Brainteaser

D9.8 Evaluation challenge: report on the analysis of the experimental results, proceedings, and integration with EOSC (M36)

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

Project Title	BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis
Grant Agreement No	GA101017598
Contract start date	01/01/2021
Contract duration	48 Months

Document ID	BRAINTEASER_D9.8_Evaluation challenge: report on the analysis of the experimental results, proceedings, and integration with EOSC (M36)
Deliverable leader	UNIPD
Due date	31/12/2023
Deliverable date	09/01/2024
Dissemination level	PUBLIC

AUTHORS – CONTRIBUTORS

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

PEER – REVIEWERS

Name	Organization
Giovanni Birolo	UNITO

DOCUMENT HISTORY

Version	Date	Author/Organization	Modifications	Status
0.1	20/09/2023	UNIPD	Initial outline	Draft
0.2	30/09/2023	UNIPD	First draft	Draft
0.3	15/10/2023	UNIPD	Second draft	Draft
0.4	30/10/2023	UNIPD	Formatting	Draft

Version	Date	Author/Organization	Modifications	Status
0.5	15/11/2023	UNIPD	Revising	Pre-final
1.0	09/01/2024	Maria F. Cabrera/UPM	Final review and final version	Final

Disclaimer

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

Copyright message

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Acronym	Meaning
ALS	Amyotrophic Lateral Sclerosis
ALSFRS	ALS Functional Rating Scale
ALSFRS-R	ALSFRS Revised
ESCO	European Skills, Competences, Qualifications and Occupations
FVC	Forced Vital Capacity
NIV	Non-Invasive Ventilation
PEG	Percutaneous Endoscopic Gastrostomy

EXECUTIVE SUMMARY

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 pre-emptively 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 2023 was organised into three tasks, two of which (Tasks 1 and 2) pertained to Multiple Sclerosis (MS), and one (Task 3) concerned the evaluation of the impact of environmental factors in Amyotrophic Lateral Sclerosis (ALS) progression, and how to use environmental data at prediction time. 10 teams took part in the iDPP@CLEF 2023 Lab, submitting a total of 163 runs with multiple approaches to the disease progression prediction task, including Survival Random Forests and Penalized Cox Proportional Hazard Regression Models.

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 patients, caregivers and clinicians alike. Furthermore, the timing of worsening or significant events – such as the need for specific treatments, as the Non-Invasive Ventilation (NIV) or Percutaneous Endoscopic Gastrostomy (PEG) in the case of Amyotrophic Lateral Sclerosis (ALS) – is uncertain and hard to predict. Being able to pre-emptively recognize signals of the worsening of the disease as well as the need for specific medical treatments would have significant implications for the quality of life of patients. Therefore, it us 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 disease progression in a probabilistic, time-dependent way;
- describe better and in an explainable fashion the mechanisms underlying Multiple Sclerosis (MS) and ALS diseases.

In this context, it is crucial to develop shared approaches, promote common benchmarks, and foster experiment comparability and replicability, which is currently not so common in this domain. The Intelligent Disease Progression Prediction at CLEF (iDPP) lab[1] aims to provide an evaluation infrastructure for the development of such AI algorithms. Unlike previous challenges in the field, intelligent Disease Progression Prediction at CLEF (iDPP) 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 also deals with providing clinicians with structured and understandable data. iDPP@CLEF 2023 encompassed three primary tasks, with two focused on MS and one centred around ALS. Concerning MS, these tasks revolved around predicting the risk of incurring a disease worsening, either in terms of probability or as a cumulative probability over increasing time periods. Furthermore, each MS task was further divided into two subtasks, each with its specific definition of worsening. The outcomes for the MS tasks have been notably promising, as participating teams achieved remarkable results, including an impressive AUC of up to 92.4% and an O/E ratio of 0.946. The third task was dedicated to ALS and built upon the tasks explored in iDPP@CLEF 2022. Specifically, participants were asked to predict the occurrence of two essential medical treatments, namely NIV and PEG, as well as the predicted time of death. Each prediction was addressed as a distinct subtask. Notably, for this year's ALS task, participants were provided with environmental data, allowing them to investigate whether incorporating such information could lead to improved predictive models. However, despite the inclusion of environmental data, the models submitted by participants did not demonstrate a statistically significant improvement. This suggests that further exploration and investigation in this domain are necessary to fully understand

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

the potential impact of environmental factors on ALS prediction models. The deliverable is organized as follows: Section 2 describes iDPP@CLEF 2023 tasks; Section 3 briefly introduces the provided dataset; Section 4 and introduces the participants and describes the proposed approaches; finally, Section 5 draws some conclusions and outlooks some future work.

2 TASKS

iDPP@CLEF 2023 was the second iteration of the iDPP lab, expanding its scope to include both ALS and MS in the study of disease progression. The activities in iDPP@CLEF 2023 focused 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 MS progression, focusing on its worsening;
- Position papers (Task 3) on the impact that environmental data can have on ALS progression.

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

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

Task 1 focused on MS and required 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 based on the Expanded Disability Status Scale (EDSS) (Kurtzke 1983), according to clinical standards. We considered 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 ≤ 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 were 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.

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

Task 2 refined 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 considered 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 ≤ 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.

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

Participants in Task 3 were 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 was based on the same design as Task 1 in iDPP@CLEF 2022 and employed the same data as well. Therefore, both training and test data were available immediately. Compared to iDPP@CLEF 2022, the dataset was complemented with environmental data to investigate the impact of exposition to pollutants on the prediction of disease progression. The task consisted 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 were required to produce position papers in which they describe their approaches and findings concerning the link between environmental factors and ALS progression.

3 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 was defined as the time of the last EDSS recorded before the date of the first recorded EDSS plus 2.5 years. Patients that were not diagnosed with MS within the time window going from the first EDSS date to 2.5 years after it had a different definition of Time 0, specifically, the first EDSS for which the patient had a MS diagnosis within 2.5 years was considered instead of the first recorded one for their Time 0 definition. Patients for which it was not possible to find suitable EDSS according to this scheme were excluded from the analysis as it was not possible to correctly define a Time 0 for them. In the context of ALS, Time 0 represents the date of the first ALSFRS-R questionnaire.

3.1 MS Data

The original MS data contained minor inconsistencies and typos. Therefore, to avoid introducing noise and spurious information within datasets, we first processed the data removing records that were likely wrong or did not provide enough information for AI methods to perform predictions. In terms of patients, we removed those where the following pieces of information were absent or out of range: onset date; first visit date; functional systems scores and corresponding EDSS scores. For each removed patient, we discarded all their records related to EDSS, evoked potentials, MRIs, and MS courses. 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. Finally, where needed, we removed duplicated records, records associated with patients without demographic and onset data, or records with missing dates. We removed patients with no clinical EDSS evaluation. Having at least one clinical EDSS evaluation was an inclusion criterion for patients in retrospective data.

3.2 ALS Data

ALS datasets are the same as the ones provided for iDPP@CLEF 2022. Their description is available at (A. Guazzo et al. 2022a, 2022b). Compared to iDPP@CLEF 2022, the ALS datasets used for Task 3 in iDPP@CLEF 2023 have been updated as follows: i) records associated with invalid event date (i.e., patients with censoring time equal to 0) have been removed; ii) environmental data has been added.

3.2.1 Updates over iDPP@CLEF 2022

Table 1. Patients removed from the iDPP 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)

In the 2023 version of the dataset, a small subset of patients (less than 50) has been removed from the dataset used for iDPP 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 ALS dataset. Table 1 reports the number of removed patients compared to the original iDPP 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.

4 PARTICIPANTS AND PROPOSED APPROACHES

We provide in this section some statistics about the participation to iDPP@CLEF 2023 and the approaches proposed by the participants.

4.1 Participants

Overall, 45 teams registered for participating in iDPP but only 10 of them managed to submit runs for at least one of the offered tasks. Table 2 reports the details about the participating teams.

Table 2 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).

The repositories of the participants are publicly available on Zenodo at the following link: https://zenodo.org/records/10210125, with DOI 10.5281/zenodo.10210125.

