Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023

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

[1] 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
[2] University of Lisbon, Portugal
mamedemg@mail.telepac.pt, mgromichosilva@medicina.ulisboa.pt, {haidos, sacmadeira}@fc.ul.pt
[3] University of Pavia, Italy
roberto.bergamaschi@mondino.it, arianna.dagliati@unipv.it
[4] University of Turin, Italy
{adriano.chio, giovanni.birolo, piero.fariselli, umberto.manera}@unito.it
[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. Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) are chronic diseases that cause progressive or alternating neurological impairments in motor, sensory, visual, and cognitive functions. Affected patients must manage hospital stays and home care while facing uncertainty and significant psychological and economic burdens that also affect their caregivers. To ease these challenges, clinicians need automatic tools to support them in all phases of patient treatment, suggest personalized therapeutic paths, and preemptively indicate urgent interventions.

iDPP@CLEF aims at developing an evaluation infrastructure for AI algorithms to describe ALS and MS mechanisms, stratify patients based on their phenotype, and predict disease progression in a probabilistic, time-dependent manner.

iDPP@CLEF 2022 ran as a pilot lab in CLEF 2022, with tasks related to predicting ALS progression and explainable AI algorithms for prediction.

[*] These authors contributed equally.

iDPP@CLEF 2023 will continue in CLEF 2023, with a focus on predicting MS progression and exploring whether pollution and environmental data can improve the prediction of ALS progression.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) and *Multiple Sclerosis (MS)* are severe chronic diseases that cause progressive neurological impairment. They exhibit high heterogeneity in terms of symptoms and disease progression, leading to differing needs for patients. The heterogeneity of these diseases partly explains the lack of effective prognostic tools and the current lack of therapies that can effectively slow or reverse their course. This poses challenges for caregivers and clinicians alike. Furthermore, the timing of worsening or significant events – such as the need for *Non-Invasive Ventilation (NIV)* or *Percutaneous Endoscopic Gastrostomy (PEG)* in the case of ALS – is uncertain and hard to predict. Being able to preemptively recognize the need for specific medical treatments would have significant implications for the quality of life of patients. Therefore, it would be of uttermost importance to devise automatic tools that could aid clinicians in their decision-making in all phases of disease progression and facilitate personalized therapeutic choices.

To address these challenges and develop *Artificial Intelligence (AI)* predictive algorithms researchers need a framework to design and evaluate approaches to:

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

In this context, it is crucial to develop shared approaches, promote common benchmarks, and foster experiment comparability and replicability, even if not yet so common. The *Intelligent Disease Progression Prediction at CLEF (iDPP@CLEF)* lab[8] aims to provide an evaluation infrastructure for the development of such AI algorithms. Unlike previous challenges in the field, 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.

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.

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

2 Related Challenges

Within CLEF, there have been no other labs on this or similar topics before the start of iDPP@CLEF. iDPP@CLEF 2022, whose details are summarized below, was the first iteration of the lab and the current is the second one.

Outside CLEF, there have been a recent challenge on Kaggle[9] in 2021 and some older ones, the DREAM 7 ALS Prediction challenge[10] in 2012 and the DREAM ALS Stratification challenge[11] in 2015. The Kaggle challenge used a mix of clinical and genomic data to seek insights about the mechanisms of ALS and the difference between people with ALS who progress faster versus those who develop it more slowly. The DREAM 7 ALS Prediction challenge [15] asked to use 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)* [5], a functional scale that ranges between 0 and 40. The DREAM ALS Stratification challenge asked participants to stratify ALS patients into meaningful subgroups, to enable better understanding of patient profiles and application of personalized ALS treatments. Differently from these previous challenges, iDPP@CLEF focuses on explainable AI and on temporal progression of the disease.

Finally, when it comes to *Multiple Sclerosis (MS)*, studies are mostly conducted on closed and proprietary datasets and iDPP@CLEF represents one of the first attempts to create a public and shared dataset.

2.1 iDPP@CLEF 2022

iDPP@CLEF 2022 ran as a pilot lab for the first time in CLEF 2022[12] [7, 8] and focused on activities aimed at ALS progression prediction as well as at an understanding of the challenges and limitations to refine and tune the labs itself for future iterations. iDPP@CLEF 2022 consisted of the following tasks:

- **Pilot Task 1 - Ranking Risk of Impairment**: it focused on ranking patients based on the risk of impairment. We used the ALSFRS-R scale [5] to monitor speech, swallowing, handwriting, dressing/hygiene, walking and respiratory ability in time and asked participants 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, we assessed model calibration in terms of the ability of the proposed algorithms to estimate a probability of an event close to the true probability within a specified time-window.
- **Position Paper Task 3 - Explainability of AI algorithms**: we evaluated proposals of different frameworks able to explain the multivariate nature of the data and the model predictions.

[9] https://www.kaggle.com/alsgroup/end-als
[10] https://dreamchallenges.org/dream-7-phil-bowen-als-prediction-prize4life/
[11] https://dx.doi.org/10.7303/syn2873386.
[12] https://brainteaser.health/open-evaluation-challenges/idpp-2022/

iDPP@CLEF 2022 created 3 datasets, for the prediction of specific events related to ALS, consisting of fully anonymized data from 2,250 real patients from medical institutions in Turin, Italy, and Lisbon, Portugal. The datasets contain both static data about patients, e.g. age, onset date, gender, ... and event data, i.e. 18,512 ALSFRS-R questionnaires and 4,015 spyrometries. 6 groups participated in iDPP@CLEF 2022 and submitted a total of 120 runs.

3 Tasks

iDPP@CLEF 2023 is the second iteration of the lab, expanding its scope to include both ALS and MS in the study of disease progression. The activities in iDPP@CLEF 2023 focus on two objectives: exploring the prediction of MS worsening and conducting a more in-depth analysis of ALS compared to iDPP@CLEF 2022, with the addition of environmental data.

Following iDPP@CLEF 2022, iDPP@CLEF 2023 targets three tasks:

- Pilot tasks (Task 1 and Task 2) on predicting the progression of the MS, focusing on its worsening;
- Position papers (Task 3) on the impact that environmental data can have on the progression of the ALS.

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

3.1 Task 1: Predicting Risk of Disease Worsening (MS)

Task 1 focuses on MS and requires ranking subjects based on the risk of worsening, setting the problem as a survival analysis task. More specifically the risk of worsening predicted by the algorithm should reflect how early a patient experiences the "worsening" event and should range between 0 and 1.

Worsening is defined on the basis of the *Expanded Disability Status Scale (EDSS)* [16], according to clinical standards. In particular, we consider two different definitions of worsening corresponding to two different sub-tasks:

- Task1a: the patient crosses the threshold EDSS ≥ 3 at least twice within a one-year interval;
- Task1b: the second definition of worsening depends on the first recorded value, according to current clinical protocols:
  - if the baseline is EDSS < 1, then the worsening event occurs when an increase of EDSS by 1.5 points is first observed;
  - if the baseline is $1 \leq$ EDSS < 5.5, then the worsening event occurs when an increase of EDSS by 1 point is first observed;
  - if the baseline is EDSS ≥ 5.5, then the worsening event occurs when an increase of EDSS by 0.5 points is first observed.

For each sub-task, participants are given a dataset containing 2.5 years of visits, with the occurrence of the worsening event and the time of occurrence pre-computed by the challenge organizers.

