the progression is heterogeneous, determining a large variability in several aspects including the treatment of the patients, the outcome, the quality of life, and, in general, their needs. This variability represents a challenge not only for patients but also for clinicians and caregivers alike. Indeed, patients with ALS tend to require, at a certain point of the progression of their disease, some specific treatment, such as *Non-Invasive Ventilation (NIV)* or *Percutaneous Endoscopic Gastrostomy (PEG)*. Similarly, patients living with MS tend to undergo impairing relapses that may cause severe drops in their quality of life. Therefore, it would beneficial to know beforehand the needs of a person affected by one of these diseases. Nevertheless, due to their heterogeneity, it is challenging to develop effective prognostic tools. This motivates the importance of developing automatic tools to aid clinicians in their decision-making in all phases of disease progression and facilitate personalized therapeutic choices. In particular, when developing new automatic predictive approaches based on *Artificial Intelligence (AI)*, researchers need a proper framework that allows for designing and evaluating approaches for different tasks, such as:

- stratifying patients according to their phenotype all over the disease evolution;
- predicting the progression of the disease in a probabilistic, time-dependent way;
- describing better and in an explainable fashion the mechanisms underlying MS and ALS diseases.

Nonetheless, it is of uttermost importance that such approaches are based on shared resources that allow for proper benchmarking and comparable and reproducible experimentation. The *Intelligent Disease Progression Prediction at CLEF (iDPP@CLEF)* lab[8] aims to provide an evaluation infrastructure for

[8] https://brainteaser.health/open-evaluation-challenges/

the development of such AI algorithms. Differently from previous efforts in this domain, iDPP@CLEF systematically addresses issues related to the application of AI in clinical practice for ALS and MS. Apart from defining risk scores based on the probability of events occurring in the short or long term, iDPP@CLEF also deals with providing clinicians with structured and understandable data. iDPP@CLEF 2024 [2] is the last iteration of an evaluation cycle of three challenges aimed at fostering reproducible and comparable evaluation of AI based approaches to predict the progression of ALS and MS. The first edition, iDPP@CLEF 2022 focused on ALS, asking participants to predict the probability that patients would incur the need for specific medical treatments based on their medical history. The second edition iDPP@CLEF 2023 focused on extending the dataset of iDPP@CLEF 2022 with environmental data, to determine the impact that the environment might have on the needs of the patients. Furthermore, a new task based on predicting the risk of worsening of MS patients was proposed. This final edition extends iDPP@CLEF 2023 by providing environmental data for patients affected by MS, to measure the impact that pollution and the external environment can have on the progression of MS. Furthermore, two new tasks have been proposed in iDPP@CLEF 2024. These tasks required participants to predict the progression of ALS, measured according to the ALSFRS-R scale, based on the clinical history of the patients, as well as measurements obtained via 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; finally, Section 8 draws some conclusions and outlooks some future work.

2 Related Challenges

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

While no major challenges – besides iDPP@CLEF 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[9] in 2012 and the DREAM ALS Stratification challenge[10] in 2015 and a Kaggle challenge[11] 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 *ALS Functional Rating Scale Revisited (ALSFRS-R)* [4]. Later on, the DREAM ALS Stratification challenge [9] required participants to stratify

[9] https://dreamchallenges.org/dream-7-phil-bowen-als-prediction-prize4life/
[10] https://dx.doi.org/10.7303/syn2873386.
[11] https://www.kaggle.com/alsgroup/end-als

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@CLEF 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@CLEF 2022

iDPP@CLEF 2022[12] [7, 8] 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@CLEF 2022 was organized into 3 tasks:

- **Pilot Task 1 - Ranking Risk of Impairment**: The focus of the first task of iDPP@CLEF 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 in time, measured according to the ALSFRS-R scale [4], 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). The task focused on assessing model calibration in terms of the ability 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 AI 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@CLEF 2022 were the three 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 spyrometries). six groups participated in iDPP@CLEF 2022 and submitted a total of 120 runs.