Table 2. 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

4.2 Approaches

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

4.2.1 Tasks 1 and 2

CompBioMed (Rossi, Birolo, and Fariselli 2023) experiments with Penalized Cox Proportional Hazard Regression Models (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, Birolo, and Fariselli (2023) 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 (Branco, Valente, et al. 2023) 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 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 (Ramos, Martínez, and González-Carrasco 2023) investigates the effectiveness of Random Survival Forest and Cox regression with Elastic Net regularization methods on MS worsening prediction. As well as other groups, Ramos, Martínez, and González-Carrasco (2023) perform a data preprocessing phase involving data cleaning, format transformation, normalization, and outlier removal. In particular, the preprocessing step removes all the dynamic features containing a high number of missing values.

Onto-Med (Asamov et al. 2023) 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. (2023) 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 (Alessandro Guazzo et al. 2023) 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 to obtain more relevant input features and augment model discrimination with the aim of obtaining satisfactory results.

Stefagroup (Buonocore et al. 2023) 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. (2023) 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 (Lombardi et al. 2023) 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. (2023) 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).

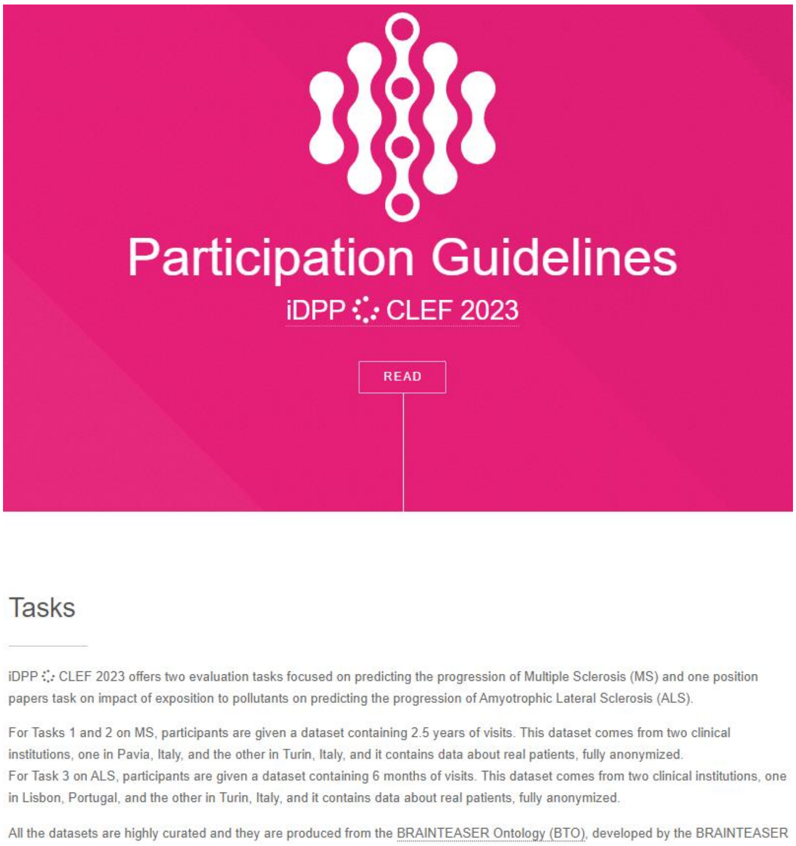
UWB (Hanzl and Picek 2023) evaluates various ML methods – such as Random Forest and Gradient Boosting – for survival analysis, as well as a Deep Learning survival analysis method based on the Transformer architecture: SurfTRACE. Among the different methods, the authors report top performance with Random Forest. Hanzl and Picek (2023) observe that three aspects are instrumental to achieving good performance: (i) data preprocessing, (ii) hyper-parameter tuning, and (iii) validation.

4.2.2 Task 3

FCOOL (Branco, Soares, et al. 2023) 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, Soares, et al. (2023) 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 (Karray 2023) Proposes an approach to stratify patients relying on the disease progression patterns according to features extracted from applying staging systems on clinical evaluation data. Clusters of patients are then profiled to determine their common characteristics: clinical, demographic, and environmental. A second clustering procedure is carried on detecting clusters of patients with similar exposure concentrations to 3 different air pollutants. Then, Karray (2023) performs risk prediction on each cluster separately and combines the predictions. Karray (2023) relies on two ensembles of classifiers trained on a different data representation (data with Environmental Features and data without Environmental Features). Furthermore, they also explored Survival Random Forests. As for Branco, Soares, et al. (2023), the introduction of environmental features does not seem to benefit both models and causes performance deterioration.

5 iDPP@CLEF 2023 WORKSHOP



Tasks

iDPP ⁛ CLEF 2023 offers two evaluation tasks focused on predicting the progression of Multiple Sclerosis (MS) and one position papers task on impact of exposition to pollutants on predicting the progression of Amyotrophic Lateral Sclerosis (ALS).

For Tasks 1 and 2 on MS, participants are given a dataset containing 2.5 years of visits. This dataset comes from two clinical institutions, one in Pavia, Italy, and the other in Turin, Italy, and it contains data about real patients, fully anonymized.
For Task 3 on ALS, participants are given a dataset containing 6 months of visits. This dataset comes from two clinical institutions, one in Lisbon, Portugal, and the other in Turin, Italy, and it contains data about real patients, fully anonymized.

All the datasets are highly curated and they are produced from the BRAINTEASER Ontology (BTO), developed by the BRAINTEASER

Figure 1. Homepage of the iDPP@CLEF 2023 website.

As for the previous years, participants to the iDPP@CLEF challenge had access to the participation guidelines to the iDPP@CLEF website (reachable here: https://brainteaser.dei.unipd.it/challenges/idpp2023).
The challenge had the following schedule:

- Registration closes: April 28, 2023
- Test data release: May 3, 2023
- Runs submission deadline: May 10, 2023
- Evaluation results out: May 26, 2023
- Participant and position paper submission deadline: June 5, 2023
- Notification of acceptance for participant and position papers: June 23, 2023
- Camera-ready participant papers submission: July 7, 2023
- iDPP@CLEF Workshop: September 18-21, 2023 during the CLEF Conference in Thessaloniki, Greece

To foster inclusivity, the iDPP@CLEF workshop was held in dual modality, with both online and physical participants. As for iDPP@CLEF 2022, the challenge was first presented during the plenary session (Figure 1), where one of the organizers of the challenge introduced the tasks and the main objectives, to favour additional participation in future sessions, as well as to disseminate the findings to the scientific community (Figure 2).