3.2 Task 2: Predicting Cumulative Probability of Worsening (MS)

Task 2 refines Task 1 by asking participants to explicitly assign the cumulative probability of worsening at different time windows, i.e., between years 0 and 2, 0 and 4, 0 and 6, 0 and 8, 0 and 10. In particular, as in Task 1, we consider two different definitions of worsening corresponding to two different sub-tasks:

- Task2a: the patient crosses the threshold EDSS ≥ 3 at least twice within a one-year interval;
- Task2b: the second definition of worsening depends on the first recorded value, according to current clinical protocols:
  - if the baseline is EDSS < 1, then the worsening event occurs when an increase of EDSS by 1.5 points is first observed;
  - if the baseline is $1 \leq$ EDSS < 5.5, then the worsening event occurs when an increase of EDSS by 1 point is first observed;
  - if the baseline is EDSS ≥ 5.5, then worsening event occurs when an increase of EDSS by 0.5 points is first observed.

For each sub-task, participants are given a dataset containing 2.5 years of visits, with the occurrence of the worsening event and the time of occurrence pre-computed by the challenge organizers.

3.3 Task 3: Position Papers on the Impact of Exposition to Pollutants (ALS)

Participants in Task 3 are required to propose approaches to assess if exposure to different pollutants is a useful variable to predict time to PEG, NIV, and death in ALS patients. This task is based on the same design as Task 1 in iDPP@CLEF 2022 and employs the same data as well. Therefore, both training and test data are available immediately. Compared to iDPP@CLEF 2022, the dataset is complemented with environmental data to investigate the impact of exposition to pollutants on the prediction of disease progression. The task consists in ranking subjects based on the risk of early occurrence of:

- Task3a: NIV or (competing event) death, whichever occurs first;
- Task3b: PEG or (competing event) Death, whichever occurs first;
- Task3c: Death.

Since test data were already released at the end of iDPP@CLEF 2022 it is impossible to produce a fair leaderboard. Therefore, participants are required to produce position papers in which they describe their approaches and findings concerning the link between environmental factors and ALS progression.

4 Dataset

For iDPP@CLEF 2023, we provided 5 datasets, two for MS and three for ALS, using data from three clinical institutions in Turin and Pavia, Italy, and Lisbon,

Portugal. The datasets are fully anonymized: identifiers and pseudo-identifiers, e.g. place of birth or city of residence, have been removed; dates are reported as relative spans in days with respect to a `Time 0`, i.e., a reference moment in time that depends on the considered disease. For MS, `Time 0` denotes the first visit to assess EDSS after the patient has received a confirmed diagnosis of MS. In the context of ALS, `Time 0` represents the date of the first ALSFRS-R questionnaire.

4.1 Task 1 and Task 2: MS Datasets

Tasks 1 and 2 share the same datasets – each MS dataset corresponds to a specific sub-task (a and b). As training features, we provide:

- Static data, containing information on patient's demographics, diagnostic delay, and symptoms at the onset;
- Dynamic data (2.5 years), containing information on: relapses, EDSS scores, evoked potentials, MRIs, and MS course.

The following data are available as ground-truth:

- The worsening occurrence, as defined in Section 3, expressed as a Boolean variable with 0 meaning "not occurred" and 1 meaning "occurred".
- The time-of-occurrence, expressed as relative delta with respect to `Time 0` in years (also fractions).

Each of dataset contains the following groups of variables:

- `static vars.`, representing static variables associated with a patient. The complete list of available static variables is available at http://brainteaser.dei.unipd.it/challenges/idpp2023/assets/other/ms/static-vars.txt.
- `MS type`, containing information about the MS type and the (relative) date when the MS type has been observed.
- `relapses` consisting of the (relative) initial dates of relapses.
- `EDSS`, containing EDSS scores and the (relative) date when they were recorded.
- `evoked potentials`, reporting the results of evoked potential tests. The complete list of variables for each evoked potential test is available at http://brainteaser.dei.unipd.it/challenges/idpp2023/assets/other/ms/evoked-potentials.txt.
- `MRI`, containing the data involving MRIs; e.g., the area on which MRIs have been performed and the observed lesions. The complete list of variables about MRIs is available at http://brainteaser.dei.unipd.it/challenges/idpp2023/assets/other/ms/mri.txt.
- `outcomes`, detailing the patients' worsening occurrence, together with the time of occurrence. More in detail, `outcomes` contain one record for each patient where:
  - The first column is the patient ID;
  - The second column indicates if the worsening occurred (`1`) or not (`0`).
  - The third column is the time of occurrence, defined as a floating point number in the range `[0,15]`.

Table 1 reports the number of records for each group of variables for training and test sets for each sub-task.

Table 1: Training and test datasets for MS tasks.

Training						
Sub-task	Patients	Relapses	EDSS	Evoked Potentials	MRIs	MS courses
Sub-task a	440	480	2,660	1,210	960	310
Sub-task b	510	552	3,068	1,521	965	324
Test						
Sub-task	Patients	Relapses	EDSS	Evoked Potentials	MRIs	MS courses
Sub-task a	110	94	674	277	236	68
Sub-task b	128	124	812	298	265	74

Creation of the datasets To obtain the iDPP@CLEF 2023 MS datasets, we processed two datasets coming from Turin and Pavia research centres. The source datasets contained approximately 1,800 records linked to patients, with approximately 6,700 records for relapses, 28,600 records on EDSS, 6,200 on evoked potentials, 10,300 on MRIs, and 3,700 on MS courses. To remove minor inconsistencies and typos present in the original data, we first processed the data removing records that were likely wrong or did not provide enough information for AI methods to perform predictions. We removed patients' records without:

- onset date;
- first visit date;
- functional systems scores and corresponding EDSS scores.

Other records associated with such patients (e.g., EDSS or MRIs) have been discarded as well. As for relapses, we removed those records where no information about the relapse was given. We removed MRI records not reporting information about T1 and T2 lesions. After cleaning, to generate the challenge datasets, we restricted visits data to a 2.5 years window prior to `Time 0`.

Split into training and test Each of the two MS datasets underwent a division into a training set and a test set, with proportions of 80% and 20% respectively. In order to ensure a well-stratified distribution of variables across the datasets and to avoid any biases during the splitting process, the data were randomly partitioned 100 times using 100 different random seeds. To assess the appropriateness of the stratification, a comparison of variable distributions was conducted for each training/test pair. Statistical tests were performed on each variable based on its type: the Kruskal-Wallis test [18] was applied to continuous variables, while the Chi-squared test [22] was employed for categorical and ordinal variables. A variable was considered well-stratified depending on the test result. For each split, the percentage of well-stratified variables was calculated using Eq. 1.

$$perc_{well-stratified} = \frac{number\ of\ positive\ tests}{total\ number\ of\ variables} * 100 \quad (1)$$

To identify the split that achieved the best stratification between those that achieved the highest percentage, equal to 97%, a visual inspection was then conducted. Density plots were used for continuous variables, bar plots for categorical and ordinal variables, and Kaplan-Meier curves [20] for the outcome time in the survival setting. A careful examination of the outcome occurrence and time was performed to ensure that the models' performance would not be influenced by the data splitting. Furthermore, special attention was given to sparsely observed levels in categorical variables. The splitting process allowed for the possibility that a rare level might only appear in the training set, but not vice versa.Table 2 and 3 report the comparison of the variables' distributions in the training and test sets for sub-task a[13]. Since the distributions are similar, we concluded that the training/test split provided to the participants met best-practice quality standards.