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

2.2 iDPP@CLEF 2023

Similarly to iDPP@CLEF 2022, also iDPP@CLEF 2023[13] [5, 6] were organized into three tasks, focusing on either ALS or MS. More in detail, Task 1 and Task 2 of iDPP@CLEF 2023 concerned MS, while Task 3 built upon iDPP@CLEF 2022 and extended the ALS tasks of the previous iteration of the Lab. To summarize iDPP@CLEF 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@CLEF 2022 and concerned the ranking of the patients based on the risk of impairment. The major difference to iDPP@CLEF 2022 was that participants were given environmental data to determine if such data was a good predictor of the risk of impairment.

iDPP@CLEF 2023 extended the iDPP@CLEF 2022 datasets with two 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 Task 2. Ten teams submitted a total of 163 runs at the end of iDPP@CLEF 2023.

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 somewhat "subjective" evaluation usually performed by a medical doctor. This task aims to answer an open question in the research community, i.e., whether the ALSFRS-R scores can be derived from objective factors.

Participants were given the ALSFRS-R questionnaire at the first visit, including the scores for each question and the time (number of days from diagnosis)

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

when the questionnaire was taken; moreover, they were also provided with the time of the second visit (number of days from diagnosis) and 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 scores 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 are assigned by medical doctors during visits, the self-assessment scores are assigned by patients themselves using the app.

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

Participants were given the first set of self-assessed scores along with the time (number of days from diagnosis) at which the questionnaire was taken; furthermore, they were also provided with the time of the second auto-evaluation (number of days from diagnosis) and 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, which occurs 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@CLEF 2024 we release three datasets: two completely new datasets for ALS and an extension of the iDPP@CLEF 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@CLEF 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. They were prompted for the assessment once per month, though the actual frequency varied and depended on patient compliance.

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.

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).

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.

Updates over IDPP@CLEF 2023 In the 2024 dataset, EDSS data after 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 [15], 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 [14]. 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 [15]. 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

[14] https://discomap.eea.europa.eu/Index/
[15] https://www.ecad.eu/download/ensembles/download.php

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.

Finally, a subset of 380 MS patients from 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[16] 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 visits occur between the baseline and the first relapse, while environmental measurements span from January 1, 2013, to 2023. It is important to note that environmental data may have gaps due to data availability.

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. Table1 provides a comparison of variable distributions between the training and test sets, confirming that the split meets the best-practice quality standards.

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@CLEF.

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.

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

Table 1: 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%)
	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)

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, an 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 values 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.

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, an hashed version of the original patient ID (should be considered just as a string);
- Columns from 2 to 13 represent the predicted self-assessd sub-score. Each column corresponds to an ALSFRS-R question, e.g. column 2 to Q1, column 3 to Q2, and so on). Each values 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.

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.

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@CLEF 2024, out of which 8 teams were able to submit one run in at least one task. Table 2 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@CLEF 2024. The most participated task was Task 1 with 59 runs and 6 teams participating. Subsequently, Task 2 had 31 runs submitted by 6 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 2: Teams participating in iDPP@CLEF 2024.

Team Name	Affiliation	Country	Repository	Paper
BIT.UA	IEETA/DETI, LASI, University of Aveiro	Portugal	https://bitbucket.org/brainteaser-health/idpp2024-bitua	Silva and Oliveira [14]
CompBiomedUniTO	University of Torino	Italy	https://bitbucket.org/brainteaser-health/idpp2024-compbiomedunito	Barducci et. al. [1]
FCOOL	LASIGE, Faculty of Sciences, University of Lisbon	Portugal	https://bitbucket.org/brainteaser-health/idpp2024-fcool	Martins et. al. [11]
iDPPExplorers	Georgia Institute of Technology, Atlanta, GA	United States	https://bitbucket.org/brainteaser-health/idpp2024-idppexplorers	Metha et. al. [12]
Mandatory	University of Bucharest	Romania	https://bitbucket.org/brainteaser-health/idpp2024-mandatory	—
Stefagroup	University of Pavia, BMI lab "Mario Stefanelli"	Italy	https://bitbucket.org/brainteaser-health/idpp2024-stefagroup	Bosoni et. al. [3]
UBCS	University of Botswana	Botswana	https://bitbucket.org/brainteaser-health/idpp2024-ubcs	Okere et. al. [13]
UNIPD	University of Padova	Italy	https://bitbucket.org/brainteaser-health/idpp2024-unipd	Martinello et. al. [10]

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 Equation 1 and Equation 2, 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.