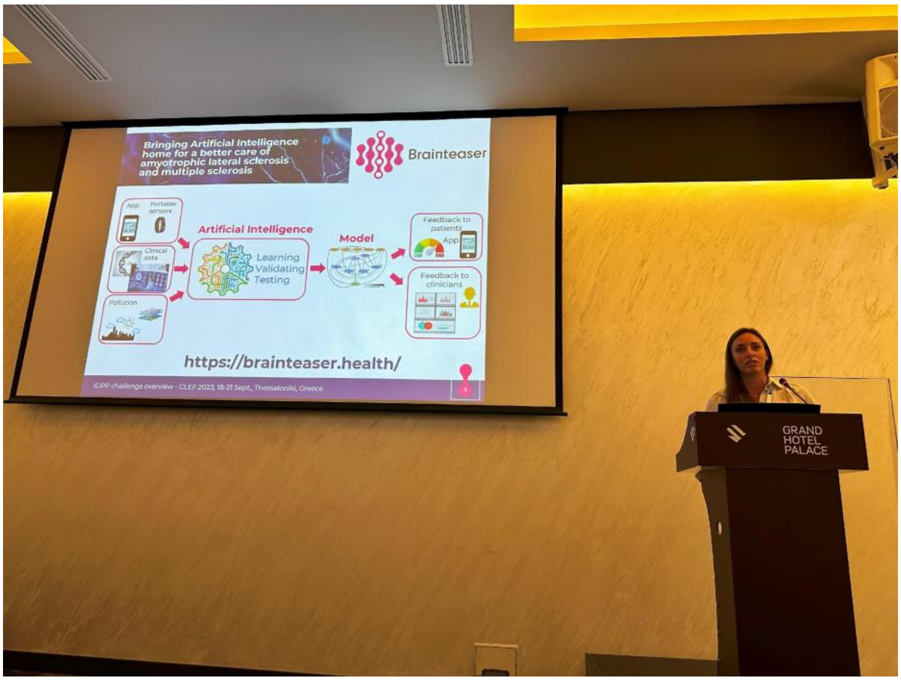
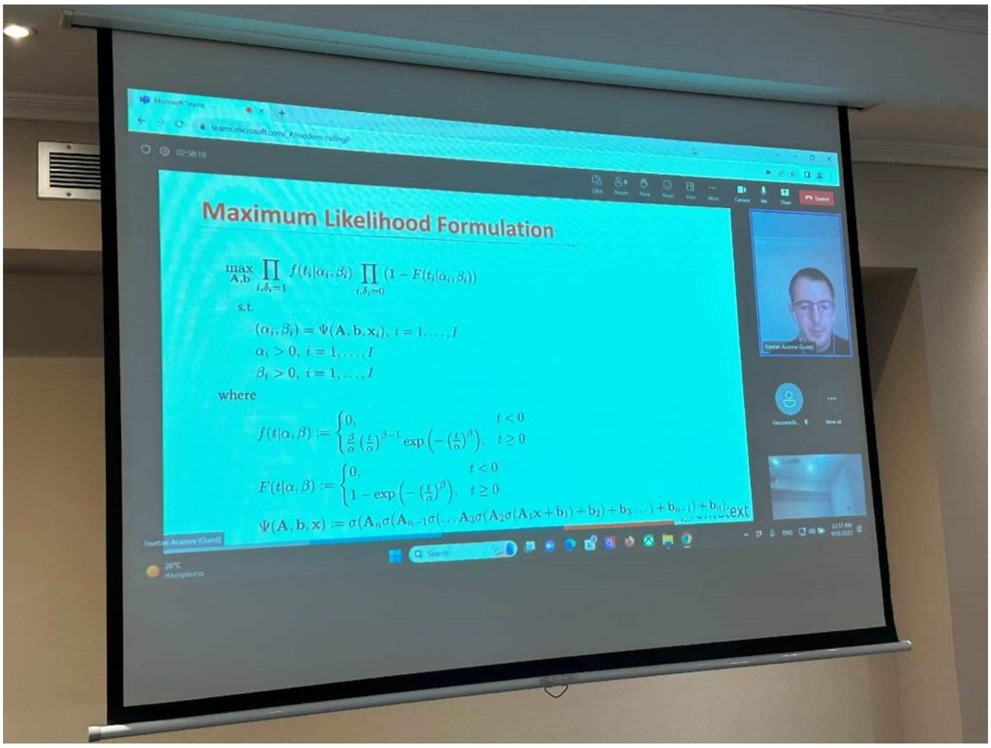

Bringing Artificial Intelligence home for a better care of amyotrophic lateral sclerosis and multiple sclerosis
Brainteaser
App
Portable sensors
Clinical data
Pollution
Artificial Intelligence
Learning
Validating
Testing
Model
Feedback to patients
App
Feedback to clinicians
https://brainteaser.health/
iDPP challenge overview - CLEF 2023, 18-21 Sept, Thessaloniki, Greece
GRAND HOTEL PALACE

Figure 2. An organizer of the iDPP@CLEF challenge describing it during the planary session of the CLEF conference.

Figure 3. The participants to the plenary session of CLEF 2023.

During the iDPP@CLEF session, physical (Figure 3) as well as virtual (Figure 4) participants described their approaches, followed by a lively discussion on what worked about their approaches, and how to improve the predictive capabilities of intelligent disease progression prediction models.

Figure 4. Physical attendees to the iDPP@CLEF 2023 workshop.



Maximum Likelihood Formulation

$$\max_{\mathbf{A},\mathbf{b}} \prod_{i,\delta_i=1} f(t_i|\alpha_i,\beta_i) \prod_{i,\delta_i=0} (1 - F(t_i|\alpha_i,\beta_i))$$

s.t.

$$(\alpha_i,\beta_i) = \Psi(\mathbf{A},\mathbf{b},\mathbf{x}_i),\ i = 1,\dots,I$$
$$\alpha_i > 0,\ i = 1,\dots,I$$
$$\beta_i > 0,\ i = 1,\dots,I$$

where

$$f(t|\alpha,\beta) := \begin{cases} 0, & t < 0 \\ \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}\exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right), & t \geq 0 \end{cases}$$

$$F(t|\alpha,\beta) := \begin{cases} 0, & t < 0 \\ 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right), & t \geq 0 \end{cases}$$

$$\Psi(\mathbf{A},\mathbf{b},\mathbf{x}) := \sigma(\mathbf{A}_n\sigma(\mathbf{A}_{n-1}\sigma(\dots\mathbf{A}_3\sigma(\mathbf{A}_2\sigma(\mathbf{A}_1\mathbf{x}+\mathbf{b}_1)+\mathbf{b}_2)+\mathbf{b}_3\dots)+\mathbf{b}_{n-1})+\mathbf{b}_n)$$

Figure 5. A virtual participant to the iDPP@CLEF 2023 challenge, describing their approach to address tasks 1 and 2.

6 CONCLUSIONS AND FUTURE WORK

The second iteration of iDPP focused on predicting the temporal progression of MS and ALS. IDPP@CLEF 2023 comprised three tasks. The first two tasks concerned MS and participants were provided clinical data and had the objective of predicting the risk of worsening. The third task was centred around ALS and built upon the foundation laid by iDPP@CLEF 2022. This task followed 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 ALS progression.

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 Penalized Cox Proportional Hazard Regression Models.

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

7 REFERENCES

1. Asamov, Tsvetan, Anna Aksenova, Petar Ivanov, Svetla Boytcheva, and Dimitar Taskov. 2023. “Maximum Likelihood Estimation with Deep Learning for Multiple Sclerosis Progression Prediction.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
2. Branco, Ruben, Diogo Soares, Andreia Martins, Joana Valente, Eduardo Castanho, Sara Madeira, and Helena Aidos. 2023. “Investigating the Impact of Environmental Data on ALS Prognosis with Survival Analysis.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
3. Branco, Ruben, Joana Valente, Andreia Martins, Diogo Soares, Eduardo Castanho, Sara Madeira, and Helena Aidos. 2023. “Survival Analysis for Multiple Sclerosis: Predicting Risk of Disease Worsening.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
4. Buonocore, Tommaso Mario, Pietro Bosoni, Giovanna Nicora, Mahin Vazifehdan, Riccardo Bellazzi, Enea Parimbelli, and Arianna Dagliati. 2023. “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*.
5. Cedarbaum, J. M., N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi. 1999. “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.
6. Faggioli, G., A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, et al. 2023. “Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge.” In *CLEF 2023 Working Notes*, edited by M. Aliannejadi, G. Faggioli, N. Ferro, and M. Vlachos. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073.
7. Guazzo, A., I. Trescato, E. Longato, E. Hazizaj, D. Dosso, G. Faggioli, G. M. Di Nunzio, et al. 2022a. “Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022.” In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*, edited by A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C: Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, 395–422. Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany.
8. Guazzo A., et al. 2022b. “Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge.” In *CLEF 2022 Working Notes*, edited by G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, 1130–210. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073.
9. Guazzo, Alessandro, Isotta Trescato, Enrico Longato, Erica Tavazzi, Martina Vettoretti, and Barbara Camillo. 2023. “Baseline Machine Learning Approaches to Predict Multiple Sclerosis Disease Progression.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
10. Hagan, David H., Gabriel Isaacman-VanWertz, Jonathan P. Franklin, Lisa M. M. Wallace, Benjamin D. Kocar, Colette L. Heald, and Jesse H. Kroll. 2018. “Calibration and Assessment of Electrochemical Air Quality Sensors by Co-Location with

Regulatory-Grade Instruments.” *Atmospheric Measurement Techniques* 11 (1): 315–28. https://doi.org/10.5194/amt-11-315-2018.