4.2 Task 3: ALS Dataset

The datasets used for Task 3 in iDPP@CLEF 2023 have the same structure and most of the records as the one used in iDPP@CLEF 2022. There are three datasets concerning patients affected by ALS, Dataset ALSa, Dataset ALSb, and Dataset ALSc. Each dataset concerns a specific type of event that might to patients affected by ALS. Datasets ALSa and ALSb regard respectively the moment in which a patient undergoes NIV or PEG. While dataset ALSc concerns the death of the patient. For a detailed description of the data, cleaning procedures, and additional statistics, please refer to [7, 8].

iDPP@CLEF 2023 dataset extends the previous version by providing participants with environmental data. Furthermore, due to its release at the end of iDPP@CLEF 2022, the ground truth is available to the challenge participants since the beginning of the challenge.

Updates over iDPP@CLEF 2022 In the 2023 version of the dataset, a small subset of patients (less than 50) has been removed from the dataset used for iDPP@CLEF 2022. Indeed, such patients were characterized by the absence of relevant events (i.e., NIV, PEG or death), but did not receive further ALSFRS-R assessments after the first. Therefore, such patients were annotated with the censoring event happening at time 0 making it impossible to provide a sensible prediction. Such patients were removed from the 2023 version of the iDPP@CLEF ALS dataset. Table 4 reports the number of removed patients compared to the original iDPP@CLEF ALS dataset. Notice that, by construction, all the removed patients were labelled with event `NONE`. Spyrometries and ALSFRS-R questionnaires associated with dropped patients have been removed as well.

[13] A more complete and detailed comparison, including the information for the other sub-task, is shown in the extended overview [6].

Table 2: Sub-task a, comparison between training and test populations. Continuous variables are presented as *mean (sd)*; categorical variables as *count (percentage on available data)*, for each level. Demographic and onset-related features.

	Variable	**Level**	**Levels train**	**Levels test**
static vars.	sex	Female	305 (69.32%)	76 (69.09%)
		Male	135 (30.68%)	34 (30.91%)
	residence_classification	Cities	120 (27.27%)	32 (29.09%)
		Rural Area	100 (22.73%)	18 (16.36%)
		Towns	208 (47.27%)	54 (49.09%)
		NA	12 (2.73%)	6 (5.45%)
	ethnicity	Caucasian	424 (96.36%)	99 (90.00%)
		Hispanic	-	4 (3.64%)
		Black_African	-	2 (1.82%)
		NA	16 (3.64%)	5(4.55%)
	ms_in_pediatric_age	FALSE	410 (93.18%)	103 (93.64%)
		TRUE	30 (6.82%)	7 (6.36%)
	age_at_onset	mean (sd)	31 (9.427)	30 (8.775)
	diagnostic_delay	mean (sd)	1029 (1727.8)	967 (1447.6)
		NA	12 (2.73%)	1 (0.91%)
	spinal_cord_symptom	FALSE	348 (79.09%)	83 (75.45%)
		TRUE	92 (20.91%)	27 (24.55%)
	brainstem_symptom	FALSE	305 (69.32%)	79 (71.82%)
		TRUE	135 (30.68%)	31 (28.18%)
	eye_symptom	FALSE	318 (72.27%)	82 (74.55%)
		TRUE	122 (27.73%)	28 (25.45%)
	supratentorial_symptom	FALSE	301 (68.41%)	74 (67.27%)
		TRUE	139 (31.59%)	36 (32.73%)
	other_symptoms	False	431 (97.95%)	107 (97.27%)
		RM+	3 (0.68%)	2 (1.82%)
		Sensory	4 (0.91%)	1 (0.91%)
		Epilepsy	2 (0.45%)	0 (—)
	time_since_onset	mean (sd)	2524 (2448.3)	2446 (2235.9)
MS type	multiple_sclerosis_type	CIS	99 (32.04%)	18 (26.87%)
		RR	210 (67.96%)	49 (73.13%)
	delta_observation_time0	mean (sd)	-718 (210.2)	-715 (237.6)

Table 3: Table 2 contd. Dynamical assessments and outcome features.

	Variable	Level	Levels train	Levels test
edss	edss_as_evaluated_by_clinician	mean (sd)	2 (0.716)	2 (0.655)
		NA	37 (1.39%)	3 (0.45%)
	delta_edss_time0	mean (sd)	-499 (251.6)	-499 (254.4)
evoked potentials	altered_potential	Auditory	280 (23.14%)	58 (20.94%)
		Motor	101 (8.35%)	19 (6.86%)
		Somatosensory	482 (39.83%)	111 (40.07%)
		Visual	347 (28.68%)	89 (32.13%)
	potential_value	mean (sd)	0 (0.401)	0 (0.415)
	location	left	311 (25.70%)	73 (26.35%)
		lower left	126 (10.41%)	29 (10.47%)
		lower right	136 (11.24%)	31 (11.19%)
		right	316 (26.12%)	74 (26.71%)
		upper left	156 (12.89%)	34 (12.27%)
		upper right	165 (13.64%)	36 (13.00%)
	delta_evoked_potential_time0	mean (sd)	-712 (206.3)	-731 (213.3)
relapses	delta_relapse_time0	mean (sd)	-561 (286.1)	-551 (286.5)
magnetic resonance image	mri_area_label	Brain Stem	681 (71.01%)	164 (69.79%)
		Cervical Spinal Cord	62 (6.47%)	25 (10.64%)
		Spinal Cord	201 (20.96%)	36 (15.32%)
		Thoracic Spinal Cord	15 (1.56%)	10 (4.26%)
	lesions_T1	FALSE	175 (18.25%)	45 (19.15%)
		TRUE	149 (15.54%)	29 (12.34%)
		NA	635 (66.21%)	161 (68.51%)
	lesions_T1_gadolinium	FALSE	575 (59.96%)	145 (61.70%)
		TRUE	247 (25.76%)	51 (21.70%)
		NA	137 (14.29%)	39 (16.1%)
	number_of_lesions_T1_gadolinium	mean (sd)	0 (1.0)	0 (1.0)
		NA	187 (19.5%)	48 (20.43%)
	new_or_enlarged_lesions_T2	FALSE	377 (39.31%)	107 (45.53%)
		TRUE	240 (25.03%)	52 (22.13%)
		NA	342 (35.66%)	76 (32.34%)
	number_of_new_or_enlarged_lesions_T2	mean (sd)	1 (1.486)	1 (1.401)
		NA	349 (36.39%)	76 (32.34%)
	lesions_T2	FALSE	55 (5.74%)	10 (4.26%)
		TRUE	275 (28.68%)	62 (26.38%)
		NA	629 (65.59%)	163 (69.36%)
	number_of_total_lesions_T2	0	55 (7.74%)	10 (4.26%)
		1-2	66 (6.88%)	14 (5.96%)
		>=3	70 (7.30%)	14 (5.96%)
		>=9	139 (14.49%)	24 (14.47%)
		NA	629 (65.59%)	163 (69.36%)
	delta_mri_time0	mean (sd)	-512 (282.0)	-534 (275.5)
outcomes	outcome_occurred	0	367 (83.41%)	93 (84.55%)
		1	73 (16.59%)	17 (15.45%)
	outcome_time	mean (sd)	5 (4.4)	5 (4.1)

Table 4: Patients removed from the iDPP@CLEF ALS dataset 2023 due to having an unrealistic censoring event time. Between parentheses the original number of patients available in the dataset.