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

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

Table 3: Number of runs submitted by each participant team in iDPP@CLEF 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

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 simulation 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 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 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.

Four teams - iDPPExplorers, Mandatory, FCOOL, and UNIPD - employed this strategy in one of their runs for Task 1, achieving the lowest errors with

Table 4: For both MAE and RMSE, we report 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 rankings

team	run	MAE	RANK(MAE)	RMSE	RANK(RMSE)
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
bitua	temporalAnalysis	0.29±0.24	5	0.61±0.24	6
unipd	average	0.33±0.18	6	0.60±0.18	5
unipd	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
mandatory	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
mandatory	1hist03	0.42±0.24	24	0.69±0.24	19
mandatory	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
mandatory	1hist08	0.44±0.26	27	0.71±0.26	23
idppexplorers	et	0.44±0.24	27	0.78±0.24	36
idppexplorers	dt	0.44±0.22	28	0.72±0.22	31
idppexplorers	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
idppexplorers	lstm	0.48±0.27	32	0.82±0.27	43
mandatory	1hist07	0.48±0.21	33	0.75±0.21	34
mandatory	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
fcool	XGBClassifier	0.57±0.15	42	0.83±0.15	44
fcool	majority-class	0.66±0.52	43	1.09±0.52	50

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.

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.

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 1 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_t3_rf` run [3], 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) 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_t3_ridge_noenv` run, which excludes environmental variables [10].

7.4 Approaches

In this section, we provide a short summary of the approaches adopted by participants in iDPP@CLEF. 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.

Tasks 1 and 2 Silva and Oliveira [14] (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

Table 5: For both MAE and RMSE, results are reported as the average error across all twelve ALSFRS-R scores, followed by their average standard deviation (computed by bootstrapping the test set one thousand times), and their respective rankings

metric team run	MAE	RANK(MAE)	RMSE	RANK(RMSE)
fcool locf	0.29±0.15	1	0.58±0.15	1
unipd hold	0.29±0.15	1	0.58±0.15	1
CBMUnito RF-MonoWindow	0.31±0.16	2	0.60±0.16	2
bitua ensemble-max	0.33±0.14	3	0.61±0.14	3
bitua moremetrics	0.37±0.17	4	0.65±0.17	4
bitua mean	0.39±0.18	5	0.71±0.18	8
bitua median	0.40±0.21	6	0.69±0.21	5
fcool 2nd-best-both-metrics	0.41±0.15	7	0.71±0.15	6
bitua ensemble-avg	0.42±0.22	8	0.71±0.22	7
bitua idpp2024-bitua	0.43±0.24	9	0.72±0.24	9
unipd average	0.49±0.20	10	0.78±0.20	12
fcool 3rd-best-both-metrics	0.50±0.13	11	0.78±0.13	10
unipd logistic-ALSFRS	0.50±0.19	11	0.85±0.19	18
bitua ensemble-min	0.50±0.24	12	0.82±0.24	14
unipd rf	0.52±0.20	13	0.78±0.20	11
unipd rf-reg	0.52±0.12	14	0.82±0.12	13
fcool best-both-metrics	0.53±0.20	15	0.84±0.20	15
fcool RFClassifier	0.53±0.24	16	0.85±0.24	17
unipd ridge	0.55±0.27	17	0.85±0.27	16
fcool LogisticRegression	0.57±0.21	18	0.89±0.21	19
fcool XGBClassifier	0.59±0.17	19	0.93±0.17	20
unipd optrun	0.61±0.27	20	0.96±0.27	21
unipd logistic	0.66±0.29	21	0.99±0.29	22
fcool SVC	0.67±0.19	22	1.01±0.19	23
ubcs features100	0.82±0.43	23	1.20±0.43	26
ubcs featuresall	0.89±0.41	24	1.25±0.41	27
ubcs features10	0.94±0.49	25	1.33±0.49	28
ubcs features25	0.96±0.21	26	1.14±0.21	24
ubcs features20	1.02±0.24	27	1.18±0.24	25
fcool majority-class	1.03±0.44	28	1.47±0.44	29
ubcs features50	1.11±0.51	29	1.51±0.51	30