11. Hanley, J A, and B J McNeil. 1982. “The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve.” *Radiology* 143 (1): 29–36.
12. Hanzl, Marek, and Lukáš Picek. 2023. “Predicting Risk of Multiple Sclerosis Worsening.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
13. Harrell, Jr, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. 1982. “Evaluating the Yield of Medical Tests.” *JAMA* 247 (18): 2543–46.
14. Karray, Mohamed Chiheb. 2023. “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 *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
15. Kurtzke, J. F. 1983. “Rating Neurologic Impairment in Multiple Sclerosis.” *Neurology* 33 (11): 1444–44. https://doi.org/10.1212/WNL.33.11.1444.
16. Küffner, R., N. Zach, R. Norel, J. Hawe, D. Schoenfeld, L. Wang, G. Li, et al. 2015. “Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression.” *Nature Biotechnology* 33 (1): 51–57.
17. Lombardi, Angela, Luigia Maria Natalia De Bonis, Giuseppe Fasano, Alessia Sportelli, Tommaso Colafiglio, Domenico Lofù, Paolo Sorino, Fedelucio Narducci, Eugenio Di Sciascio, and Tommaso Di Noia. 2023. “Time-to-Event Interpretable Machine Learning for Multiple Sclerosis Worsening Prediction: Results from iDPP@CLEF 2023.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
18. McKight, Patrick E, and Julius Najab. 2010. “Kruskal-Wallis Test.” *The Corsini Encyclopedia of Psychology*, 1–1.
19. Ramos, Alberto, Paloma Martínez, and Israel González-Carrasco. 2023. “HULAT@IDDP CLEF 2023: Intelligent Prediction of Disease Progression in Multiple Sclerosis Patients: Report for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2023.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
20. Rich, Jason T, J Gail Neely, Randal C Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W Wang. 2010. “A Practical Guide to Understanding Kaplan-Meier Curves.” *Otolaryngology—Head and Neck Surgery* 143 (3): 331–36.
21. Rossi, Ivan, Giovanni Birolo, and Piero Fariselli. 2023. “iDPP@CLEF 2023 Results from DSM-COMPBIO UNITO.” In *CLEF 2023 Working Notes*, edited by Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos.
22. Tallarida, Ronald J, Rodney B Murray, Ronald J Tallarida, and Rodney B Murray. 1987. “Chi-Square Test.” *Manual of Pharmacologic Calculations: With Computer Programs*, 140–42.
23. Vogt, Matthias, Philipp Schneider, Nuria Castell, and Paul Hamer. 2021. “Assessment of Low-Cost Particulate Matter Sensor Systems Against Optical and Gravimetric Methods in a Field Co-Location in Norway.” *Atmosphere* 12 (8): 961. https://doi.org/10.3390/atmos12080961.

ANNEX A. RELATED CHALLENGES

Within CLEF, there have been no other labs on this or similar topics before the start of iDPP. 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[2] in 2021 and some older ones, the DREAM 7 ALS Prediction challenge[3] in 2012 and the DREAM ALS Stratification challenge[4] 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 (Küffner et al. 2015) 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 ALSFRS-R (Cedarbaum et al. 1999), 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 focuses on explainable AI and on temporal progression of the disease.

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

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

ANNEX B. iDPP@CLEF 2022

iDPP@CLEF 2022 ran as a pilot lab for the first time in CLEF@CLEF 2022[5] (A. Guazzo et al. 2022a, 2022b) 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 (Cedarbaum et al. 1999) 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.

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 2022 and submitted a total of 120 runs.

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

ANNEX C. DATASET CREATION

C.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 2, 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 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 3 reports the number of records for each group of variables for training and test sets for each sub-task.

Table 3. 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 a	110	94	674	277	236	68
Sub-task b	128	124	812	298	265	74

C.1.1 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's test (McKight and Najab 2010) was applied to continuous variables, while the Chi-squared test (Tallarida et al. 1987) 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 the following equation:

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

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 (Rich et al. 2010) 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. For each variable, we enforced the test set to not contain levels that were observed in the training set for the same variable. Table 8 and Table 9 show the comparison of the variables' distributions in the training and test sets for sub-tasks a and b. Since the distributions are similar, we concluded that the training/test split provided to the participants met best-practice quality standards.

Table 4. Variables' distribution for the MS dataset used for sub-task A.

Variable		Level	Levels train	Levels test
static variables	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%)

	Variable	Level	Levels train	Levels test
		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)
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%)

	Variable	Level	Levels train	Levels test
		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	107 (45.53%)	377 (39.31%)
		TRUE	52 (22.13%)	240 (25.03%)
		NA	76 (32.34%)	342 (35.66%)
	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	mean (sd)	629 (65.59%)	76 (32.34%)
	delta_mri_time0	mean (sd)	-512 (282.0)	-534 (275.5)
outcomes	outcome_occurred	FALSE	367 (83.41%)	93 (84.55%)
		TRUE	73 (16.59%)	17 (15.45%)
	outcome_time	mean (sd)	5 (4.4)	5 (4.1)

Table 5. Variables' distribution for the MS dataset used for sub-task B.

	Variable	Level	Levels train	Levels test
static vars.	sex	female	355 (69.61%)	85 (66.41%)
		male	155 (30.39%)	43 (33.59%)
	residence_classification	Cities	152 (29.8%)	37 (28.91%)
		RuralArea	106 (20.78%)	28 (21.88%)
		Towns	236 (46.27%)	56 (43.75%)
		NA	16 (3.14%)	7 (5.47%)
	ethnicity	Caucasian	491 (96.27%)	122 (95.31%)
		Black_African	---	2 (1.56%)
		Hispanic	---	3 (2.34%)
		NA	19 (3.73%)	1 (0.78%)
	ms_in_pediatric_age	FALSE	483 (94.71%)	116 (90.62%)
		TRUE	27 (5.29%)	12 (9.38%)

	Variable	Level	Levels train	Levels test
	age_at_onset	mean (sd)	31 (9.816)	31 (10.642)
	diagnostic_delay	mean (sd)	1094 (1809.46)	1332 (2092)
		NA	9 (1.76%)	5 (3.91%)
	spinal_cord_symptom	FALSE	389 (76.27%)	95 (74.22%)
		TRUE	121 (23.73%)	33 (25.78%)
	brainstem_symptom	FALSE	367 (71.96%)	85 (66.41%)
		TRUE	143 (28.04%)	43 (33.59%)
	eye_symptom	FALSE	370 (72.55%)	95 (74.22%)
		TRUE	140 (27.45%)	33 (25.78%)
	supratentorial_symptom	FALSE	355 (69.61%)	91 (71.09%)
		TRUE	155 (30.39%)	37 (28.91%)
	other_symptoms	epilepsy	2 (0.39%)	---
		FALSE	498 (97.65%)	126 (98.44%)
		sensory	5 (0.98%)	---
		RM+	5 (0.98%)	2 (1.56%)
	time_since_onset	mean (sd)	2871 (2775.14)	3773 (3595)
MSt ype	multiple_sclerosis_type	CIS	108 (33.33%)	22 (29.73%)
		RR	216 (66.67%)	48 (64.86%)
		PR	---	1 (1.35%)
		SP	---	3 (4.05%)
	delta_observation_time0	mean (sd)	-726 (193.54)	-726 (226.50)
edss	edss_as_evaluated_by_clinician	mean (sd)	2 (1.2)	3 (1.7)
		NA	39 (1.27%)	7 (0.86%)
	delta_edss_time0	mean (sd)	-501 (248.58)	-494 (253.84)
evoked potentials	altered_potential	Auditory	341 (22.42%)	68 (22.82%)
		Motor	130 (8.55%)	22 (7.38%)
		Somatosensory	625 (41.09%)	130 (43.62%)
		Visual	425 (27.94%)	78 (26.17%)
	potential_value	FALSE	1193 (78.44%)	237 (79.53%)
		TRUE	328 (21.56%)	61 (20.47%)
	location	left	379 (24.92%)	73 (24.5%)
		lowerleft	167 (10.98%)	37 (12.42%)