	Train	Test	Total
Dataset ALSa	22 (orig. 1454)	4 (orig. 350)	26 (orig. 1806)
Dataset ALSb	36 (orig. 1715)	8 (orig. 430)	44 (orig. 2145)
Dataset ALSc	40 (orig. 1756)	8 (orig. 494)	48 (orig. 2250)

Environmental Data One of the primary objectives of iDPP@CLEF 2023 is to promote research on the influence of environmental factors on the progression of ALS disease. Task 3, which specifically focuses on this aspect, requires participants to submit position papers investigating the impact of exposure to pollutants.

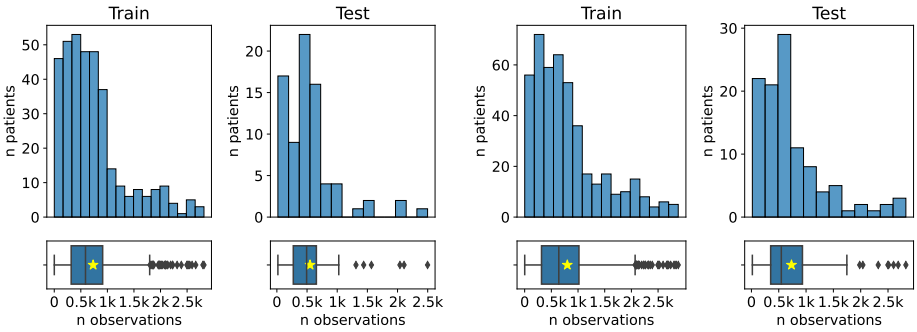
To address this objective, the iDPP@CLEF 2022 datasets have been expanded to include information about patients' exposure to environmental agents. This includes various environmental factors such as daily mean, minimum, and maximum temperatures, daily precipitation, daily averaged sea level pressure and relative humidity, daily mean wind speed, and daily mean global radiation. Additionally, the iDPP@CLEF 2023 ALS datasets also provide information on the concentration of seven pollutants: PM10, PM25, O_3, C_6H_6, CO, SO_2, and NO_2. For each environmental parameter, both the raw observations collected each day and the calibrated version of the observations, following best practices [10, 23], are made available.

It is important to note that not all patients have the same amount of environmental information due to varying diagnosis times and data availability. Several patients could not be associated with environmental data, as their disease progression occurred before public environmental data repositories were established. Approximately 20% of the iDPP@CLEF 2023 ALS datasets, corresponding to 434 to 574 patients, are linked to environmental data.

Considering that the impact of environmental factors may occur well before the diagnosis, we include the maximum amount of available information before `Time 0` for all patients with historical records. Depending on the patient, this corresponds to a maximum of 4 to 6 years of data. However, no more than 6 months of data after `Time 0` are considered. If a patient has more than 180 days of information after the first ALSFRS-R assessment, the subsequent days are excluded from the released dataset.

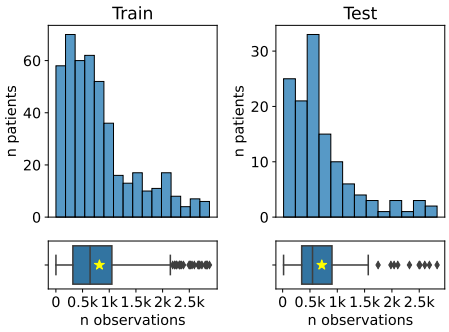
Figure 1 reports the number of patients associated with environmental data as well as the number of records of environmental observations available. It is possible to observe that on average, on the training set, there are 732, 799 and 856 days of observations in the case of Datasets ALSa ALSb, and ALSc respectively. Patients within the test set contain slightly lower numbers of records.

Figure 2 shows the proportion of patients (among those with environmental data) having observations for a given day in (their) history. For example, it is



(a) Dataset ALSa

(b) Dataset ALSb

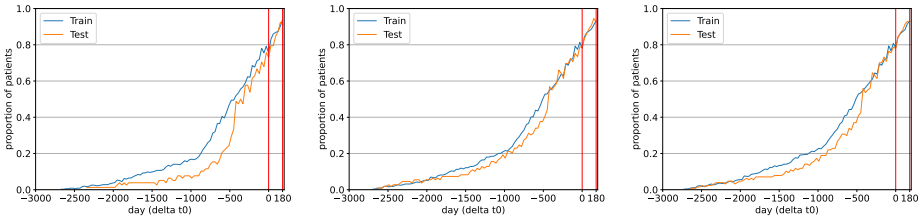


(c) Dataset ALSc

	Dataset ALSa		Dataset ALSb		Dataset ALSc	
	Train	Test	Train	Test	Train	Test
n. patients	356	78	444	109	447	127
n. obs. (q3)	318	272	315	352	318	349
n. obs. (median)	588	493	640	547	641	547
n. obs. (q1)	911	655	1017	934	1050	903
n. obs. (mean)	732	550	799	733	856	716

(d) Number of patients with at least one environmental observation and statistics on the number of observations per patient.

Fig. 1: Statistics on environmental observations available. The star in the boxplots indicates the mean.



(a) Dataset ALSa

(b) Dataset ALSb

(c) Dataset ALSc

Fig. 2: Proportion of patients having environmental observations on a given day in (their relative) time.

possible to observe that roughly 80% of the patients have a record of their `Time 0`, this number grows to approximately 95% if we consider the `Time 180`, the last day for which we release information. Going back in time, we observe that, for roughly 40% of the patients, we have at least 2 years (`Time -730`) of information before their `Time 0`.

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. In the remainder, we describe the guidelines provided to participating teams.

5.1 Guidelines

- The runs should be submitted in the textual format described below;
- Each group can submit a maximum of 10 runs for each subtask, thus amounting to maximum 20 runs for each of Task 1 and Task 2 and 30 runs for Task 3.

Task 1 Run Format

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

```
100619256189067386770484450960632124211 0.897 upd_T1a_survRF
101600333961427115125266345521826407539 0.773 upd_T1a_survRF
102874795308599532461878597137083911508 0.773 upd_T1a_survRF
123988288044597922158182615705447150224 0.615 upd_T1a_survRF
100381996772220382021070974955176218231 0.317 upd_T1a_survRF
...
```

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);
- The second column is the risk score. It is expected to be a floating point number in the range [0, 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.