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 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. [1] (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 7 days before the considered questionnaire. The other approach instead considers two windows: the first window is the same as before, and the second window in-

Table 6: RMSE and MAE results for all the submitted runs for Task 3

Team	RMSE	MAE
UNIPD_t3_ridge	89.84	68.59
UNIPD_t3_rf_reg	79.74	66.63
UNIPD_t3_average	79.26	65.80
UNIPD_t3_ridge_noenv	78.62	61.37
UNIPV_t3_lmer_last	72.51	47.74
UNIPV_t3_lmer_first	48.07	28.05
UNIPV_t3_rf	41.52	22.49

stead 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 3 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 possibile 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. [11] (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. [12] (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. 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

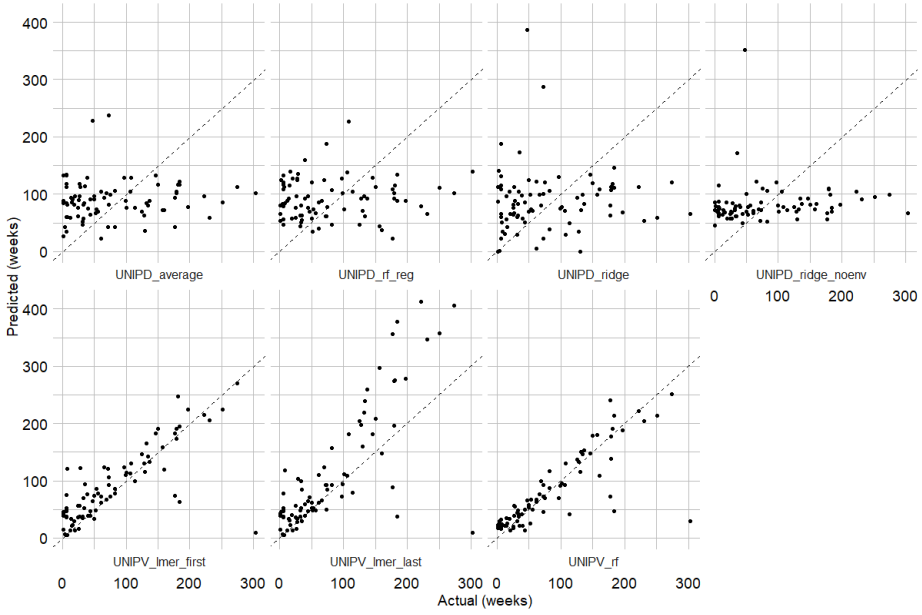

Fig. 1: Actual versus Predicted values for each run submitted for Task 3

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, Martinello et. al. [10] (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. [13] (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.

Task 3 Bosoni et. al. [3] (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 Martinello et. al. [10] (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 Conclusions and Future Work

iDPP@CLEF 2024 is the third and last iteration of the iDPP@CLEF 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@CLEF 2024 operated in continuation with iDPP@CLEF 2022 and iDPP@CLEF 2023, expanding previously proposed tasks, but also identifying novel tasks. In particular, iDPP@CLEF was organized into three 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@CLEF 2024 the first edition of making use of data collected on patients currently involved in the BRAINTEASER project. The third task of iDPP@CLEF 2024 built upon the results of iDPP@CLEF 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 improvement over the previous iDPP edition, this year participants were also provided with environmental data that could have been 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@CLEF, we envision several possible research paths for which iDPP@CLEF paved the way. First of all, novel and more effective AI approaches can be developed in the future, by using iDPP@CLEF data as training and evaluation sets. Secondly, iDPP@CLEF 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.

Acknowledgments

The work reported in this paper has been partially supported by the BRAINTEASER[17] project (contract n. GA101017598), as a part of the European Union's Horizon 2020 research and innovation programme.