	Variable	Level	Levels train	Levels test
		lowerright	177 (11.64%)	36 (12.08%)
		right	387 (25.44%)	73 (24.5%)
		upperleft	201 (13.21%)	40 (13.42%)
		upperright	210 (13.81%)	39 (13.09%)
	delta_evoked_potential_time0	mean (sd)	-714 (196.78)	-656 (252.93)
relapses	delta_relapse_time0	mean (sd)	-561 (280.915)	-595 (279.73)
magnetic resonance image	mri_area_label	BrainStem	688 (71.3%)	188 (70.94%)
		CervicalSpinal Cord	67 (6.94%)	15 (5.66%)
		SpinalCord	191 (19.79%)	57 (21.51%)
		ThoracicSpinal Cord	19 (1.97%)	5 (1.89%)
	lesions_T1	FALSE	155 (16.06%)	37 (13.96%)
		TRUE	164 (16.99%)	56 (21.13%)
		NA	646 (66.94%)	172 (64.91%)
	lesions_T1_gadolinium	FALSE	566 (58.65%)	162 (61.13%)
		TRUE	243 (25.18%)	57 (21.51%)
		NA	156 (16.17%)	46 (17.36%)
	number_of_lesions_T1_gadolinium	mean (sd)	0 (1.049)	0 (0.772)
		NA	222 (23.01%)	57 (21.51%)
	new_or_enlarged_lesions_T	FALSE	363 (37.62%)	116 (43.77%)
		TRUE	222 (23.01%)	55 (20.75%)
		NA	383 (39.69%)	94 (35.47%)
	number_of_new_or_enlarged_lesions_T	mean (sd)	1 (1.54)	1 (1.32)
		NA	383 (39.69%)	94 (35.47%)
	lesions_T	FALSE	61 (6.32%)	12 (4.53%)
		TRUE	256 (26.53%)	65 (24.53%)
		NA	648 (67.15%)	188 (70.94%)
	number_of_total_lesions_T	0	61 (6.32%)	12 (4.53%)
		1-2	57 (5.91%)	12 (4.53%)
		>=3	53 (5.49%)	13 (4.91%)
		>=9	146 (15.13%)	40 (15.09%)

Variable		Level	Levels train	Levels test
		NA	648 (67.15%)	188 (70.94%)
	delta_mri_time0	mean (sd)	-526 (280.304)	-525 (280.263)
outcomes	outcome_occurred	0	384 (75.29%)	97 (75.78%)
		1	126 (24.71%)	31 (24.22%)
	outcome_time	mean (sd)	5 (4.396)	5 (4.396)

C.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 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 (A. Guazzo et al. 2022a, 2022b).

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.

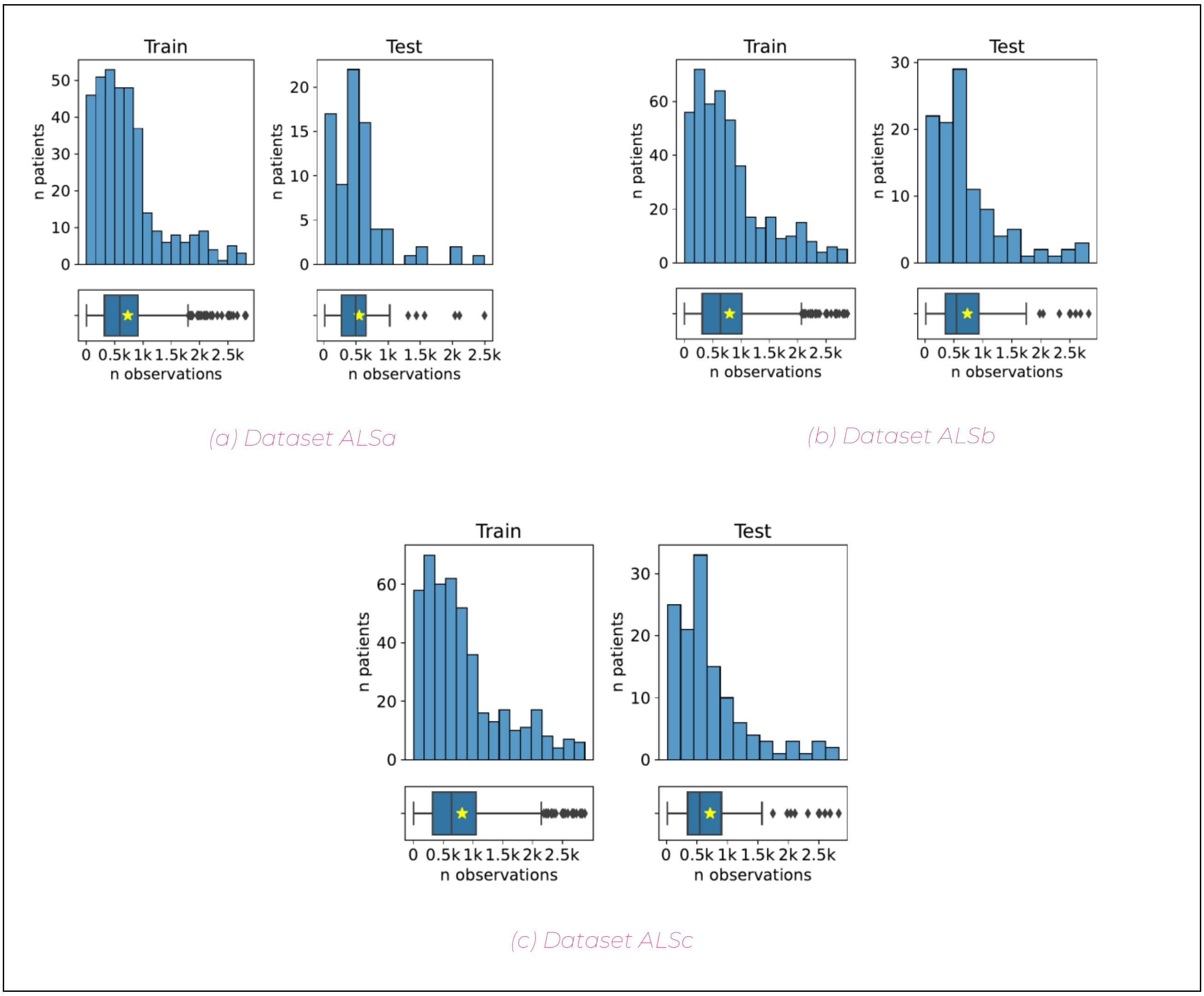
C.2.1 Environmental Data

One of the primary objectives of iDPP@CLEF 2023 is to promote research on the influence of environmental factors on ALS progression. 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 (Vogt et al. 2021; Hagan et al. 2018), 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.

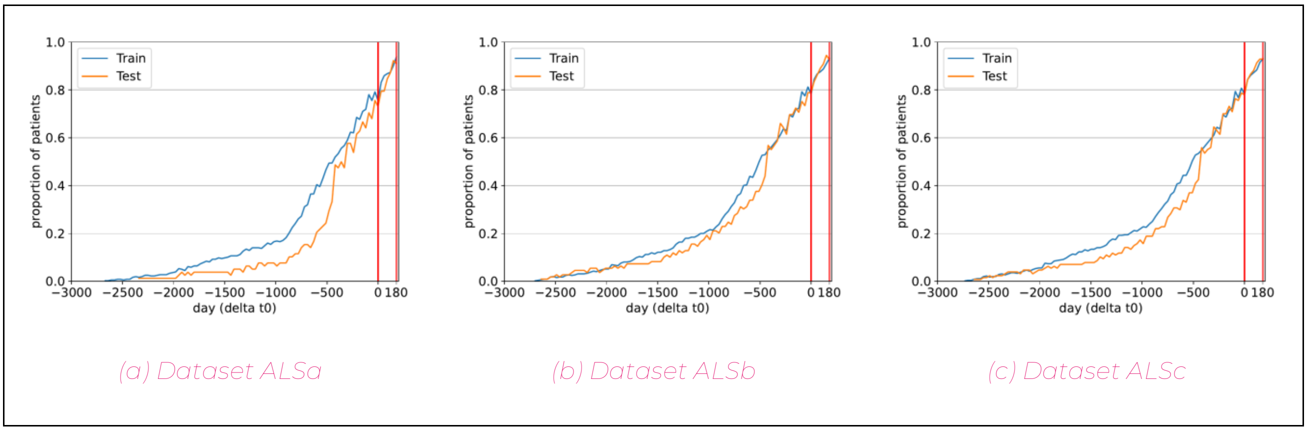


(a) Dataset ALSa

(b) Dataset ALSb

(c) Dataset ALSc

Figure 6. Distributions of the amount of available environmental observations.



(a) Dataset ALSa

(b) Dataset ALSb

(c) Dataset ALSc

Figure 7. Distribution of days covered by environmental data for different patients.

Figure 6reports 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 7 shows the proportion of patients (among those with environmental data) having observations for a given day in (their) history. For example, it is 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.