Task 2 Run Format

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

```
10061925618906... 0.221 0.437 0.515 0.817 0.916 upd_T2b_survRF
10160033396142... 0.213 0.617 0.713 0.799 0.822 upd_T2b_survRF
10287479530859... 0.205 0.312 0.418 0.781 0.856 upd_T2b_survRF
12398828804459... 0.197 0.517 0.617 0.921 0.978 upd_T2b_survRF
10038199677222... 0.184 0.197 0.315 0.763 0.901 upd_T2b_survRF
...
```

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 cumulative probability of worsening between years 0 and 2. It is expected to be a floating point number in the range [0, 1].
- The third column is the cumulative probability of worsening between years 0 and 4. It is expected to be a floating point number in the range [0, 1].
- The fourth column is the cumulative probability of worsening between years 0 and 6. It is expected to be a floating point number in the range [0, 1].
- The fifth column is the cumulative probability of worsening between years 0 and 8. It is expected to be a floating point number in the range [0, 1].
- The sixth column is the cumulative probability of worsening between years 0 and 10. It is expected to be a floating point number in the range [0, 1].
- The seventh 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:

```
0x4bed50627d141453da7499a7f6ae84ab 0.897 upd_T3a_EW6_survRF
0x4d0e8370abe97d0fdedbded6787ebcfc 0.773 upd_T3a_EW6_survRF
0x5bbf2927feefd8617b58b5005f75fc0d 0.773 upd_T3a_EW6_survRF
0x814ec836b32264453c04bb989f7825d4 0.615 upd_T3a_EW6_survRF
0x71dabb094f55fab5fc719e348dffc85x 0.317 upd_T3a_EW6_survRF
...
```

where:

- Columns are separated by a white space;

- The first column is the `patient ID`, a 128 bit hex number (should be considered just as a string);
- The second column is the risk score. It is expected to be a floating point number in the range [0, 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. Since different time windows may be considered, participants are allowed to submit predictions for a variable number of patients. We encourage participants to submit predictions for as many patients as possible. To avoid favoring runs that consider only a few patients, submitted runs will be evaluated based on their correctness as well as the number of patients included. The number of patients included is also reported in the output of the evaluation scripts.

Submission Upload

Runs should be uploaded in 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><a|b|c>_[type_]<freefield>`

where:

- `teamname` is the name of the participating team;
- `T<1|2><a|b|c>` is the identifier of the task the run is submitted to, e.g. `T1b` for Task 1, subtask b;
  - `type` describes the type of run only in the case of Task 3 (it can be omitted for Task 1 and 2). It should be one among:
  - `base` for a baseline run;
  - `EW6` when using environmental data in a time window of 6 months before and after `Time 0`;
  - `EWP` when using environmental data in a time windows chosen by the participant; in this case it is suggested to use `freefield` to provide information about the adopted time window;
- `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_T3a_EW6_survRF`

where:

- `upd` is the University of Padua team;
- `T3a` means that the run is submitted for Task 3, subtask a;
- `EW6` means that environmental data in a time window of 6 months before and after `Time 0` have been used;
- `survRF` suggests that participants have used survival random forests as a prediction method.

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_T3a_EW6_survRF.txt
```

Run Scores

Performance scores for the submitted runs will be returned by the organizers in the `score` folder, which follows the same structure as the `submission` folder.

For each submitted run, participants will find a file named

```
<teamname>_T<1|2|3><a|b|c>_[type_]<freefield>.score.txt
```

where `<teamname>_T<1|2|3><a|b|c>_[type_]<freefield>` matches the corresponding run. The file will contain performance scores for each of the evaluation measures described below. In the above example:

```
upd_T3a_EW6_survRF.score.txt
```

5.2 Participants

Overall, 45 teams registered for participating in iDPP@CLEF but only 10 of them actually managed to submit runs for at least one of the offered tasks. Table 5 reports the details about the participating teams.

Table 6 provides breakdown of the number of runs submitted by each participant for each task and sub-task. Overall, we have received 163 runs with a prevalence of submissions for Task 1 (76 runs), followed by Task 2 (48 runs), and lastly, Task 3 (49 runs).

6 Evaluation Measures

iDPP@CLEF adopted several state-of-the-art evaluation measures to assess the performance of the prediction algorithms, among which:

- *Area Under the ROC Curve (AUC)* [11] to show the trade-off between clinical sensitivity and specificity for every possible cut-off of the risk scores;
- *Harrel's Concordance Index (C-index)* [13] to summarize how well a predicted risk score describes an observed sequence of events.
- *O/E ratio* to assess whether or not the observed event rates match expected event rates in subgroups of the model population.

To ease the computation and reproducibility of the results, scripts for computing the measures are available in the following repository: https://bitbucket.org/brainteaser-health/idpp2023-performance-computation.

Table 5: Teams participating in iDPP@CLEF 2023.

Team Name	Description	Country	Repository	Paper
CompBioMed	Department of Medical Sciences, University of Turin	Italy	https://bitbucket.org/brainteaser-health/idpp2023-compbiomed	Rossi et al. [21]
FCOOL	Faculty of Sciences of the University of Lisbon	Portugal	https://bitbucket.org/brainteaser-health/idpp2023-fcool	Branco et al. [2, 3]
HULAT-UC3M	Polytechnic School Universidad Carlos III de Madrid	Spain	https://bitbucket.org/brainteaser-health/idpp2023-hulat-u3m	Ramos et al. [19]
NeuroTN	Independent Researcher, Sfax	Tunisia	https://bitbucket.org/brainteaser-health/idpp2023-neurotn	Karray [14]
Onto-Med	Ontomed	Bulgaria	https://bitbucket.org/brainteaser-health/idpp2023-onto-med	Asamov et al. [1]
SBB	University of Padua	Italy	https://bitbucket.org/brainteaser-health/idpp2023-sbb	Guazzo et al. [9]
Stefagroup	University of Pavia, BMI lab "Mario Stefanelli"	Italy	https://bitbucket.org/brainteaser-health/idpp2023-stefagroup	Buonocore et al. [4]
SisInfLab_AIBio	Polytechnic University of Bari	Italy	https://bitbucket.org/brainteaser-health/idpp2023-sisinfo-aibio	Lombardi et al. [17]
UHU-ETSI-1	Universidad de Huelva	Spain	https://bitbucket.org/brainteaser-health/idpp2023-uhu-etsi	Not Submitted
UWB	University of West Bohemia	Czech Republic	https://bitbucket.org/brainteaser-health/idpp2023-uwb	Hanzl and Picek [12]

6.1 Task 1: Measures to evaluate the Prediction of the Risk of Disease Worsening (MS)

For Task 1, the effectiveness of the submitted runs is evaluated using Harrell's Concordance Index (C-index) [13]. This score quantifies the model's ability in ranking pairs of observations based on their predicted outcomes. A C-index value of 1 indicates perfect concordance, meaning the model can accurately distinguish between higher and lower-risk individuals. Conversely, a value of 0.5 suggests random guessing, while values below 0.5 indicate a counter-correlation.

6.2 Task 2:Measures to evaluate the Prediction of the Cumulative Probability of Worsening (MS)

The effectiveness of the submitted runs is evaluated with the following measures:

- *Area Under the ROC curve (AUROC)* at each of the time intervals (0-2, 0-4, 0-6, 0-8, 0-10 years);
- O/E Ratio: the ratio of observed to expected events at each of the time intervals (0-2, 0-4, 0-6, 0-8, 0-10 years).

Table 6: Break-down of the runs submitted by participants for each task and sub-task. Participation in Task 3 does not involve submission of runs and it is marked just with a tick.

Team	Task 1		Task 2		Task 3			Total
	a	b	a	b	a	b	c	
CompBioMed	3	3	3	2	—	—	—	11
FCOOL	5	5	—	—	9	9	9	37
HULAT-UC3M	2	1	2	1	—	—	—	6
NeuroTN	—	—	—	—	4	4	4	12
Onto-Med	5	4	5	4	—	—	—	18
SBB	3	3	3	3	—	—	—	12
SisInfLab_AIBio	5	4	5	4	—	—	—	18
Stefagroup	2	—	—	—	—	—	—	2
UHU-ETSI-1	6	7	3	3	—	—	—	19
UWB	9	9	5	5	—	—	—	28
Total	40	36	26	22	13	13	13	163

The *Receiver Operating Characteristic (ROC)* curve is a graphical representation of the model's true positive rate (sensitivity) against the false positive rate (1 - specificity) at different classification thresholds. The AUROC ranges from 0 to 1, where a value of 1 indicates a perfect model that can accurately distinguish between individuals who will experience worsening and those who will not. An AUROC value of 0.5 suggests a model that performs no better than random chance. Therefore, a higher AUROC reflects a better ability of the model to discriminate between different outcomes.

The O/E (Observed-to-Expected) ratio provides a measure of calibration for the model's predictions. It compares the actual number of observed worsening events to the number of events expected based on the model's predictions. Ideally, the O/E ratio should be close to 1, indicating good calibration and alignment between predicted and observed outcomes. A ratio significantly above 1 suggests an overestimation of the number of worsening events, while a ratio below 1 indicates an underestimation. Monitoring the O/E ratio at each time interval allows for assessing the model's calibration performance over time.

To compute the AUROC and O/E Ratio, we applied censoring to the ground truth values using the following schema. Let A, B, C, and D be four subjects, where:

- A experienced the outcome at t_A;
- B was censored at t_A;
- C experienced the outcome at t_3;
- D was censored at t_3.

The scenario is represented in Figure 3.

Table 7 reports the outcome occurrence label and outcome time for each possible scenario of censoring time, which we refer to as t_1, t_2, and t_3. When t_1

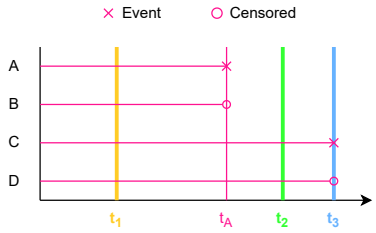


Fig. 3: The set of possible outcomes and censoring time scenarios.

Table 7: Outcome time/occurrence annotation for the example in Figure 3. * indicates that being the outcome of the subject at censoring time t_i unknown, the subject can not be considered for evaluation at censoring time t_i.

		t_1	t_2	t_3
A	outcome time	t_1	t_A	t_A
	outcome occurred	0	1	1
B	outcome time	t_1	NA	NA
	outcome occurred	0	NA*	NA*
C	outcome time	t_1	t_2	t_3
	outcome occurred	0	0	1
D	outcome time	t_1	t_2	t_3
	outcome occurred	0	0	0

is considered as censoring time, all four example subjects have yet to experience the event or be censored, as a result, their outcome occurrence label at this time is set to 0 as shown in the first column of Table 7. When t_2 is considered to perform censoring (second column of Table 7), instead, only subjects C and D have yet to experience either the even or the censoring, and their outcome label is then set to 0. In this scenario, subject A had the event before t_2 and its outcome label is then set to 1. Subject B was censored before t_2 and, as its outcome at this time is unknown, it must be excluded from performance evaluation. Finally, when t_3 is considered to perform censoring (third column of Table 7), outcome labels of subjects A and B are equal to those considered for t_2 since their situation at this time is unchanged compared to the previous one. However, subject C experienced the vent at t_3 and now its outcome label must be set to 1 and subject D was censored at t_3 and its outcome label is then set to 0.

6.3 Task 3: Measures to evaluate the Impact of Exposition to Pollutants (ALS)

The effectiveness of the submitted runs is evaluated with the following measures:

– AUROC: the area under the receiver operating characteristic curve at each of the time intervals (6, 12, 18, 24, 30, 36 months);
– C-index.

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 Risk of Disease Worsening (Multiple Sclerosis)

Figure 4 shows the C-index with its 95% confidence intervals computed for all runs submitted for Task 1 sub-task a and for the random classifier (last row)[14]. Discrimination performance varies across the different submitted runs ranging from 0.4 to above 0.8. Runs submitted by the UWB team [12] lead the pack (C-index > 0.8), followed by CompBioMed (CBMUnitTO) [21], and FCOOL [3]. The best-performing approach for UWB and FCOOL and SisInfLab_AIBio [17] are Survival Random Forests. CompBioMed [21], HULAT [19], and SBB [9] achieve the best performance with Cox regression and CoxNets.