References

1. Barducci, G., Sartori, F., Birolo, G., Sanavia, T., Fariselli, P.: Alsfrs-r score prediction for amyotrophic lateral sclerosis. In: CLEF 2024 Working Notes (2024)
2. Birolo, G., Bosoni, P., Faggioli, G., Aidos, H., Bergamaschi, R., Cavalla, P., Chiò, A., Dagliati, A., de Carvalho, M., Di Nunzio, G., Fariselli, P., García Dominguez, J., Marta Gromicho, A.G., Longato, E., Madeira, S., Manera, U., Marchesin, S., Menotti, L., Silvello, G., Tavazzi, E., Tavazzi, E., Trescato, I., Vettoretti, M., Camillo, B.D., Ferro, N.: Overview of iDPP@CLEF 2024: The Intelligent Disease Progression Prediction Challenge. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024 (2024)
3. Bosoni, P., Vazifehdan, M., Pala, D., Tavazzi, E., Bergamaschi, R., Bellazzi, R., Dagliati, A.: Predicting multiple sclerosis relapses using patient exposure trajectories. In: CLEF 2024 Working Notes (2024)
4. Cedarbaum, J.M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., Nakanishi, A.: The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. Journal of the Neurological Sciences **169**(1–2), 13–21 (October 1999)

[17] https://brainteaser.health/

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., Nunzio, G.M.D., Fariselli, P., Dominguez, J.M.G., Gromicho, M., Longato, E., Madeira, S.C., Manera, U., Silvello, G., Tavazzi, E., Tavazzi, E., Vettoretti, M., Camillo, B.D., Ferro, N.: Intelligent disease progression prediction: Overview of idpp@clef 2023. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, Lecture Notes in Computer Science, vol. 14163, pp. 343–369, Springer (2023), https://doi.org/10.1007/978-3-031-42448-9_24, URL https://doi.org/10.1007/978-3-031-42448-9_24
6. 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., Nunzio, G.M.D., Fariselli, P., Dominguez, J.M.G., Gromicho, M., Longato, E., Madeira, S.C., Manera, U., Silvello, G., Tavazzi, E., Tavazzi, E., Vettoretti, M., Camillo, B.D., Ferro, N.: Overview of idpp@clef 2023: The intelligent disease progression prediction challenge. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, CEUR Workshop Proceedings, vol. 3497, pp. 1123–1164, CEUR-WS.org (2023), URL https://ceur-ws.org/Vol-3497/paper-095.pdf
7. 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., Aidos, H., Di Camillo, B., Ferro, N.: Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022. In: Barrón-Cedeño, A., Da San Martino, G., Degli Esposti, M., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), pp. 395–422, Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany (2022)
8. 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., Aidos, H., Di Camillo, B., Ferro, N.: Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) CLEF 2022 Working Notes, pp. 1130–1210, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. http://ceur-ws.org/Vol-3180/ (2022)
9. 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., Tran, L., Vaughan, R., Stolovitzky, G., Leitner, M.L.: Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nature Biotechnology **33**(1), 51–57 (January 2015)
10. Marinello, E., Guazzo, A., Longato, E., Tavazzi, E., Trescato, I., Vettoretti, M., Camillo, B.D.: Using wearable and environmental data to improve the prediction of amyotrophic lateral sclerosis and multiple sclerosis progression: an explorative study. In: CLEF 2024 Working Notes (2024)
11. Martins, A., Amaral, D., Castanho, E., Soares, D., Branco, R., Madeira, S., Aidos, H.: Predicting the functional rating scale and self-assessment status of als patients with sensor data. In: CLEF 2024 Working Notes (2024)
12. Mehta, R., Pramov, A., Verma, S.: Machine learning for alsfrs-r score prediction: Making sense of the sensor data. In: CLEF 2024 Working Notes (2024)
13. Okere, C., Thuma, E., Mosweunyane, G.: Ubcs at idpp: Predicting patient self-assessment score from sensor data using machine learning algorithms. In: CLEF 2024 Working Notes (2024)
14. Silva, J., Oliveira, J.: Bit.ua at idpp: Predictive analytics on als disease progression using sensor data with machine learning. In: CLEF 2024 Working Notes (2024)
15. World Health Organization: WHO global air quality guidelines: Particulate matter (PM(2.5) and PM(10)), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva (2021), ISBN 978-92-4-003422-8, review