ANNEX D. 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.

D.1 Task 1 Run Format

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

```
100619256189067386770484450960632124211      0.897      upd_T1a_survRF
1016003339614271151252663455218264O7539      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, a 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.

D2. 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.

D.3 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.

D.4 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

D.5 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

ANNEX E. EVALUATION

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

- *Area Under the ROC Curve (AUC)* (Hanley and McNeil 1982) 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)* (Harrell et al. 1982) 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.

E.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) (Harrell et al. 1982). 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.

E.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).

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 8.

Table 4 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 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 4. When t_2 is considered to perform censoring (second column of Table 6), 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 4), 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.

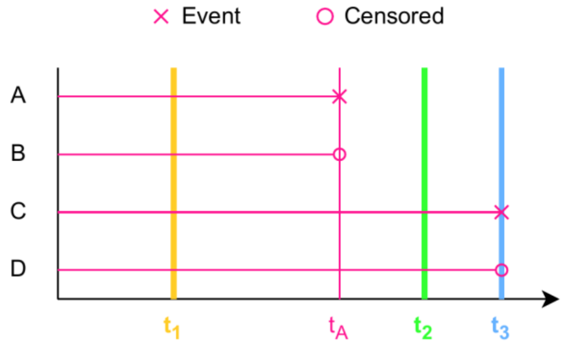


Figure 8. The set of possible outcomes and censoring time scenarios.

*Table 4. Outcome time/occurrence annotation for the example in Figure 3. * indicates that being the outcome of the subject at censoring time ti unknown, the subject cannot be considered for evaluation at censoring time ti.*

		t_1	t_2	t_3
A	outcome time	t_1	t_A	t_A
	outcome occurred	0	1	1

		t_1	t_2	t_3
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

E.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.

ANNEX F. 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 Annex E.

F.1 Task 1: Predicting Risk of Disease Worsening (MS)

Figure 9 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). Discrimination performance varies across the different submitted runs ranging from 0.4 to above 0.8. Runs submitted by the UWB team (Hanzl and Picek 2023) lead the pack (C-index > 0.8), followed by CompBioMed (CBMUnitTO) (Rossi, Birolo, and Fariselli 2023), and FCOOL (Branco, Valente, et al. 2023). The best-performing approach for UWB and FCOOL and SisInfLab_AIBio (Lombardi et al. 2023) are Survival Random Forests. CompBioMed (Rossi, Birolo, and Fariselli 2023), HULAT (Ramos, Martínez, and González-Carrasco 2023), and SBB (Alessandro Guazzo et al. 2023) achieve the best performance with Cox regression and CoxNets.

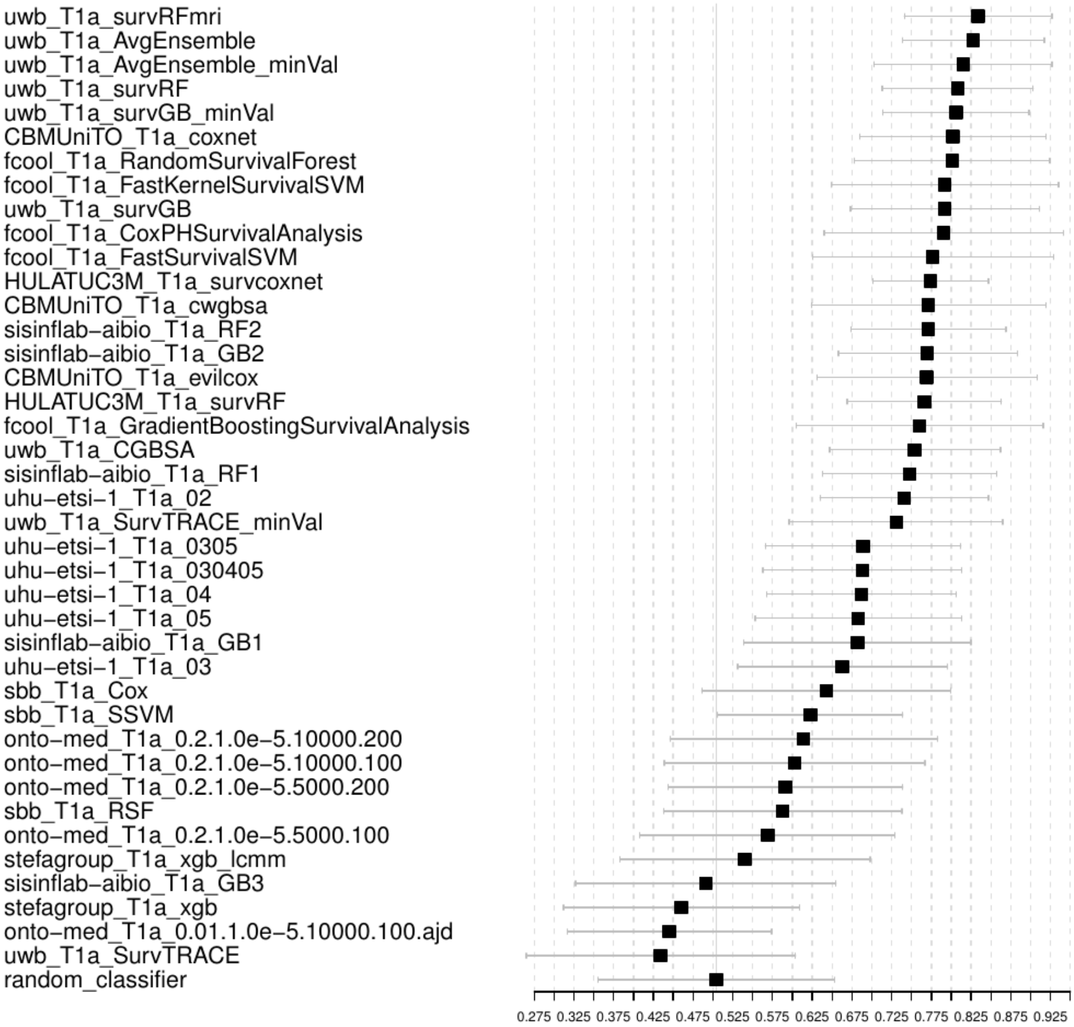


Figure 9. C-index (with 95% confidence interval) achieved by runs submitted to Task 1a.

Figure 10 shows the C-index with its 95% confidence intervals computed for all runs submitted for Task 1 sub-task b and for the random classifier (last row). For this sub-task discrimination performance varies across the different submitted runs ranging from 0.4 to above 0.7. Runs submitted by the FCOOL team (Branco, Valente, et al. 2023) lead the pack (C-index ~ 0.7), followed by CompBioMed (CBMUnitTO) (Rossi, Birolo, and Fariselli 2023), and UWB (Hanzl and Picek 2023). The best-performing approach for FCOOL is a survival SVM. CompBioMed (Rossi, Birolo, and Fariselli 2023), and SBB (Alessandro Guazzo et al. 2023) achieve the best performance with Cox regression and CoxNets. Other methodologic approaches such as gradient boosting or survival random forest show lower performance in this sub-task.

Model performance was overall lower in sub-task b with respect to sub-task a. This observation suggests that, from a model-based perspective and with the available data, the prediction of the crossing of an EDSS threshold (EDSS=3 in this study) may be simpler than the prediction of the worsening of the disease as defined by medical guidelines.

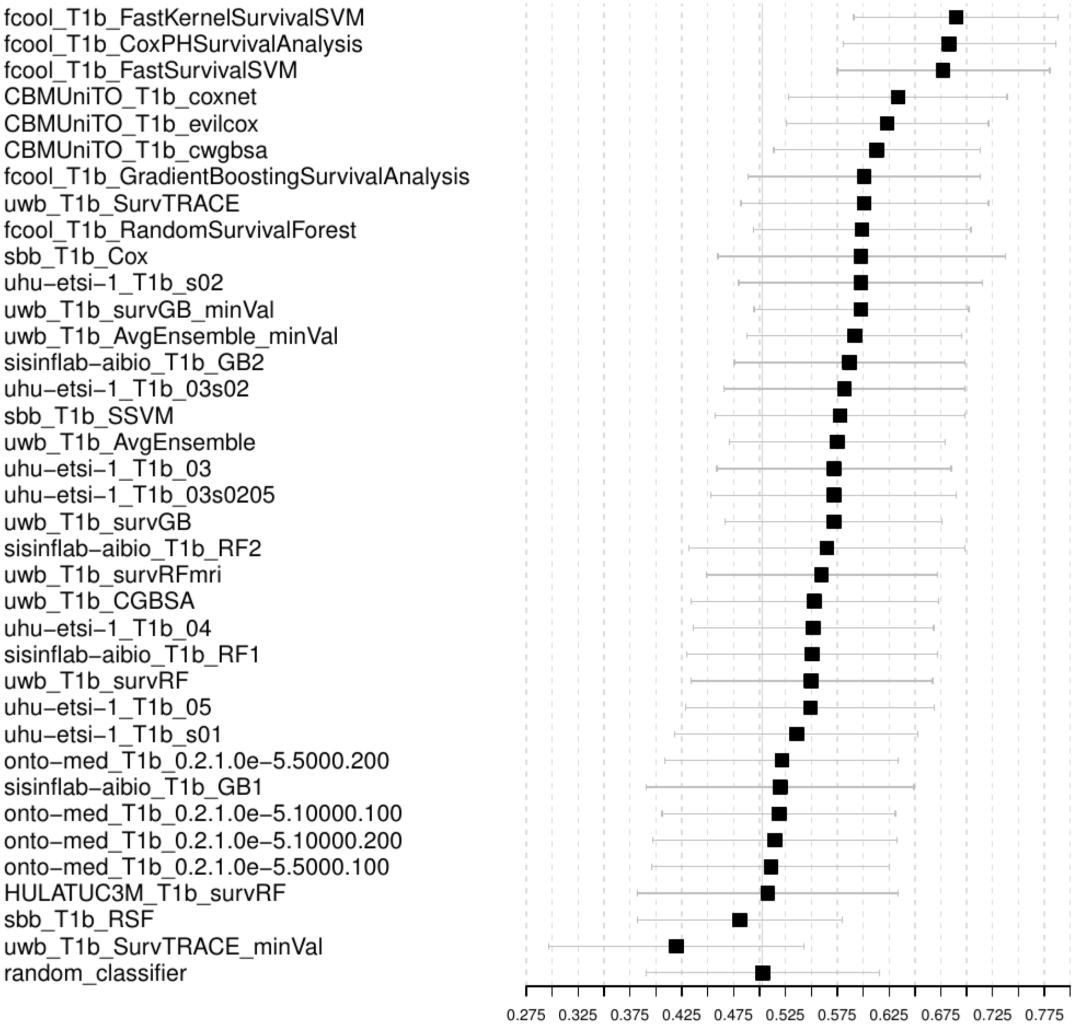


Figure 10. C-index (with 95% confidence interval) achieved by runs submitted to Task 1b.

F.2 Task 2: Predicting Cumulative Probability of Worsening (MS)

F.2.1 Sub-task a