7.2 Task 2: Predicting Cumulative Probability of Worsening (Multiple Sclerosis)

Table 8 presents the AUROC and the O/E ratios, with their 95% confidence intervals computed for all runs submitted for task 2 sub-task a. To avoid clutter-ing, we report the performance obtained for the two-year time window; complete results for subtask a and the results for sub-task b, are shown in the extended overview [6]. As highlighted in Table 8, the approach obtaining the best result in terms of AUCROC corresponds to the run `uwb_T2a_survRFmri`, while the best results for O/E ratio are shown by `uwb_T2a_survGB_minVal`. In general, survival Gradient Boosting approaches proposed by UWB achieve good performance in AUROC, with a good O/E as well.

7.3 Task 3: Position Papers on Impact of Exposition to Pollutants (Amyotrophic Lateral Sclerosis)

Figure 5 shows the C-index and 95% confidence intervals achieved on Task 1 sub-task a[15] by the submitted runs and for the random classifier (last row). As observed by Karray [14] and Branco et al. [2] runs including environmental data (runs tagged with EWP and EW6) tend to perform worse than their counterpart that does not rely on the environmental data. The best-performing approach is provided by the NeuroTN team [14] and corresponds to the classifier ensemble (see subsection 7.4).

[14] Results for sub-task b are available on the extended overview [6].
[15] Results for sub-tasks b and c are available on the extended overview [6].

Table 8: AUROC and OE ratio for all the submitted runs for task 2 subtask a, with a two-year time window. We report the measure as well as the 95% confidence interval.

identifier	**AUROC**	**O/E ratio**
CBMUniTO_T2a_coxnet	0.890 (0.739, 1.000)	0.443 (-0.018, 0.904)
CBMUniTO_T2a_cwgbsa	0.841 (0.618, 1.000)	0.467 (-0.007, 0.940)
CBMUniTO_T2a_evilcox	0.854 (0.655, 1.000)	0.449 (-0.015, 0.913)
HULATUC3M_T2a_survcoxnet	0.864 (0.770, 0.958)	0.437 (-0.021, 0.895)
HULATUC3M_T2a_survRF	0.840 (0.710, 0.969)	0.451 (-0.014, 0.917)
onto-med_T2a_0.01.1.0e-5.10000.100.adj	0.731 (0.482, 0.980)	0.133 (-0.120, 0.386)
onto-med_T2a_0.2.1.0e-5.10000.100	0.696 (0.440, 0.951)	0.269 (-0.090, 0.628)
onto-med_T2a_0.2.1.0e-5.10000.200	0.716 (0.446, 0.987)	0.234 (-0.101, 0.570)
onto-med_T2a_0.2.1.0e-5.5000.100	0.647 (0.399, 0.896)	0.380 (-0.047, 0.807)
onto-med_T2a_0.2.1.0e-5.5000.200	0.590 (0.337, 0.842)	0.358 (-0.057, 0.772)
sbb_T2a_Cox	0.708 (0.491, 0.926)	0.389 (-0.043, 0.821)
sbb_T2a_RSF	0.604 (0.386, 0.822)	0.385 (-0.045, 0.815)
sbb_T2a_SSVM	0.624 (0.461, 0.787)	0.358 (-0.057, 0.772)
sisinflab-aibio_T2a_GB1	0.677 (0.462, 0.893)	0.000 (0.000, 0.000)
sisinflab-aibio_T2a_GB2	0.782 (0.618, 0.945)	0.000 (0.000, 0.000)
sisinflab-aibio_T2a_GB3	0.481 (0.259, 0.703)	0.000 (-0.002, 0.002)
sisinflab-aibio_T2a_RF1	0.754 (0.537, 0.970)	0.017 (-0.073, 0.107)
sisinflab-aibio_T2a_RF2	0.569 (0.347, 0.791)	0.010 (-0.060, 0.081)
uhu-etsi-1_T2a_03	0.769 (0.621, 0.916)	0.678 (0.107, 1.248)
uhu-etsi-1_T2a_04	0.812 (0.690, 0.933)	0.713 (0.128, 1.298)
uhu-etsi-1_T2a_05	0.774 (0.636, 0.912)	0.697 (0.119, 1.276)
uwb_T2a_CGBSA	0.862 (0.731, 0.993)	3.106 (1.885, 4.327)
uwb_T2a_survGB	0.877 (0.745, 1.000)	0.919 (0.255, 1.583)
uwb_T2a_survGB_minVal	0.894 (0.787, 1.000)	**0.946** (0.272, 1.620)
uwb_T2a_survRF	0.914 (0.784, 1.000)	1.811 (0.879, 2.744)
uwb_T2a_survRFmri	**0.924** (0.800, 1.000)	1.889 (0.937, 2.842)

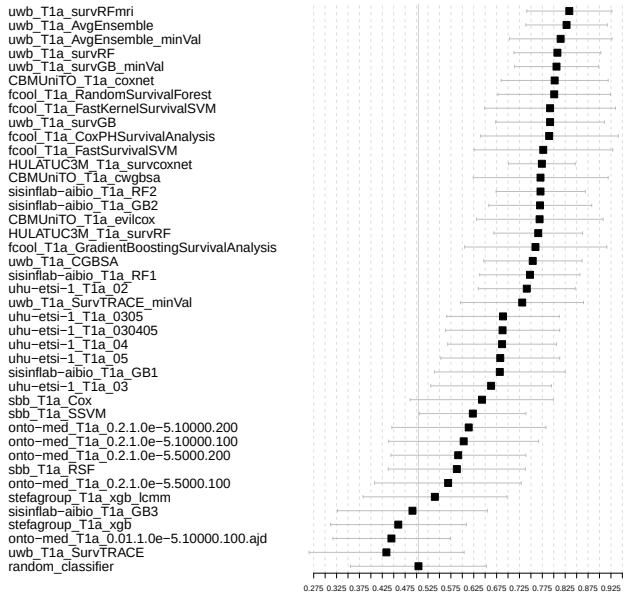


Fig. 4: C-index (with 95% confidence interval) achieved by runs submitted to Task 1a.

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 MS worsening prediction – and one for Task 3 – which concerns the impact of exposition to pollutants on the ALS progression.

Tasks 1 and 2 CompBioMed [21] experiments with CoxNet, Component-wise Gradient Boosting Survival Analysis (CWGBSA), and a hybrid method where the most important features selected by CWGBSA are used to build a CoxNet model (EvilCox). They also test non-linear methods such as Random Survival Forest and Gradient Boosting Survival Analysis, observing a tendency to overfit the training data. To assess the importance of the features, Rossi et al. [21] perform Permutation-based Feature Importance Analysis. In general, they observe that Coxnet is the best-performing approach for all tasks and subtasks. Nevertheless, they also observed that CWGBSA is resistant to over-fitting and aggressive in eliminating features. CWGBSA cross-validated performance is almost on par with that of CoxNet, despite using a smaller set of features.

FCOOL [3] explores several survival prediction methods to rank MS patients according to the risk of worsening. The considered methods are Random Survival Forest, Gradient Boosting, Fast Survival SVM, Fast Kernel Survival SVM, and

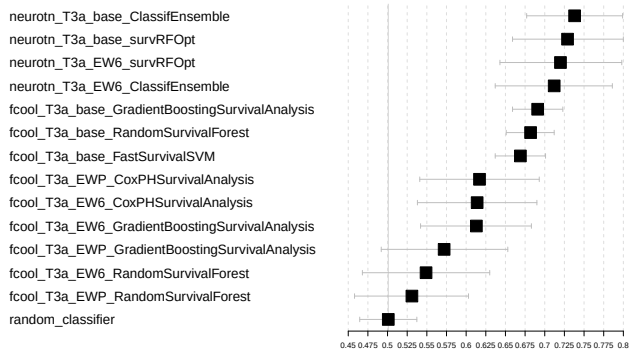


Fig. 5: C-index (with 95% confidence interval) achieved by runs submitted to Task 3a.

the Cox Proportional-Hazards model. A data preprocessing phase is conducted prior to training to manage the temporal nature of patient data by choosing relevant features and by computing additional ones – which capture the temporal progression of the disease. Overall, Random Survival Forest performs best on subtask 1a, whereas Fast Kernel Survival SVM on subtask 1b. Subtask 1b was found to be more complex because of the different definition of the worsening event.

HULAT [19] investigates the effectiveness of Random Survival Forest and Cox regression with Elastic Net regularization (CoxNet) methods on MS worsening prediction. As well as other groups, Ramos et al. [19] perform a data preprocessing phase involving data cleaning, format transformation, normalization, and outliers removal. In particular, the preprocessing step removes all the dynamic features containing a high number of missing values.

Onto-Med [1] develop a Maximum Likelihood Estimation approach to predict MS progression. The proposed method relies on patients' covariates and employs a multi-layer perceptron to approximate the optimal distribution parameters. To handle both tasks, Asamov et al. [1] used the whole training data to build a model and estimate a maximum likelihood distribution for each patient given their features. The method uses a cumulative probability estimate instead of coherent risk measures to accommodate the requirements of bot tasks.

SBB [9] develops different machine-learning approaches to predict a worsening in patient disability caused by MS. Specifically, they consider the following well-known survival analysis approaches: Cox model, random survival forests, and survival support machine. They conclude that these approaches achieve modest performance and that employing non-linear methods does not lead to a discernible advantage with respect to the gold standard Cox model. Nonetheless, they observe that improving data pre-processing may be a key operation

to perform in order to obtain more relevant input features and augment model discrimination with the aim of obtaining satisfactory results.

Stefagroup [4] explores two post-hoc model-agnostic XAI methods, namely SHAP and AraucanaXAI, to provide insights about the most predictive factors of worsening in MS patients. Buonocore et al. [4] evaluate the proposed XAI approaches using commonly adopted measures in XAI for healthcare such as identity, fidelity, separability and time. By leveraging SHAP and AraucanaXAI, the authors gained a deeper understanding of the shortcomings and limitations of their classifiers through feature importance and navigable decision trees.

SisInfLab_AIBio [17] uses Random Survival Forests, an extension of random forests specifically designed for survival analysis, and Boosting Machines for time-to-event analysis. To assess the importance of features for both ML models, the permutation feature importance is computed as well. Lombardi et al. [17] observe that, if the definition of worsening is more complex and condition-dependent (tasks 1b and 2b) significantly lower their approach performs worse than with a simpler definition of worsening (tasks 1a and 2a).