Table 5 presents the AUROC and OE ratio values for a two-year time window. In this time span, the run identified as uwb_T2a_survRFmri achieved the highest AUROC value (0.924). The best O/E ratio of 0.946 is obtained by uwb_T2a_survGB_minVal, indicating a good balance between observed and expected events.

Table 6 shows the performance measures for the same runs but with a four-year time window. Also in this case, uwb_T2a_survRFmri obtains the best AUROC score of 0.907. Regarding the O/E ratio, sisinflab-aibio_T2a_RF2 demonstrates the best balance between observed and expected events with a value of 0.927.

Table 7 displays the performance over a six-year time span. HULATUC3M_T2a_survcoxnet achieves the highest AUROC score of 0.938, while uhu-etsi-1_T2a_04 (0.825) has the best O/E ratio.

Table 8provides the performance measures at eight years. HULATUC3M_T2a_survcoxnet reaches the highest AUROC value of 0.859. In terms of the O/E ratio, uhu-etsi-1_T2a_04 (0.900) achieves the best balance between observed and expected events.

Table 9 reports the performance on the longest time span considered, i.e., at ten years. In this scenario, uwb_T2a_survRFmri (0.839) demonstrates the highest AUROC value among the submitted runs. The identifier with the highest O/E ratio is uhu-etsi-1_T2a_05 (0.816), indicating good calibration.

In Sub-task a, the identifier uwb_T2a_survRFmri consistently achieves the highest AUROC values across multiple time windows, indicating strong predictive performance. Notably, HULATUC3M_T2a_survcoxnet also demonstrates good AUROC scores in longer time spans.

When considering the balance between observed and expected events (O/E ratio), the identifiers uwb_T2a_survGB_minVal and sisinflab-aibio_T2a_RF2 stand out by achieving good equilibrium.

Table 5. 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)

identifier	AUROC	O/E ratio
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)

Table 6. AUROC and OE ratio for all the submitted runs for task 2 subtask a, with a four-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2a_coxnet	0.900 (0.779, 1.000)	0.627 (0.136, 1.117)
CBMUniTO_T2a_cwgbsa	0.864 (0.691, 1.000)	0.638 (0.143, 1.134)
CBMUniTO_T2a_evilcox	0.867 (0.711, 1.000)	0.620 (0.132, 1.109)
HULATUC3M_T2a_survcoxnet	0.898 (0.812, 0.984)	0.599 (0.119, 1.079)
HULATUC3M_T2a_survRF	0.833 (0.711, 0.956)	0.637 (0.142, 1.132)
onto-med_T2a_0.01.1.0e-5.10000.100.adj	0.804 (0.600, 1.000)	0.228 (-0.068, 0.525)
onto-med_T2a_0.2.1.0e-5.10000.100	0.733 (0.522, 0.944)	0.360 (-0.012, 0.732)
onto-med_T2a_0.2.1.0e-5.10000.200	0.760 (0.540, 0.980)	0.316 (-0.033, 0.664)
onto-med_T2a_0.2.1.0e-5.5000.100	0.627 (0.426, 0.827)	0.487 (0.055, 0.920)
onto-med_T2a_0.2.1.0e-5.5000.200	0.622 (0.409, 0.835)	0.460 (0.040, 0.881)
sbb_T2a_Cox	0.762 (0.576, 0.948)	0.577 (0.106, 1.047)
sbb_T2a_RSF	0.644 (0.459, 0.830)	0.604 (0.123, 1.086)
sbb_T2a_SSVM	0.631 (0.466, 0.796)	0.405 (0.010, 0.799)
sisinflab-aibio_T2a_GB1	0.776 (0.601, 0.950)	0.948 (0.344, 1.551)
sisinflab-aibio_T2a_GB2	0.824 (0.698, 0.951)	0.006 (-0.043, 0.055)
sisinflab-aibio_T2a_GB3	0.533 (0.336, 0.731)	0.695 (0.178, 1.212)
sisinflab-aibio_T2a_RF1	0.873 (0.757, 0.990)	0.470 (0.045, 0.895)
sisinflab-aibio_T2a_RF2	0.836 (0.705, 0.966)	0.927 (0.330, 1.524)
uhu-etsi-1_T2a_03	0.716 (0.565, 0.866)	0.842 (0.274, 1.411)
uhu-etsi-1_T2a_04	0.740 (0.590, 0.890)	0.922 (0.327, 1.517)
uhu-etsi-1_T2a_05	0.740 (0.585, 0.895)	0.873 (0.294, 1.452)
uwb_T2a_CGBSA	0.842 (0.713, 0.971)	1.975 (1.104, 2.847)
uwb_T2a_survGB	0.891 (0.796, 0.986)	1.831 (0.993, 2.670)
uwb_T2a_survGB_minVal	0.898 (0.810, 0.985)	1.759 (0.937, 2.581)
uwb_T2a_survRF	0.893 (0.798, 0.989)	2.283 (1.347, 3.220)
uwb_T2a_survRFmri	0.907 (0.816, 0.998)	2.339 (1.391, 3.287)

Table 7. AUROC and OE ratio for all the submitted runs for task 2 subtask a, with a six-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2a_coxnet	0.856 (0.722, 0.991)	0.608 (0.184, 1.031)
CBMUniTO_T2a_cwgbsa	0.821 (0.658, 0.984)	0.619 (0.191, 1.047)
CBMUniTO_T2a_evilcox	0.816 (0.655, 0.978)	0.605 (0.182, 1.027)
HULATUC3M_T2a_survcoxnet	0.938 (0.870, 1.000)	0.529 (0.133, 0.924)
HULATUC3M_T2a_survRF	0.809 (0.667, 0.951)	0.576 (0.164, 0.989)

identifier	AUROC	O/E ratio
onto-med_T2a_0.01.1.0e-5.10000.100.adj	0.687 (0.495, 0.880)	0.284 (-0.005, 0.574)
onto-med_T2a_0.2.1.0e-5.10000.100	0.655 (0.451, 0.859)	0.352 (0.029, 0.675)