UWB [12] evaluates various ML methods – such as Random Forest and Gradient Boosting – for survival analysis, as well as a Deep Learning survival analysis method based on the Transformer architecture: SurfTRACE. Among the different methods, the authors report top performance with Random Forest. Hanzl and Picek [12] observe that three aspects are instrumental to achieving good performance: (i) data preprocessing, (ii) hyper-parameter tuning, and (iii) validation.

Task 3 FCOOL [2] investigates four models to assess the importance of environmental data in predicting the risk of early occurrence of NIV, PEG or death: Cox Proportional-Hazards, Random Survival Forest, Survival SVM, and Gradient Boosting. Without the introduction of environmental data, the models perform reasonably well. Nevertheless, Branco et al. [2] observe an evident degradation in performance when providing the model with environmental and clinical data in all three tasks. For task A, they observe an even larger degradation when unconstrained amounts of environmental data are provided, compared to what was observed with only 6 months of data. This pattern does not hold for Tasks B and C, where the amount of data does not harm the results, which are, in any case, lower than what was observed without environmental data.

NeuroTN [14] Proposes an approach to stratify patients relying on the disease progression patterns according to features extracted from applying staging systems on visits data. Clusters of patients are then profiled to determine their common characteristics: clinical, demographic and environmental. A second clustering procedure is carried on to detect clusters of patients with similar exposure concentrations to 3 different air pollutants. Then, Karray [14] performs risk prediction on each cluster separately and combines the predictions. In particular Karray [14] relies on two ensembles of classifiers trained on a different data representation (data with Environmental Features and data without Environmental Features). Furthermore, they explored also Survival Random Forests. As

for Branco et al. [2], the introduction of environmental features does not seem to benefit both models and causes performance deterioration.

8 Conclusions and Future Work

The second iteration of iDPP@CLEF focuses on predicting the temporal progression of MS and ALS. In particular, iDPP@CLEF 2023 comprises three tasks. The first two tasks concern MS and participants were provided clinical data and had the objective of predicting the risk of worsening. The third task centres around ALS and builds upon the foundation laid by iDPP@CLEF 2022. This task follows a similar design, involving the prediction of NIV, PEG, or death, but with the addition of environmental data to explore the impact of pollutant exposure on the progression of ALS.

We developed 5 datasets, two for MS and three for ALS, based on the anonymized data provided by three medical institutions in Turin, Lisbon, and Pavia. Out of 45 registered participants, 10 managed to submit a total of 163 runs with a prevalence of submissions for Tasks 1 and 2. Participants adopted a range of approaches, such as Survival Random Forests and Coxnets.

The next iteration of iDPP@CLEF will maintain its dual focus on both ALS and MS. We will extend the amount of available information, by considering also time-series concerning patients' vital parameters produced by wearable devices.

Acknowledgments

The work reported in this paper has been partially supported by the BRAINTEASER[16] project (contract n. GA101017598), as a part of the European Union's Horizon 2020 research and innovation programme.

References

1. Asamov, T., Aksenova, A., Ivanov, P., Boytcheva, S., Taskov, D.: Maximum likelihood estimation with deep learning for multiple sclerosis progression prediction. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
2. Branco, R., Soares, D., Martins, A., Valente, J., Castanho, E., Madeira, S., Aidos, H.: Investigating the impact of environmental data on als prognosis with survival analysis. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
3. Branco, R., Valente, J., Martins, A., Soares, D., Castanho, E., Madeira, S., Aidos, H.: Survival analysis for multiple sclerosis: Predicting risk of disease worsening. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)

[16] https://brainteaser.health/

4. Buonocore, T., Bosoni, P., Nicora, G., Vazifehdan, M., Bellazzi, R., Parimbelli, E., Dagliati, A.: Predicting and explaining risk of disease worsening using temporal features in multiple sclerosis notebook for the idpp lab on intelligent disease progression prediction at clef 2023. In: CLEF 2023 Working Notes (2023)
5. 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)
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., Di Nunzio, G.M., Fariselli, P., García Dominguez, J.M., Gromicho, M., Longato, E., Madeira, S.C., Manera, U., Silvello, G., Tavazzi, E., Tavazzi, E., Vettoretti, M., Di Camillo, B., Ferro, N.: Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. (2023)
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. Guazzo, A., Trescato, I., Longato, E., Tavazzi, E., Vettoretti, M., Camillo, B.: Baseline machine learning approaches to predict multiple sclerosis disease progression. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
10. Hagan, D.H., Isaacman-VanWertz, G., Franklin, J.P., Wallace, L.M.M., Kocar, B.D., Heald, C.L., Kroll, J.H.: Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. Atmospheric Measurement Techniques **11**(1), 315–328 (jan 2018), ISSN 1867-8548, https://doi.org/10.5194/amt-11-315-2018

11. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology **143**(1), 29–36 (1982), pMID: 7063747
12. Hanzl, M., Picek, L.: Predicting risk of multiple sclerosis worsening. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
13. Harrell, Frank E., J., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the Yield of Medical Tests. JAMA **247**(18), 2543–2546 (05 1982), ISSN 0098-7484
14. Karray, M.: Air pollution profiling through patient stratification: Study of als staging systems usefulness in facilitating data-driven disease subtyping and discovery of hazardous ambient air pollutants. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
15. 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)
16. Kurtzke, J.F.: Rating Neurologic Impairment in Multiple Sclerosis. Neurology **33**(11), 1444–1444 (1983), ISSN 0028-3878, https://doi.org/10.1212/WNL.33.11.1444, URL https://n.neurology.org/content/33/11/1444
17. Lombardi, A., De Bonis, L., Fasano, G., Sportelli, A., Colafiglio, T., Lofù, D., Sorino, P., Narducci, F., Di Sciascio, E., Di Noia, T.: Time-to-event interpretable machine learning for multiple sclerosis worsening prediction: Results from idpp@clef 2023. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
18. McKight, P.E., Najab, J.: Kruskal-wallis test. The corsini encyclopedia of psychology pp. 1–1 (2010)
19. Ramos, A., Martínez, P., González-Carrasco, I.: Hulat@iddp clef 2023: Intelligent prediction of disease progression in multiple sclerosis patients. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
20. Rich, J.T., Neely, J.G., Paniello, R.C., Voelker, C.C., Nussenbaum, B., Wang, E.W.: A practical guide to understanding kaplan-meier curves. Otolaryngology—Head and Neck Surgery **143**(3), 331–336 (2010)
21. Rossi, I., Birolo, G., Fariselli, P.: idpp@clef 2023 results from dsm-compbio unito. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) CLEF 2023 Working Notes (2023)
22. Tallarida, R.J., Murray, R.B., Tallarida, R.J., Murray, R.B.: Chi-square test. Manual of pharmacologic calculations: with computer programs pp. 140–142 (1987)
23. Vogt, M., Schneider, P., Castell, N., Hamer, P.: Assessment of low-cost particulate matter sensor systems against optical and gravimetric methods in

a field co-location in norway. Atmosphere **12**(8), 961 (jul 2021), ISSN 2073-4433, https://doi.org/10.3390/atmos12080961