onto-med_T2a_0.2.1.0e-5.10000.200	0.702 (0.495, 0.909)	0.317 (0.011, 0.623)
onto-med_T2a_0.2.1.0e-5.5000.100	0.538 (0.351, 0.726)	0.469 (0.097, 0.842)
onto-med_T2a_0.2.1.0e-5.5000.200	0.558 (0.370, 0.746)	0.458 (0.090, 0.826)
sbb_T2a_Cox	0.728 (0.541, 0.916)	0.539 (0.124, 0.954)
sbb_T2a_RSF	0.638 (0.445, 0.830)	0.520 (0.112, 0.929)
sbb_T2a_SSVM	0.643 (0.470, 0.816)	0.357 (0.019, 0.695)
sisinflab-aibio_T2a_GB1	0.727 (0.542, 0.912)	58.75 (54.58, 62.92)
sisinflab-aibio_T2a_GB2	0.824 (0.701, 0.947)	18.06 (15.75, 20.37)
sisinflab-aibio_T2a_GB3	0.531 (0.344, 0.718)	3.783 (2.726, 4.840)
sisinflab-aibio_T2a_RF1	0.871 (0.759, 0.983)	0.493 (0.111, 0.875)
sisinflab-aibio_T2a_RF2	0.856 (0.731, 0.981)	1.373 (0.736, 2.010)
uhu-etsi-1_T2a_03	0.722 (0.561, 0.883)	0.758 (0.285, 1.231)
uhu-etsi-1_T2a_04	0.717 (0.556, 0.878)	0.825 (0.331, 1.319)
uhu-etsi-1_T2a_05	0.774 (0.631, 0.917)	0.777 (0.298, 1.256)
uwb_T2a_CGBSA	0.805 (0.670, 0.941)	1.366 (0.731, 2.002)
uwb_T2a_survGB	0.868 (0.753, 0.984)	1.739 (1.022, 2.455)
uwb_T2a_survGB_minVal	0.901 (0.800, 1.000)	1.768 (1.045, 2.490)
uwb_T2a_survRF	0.898 (0.808, 0.989)	1.796 (1.067, 2.524)
uwb_T2a_survRFmri	0.896 (0.801, 0.991)	1.797 (1.068, 2.525)

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

identifier	AUROC	O/E ratio
CBMUniTO_T2a_coxnet	0.787 (0.626, 0.948)	0.652 (0.244, 1.061)
CBMUniTO_T2a_cwgbsa	0.759 (0.587, 0.931)	0.666 (0.253, 1.079)
CBMUniTO_T2a_evilcox	0.749 (0.570, 0.927)	0.649 (0.242, 1.057)
HULATUC3M_T2a_survcoxnet	0.859 (0.735, 0.983)	0.587 (0.200, 0.975)
HULATUC3M_T2a_survRF	0.710 (0.552, 0.868)	0.653 (0.244, 1.062)
onto-med_T2a_0.01.1.0e-5.10000.100.adj	0.626 (0.446, 0.805)	0.38 (0.068, 0.692)
onto-med_T2a_0.2.1.0e-5.10000.100	0.636 (0.447, 0.825)	0.397 (0.078, 0.715)
onto-med_T2a_0.2.1.0e-5.10000.200	0.664 (0.477, 0.852)	0.366 (0.060, 0.671)
onto-med_T2a_0.2.1.0e-5.5000.100	0.538 (0.355, 0.722)	0.503 (0.144, 0.862)
onto-med_T2a_0.2.1.0e-5.5000.200	0.449 (0.267, 0.630)	0.499 (0.141, 0.856)
sbb_T2a_Cox	0.650 (0.454, 0.847)	0.547 (0.159, 0.934)
sbb_T2a_RSF	0.556 (0.354, 0.759)	0.570 (0.174, 0.965)
sbb_T2a_SSVM	0.697 (0.530, 0.865)	0.328 (0.028, 0.627)
sisinflab-aibio_T2a_GB1	0.690 (0.513, 0.867)	874.2 (859.2, 889.1)
sisinflab-aibio_T2a_GB2	0.795 (0.659, 0.931)	1142 (1125, 1159)

identifier	AUROC	O/E ratio
sisinflab-aibio_T2a_GB3	0.605 (0.420, 0.790)	23.00 (20.57, 25.43)
sisinflab-aibio_T2a_RF1	0.746 (0.586, 0.906)	0.567 (0.186, 0.948)
sisinflab-aibio_T2a_RF2	0.754 (0.591, 0.917)	1.866 (1.175, 2.557)
uhu-etsi-1_T2a_03	0.664 (0.485, 0.843)	0.874 (0.401, 1.347)
uhu-etsi-1_T2a_04	0.672 (0.499, 0.845)	0.900 (0.420, 1.380)
uhu-etsi-1_T2a_05	0.703 (0.530, 0.875)	0.870 (0.398, 1.342)
uwb_T2a_CGBSA	0.747 (0.597, 0.898)	1.312 (0.732, 1.891)
uwb_T2a_survGB	0.790 (0.641, 0.938)	1.906 (1.207, 2.604)
uwb_T2a_survGB_minVal	0.818 (0.677, 0.959)	1.926 (1.224, 2.628)
uwb_T2a_survRF	0.828 (0.702, 0.954)	1.732 (1.066, 2.398)
uwb_T2a_survRFmri	0.838 (0.713, 0.964)	1.731 (1.065, 2.396)

Table 9. AUROC and OE ratio for all the submitted runs for task 2 subtask a, with a ten-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2a_coxnet	0.796 (0.640, 0.952)	0.636 (0.257, 1.016)
CBMUniTO_T2a_cwgbsa	0.765 (0.594, 0.935)	0.643 (0.262, 1.024)
CBMUniTO_T2a_evilcox	0.757 (0.585, 0.929)	0.634 (0.255, 1.012)
HULATUC3M_T2a_survcoxnet	0.831 (0.682, 0.980)	0.582 (0.220, 0.945)
HULATUC3M_T2a_survRF	0.741 (0.567, 0.915)	0.610 (0.239, 0.982)
onto-med_T2a_0.01.1.0e-5.10000.100.adj	0.631 (0.429, 0.834)	0.366 (0.078, 0.653)
onto-med_T2a_0.2.1.0e-5.10000.100	0.682 (0.490, 0.875)	0.383 (0.089, 0.677)
onto-med_T2a_0.2.1.0e-5.10000.200	0.702 (0.518, 0.886)	0.361 (0.075, 0.647)
onto-med_T2a_0.2.1.0e-5.5000.100	0.557 (0.344, 0.770)	0.465 (0.141, 0.789)
onto-med_T2a_0.2.1.0e-5.5000.200	0.404 (0.189, 0.618)	0.456 (0.135, 0.776)
sbb_T2a_Cox	0.608 (0.388, 0.828)	0.522 (0.168, 0.877)
sbb_T2a_RSF	0.568 (0.342, 0.794)	0.491 (0.148, 0.835)
sbb_T2a_SSVM	0.659 (0.446, 0.872)	0.275 (0.018, 0.532)
sisinflab-aibio_T2a_GB1	0.784 (0.624, 0.944)	$> 10^5$ ($> 10^5$, $> 10^5$)
sisinflab-aibio_T2a_GB2	0.749 (0.564, 0.934)	2411 (2387, 2434)
sisinflab-aibio_T2a_GB3	0.510 (0.298, 0.721)	217.3 (210.3, 224.3)
sisinflab-aibio_T2a_RF1	0.745 (0.568, 0.922)	0.587 (0.223, 0.951)
sisinflab-aibio_T2a_RF2	0.698 (0.511, 0.885)	1.877 (1.226, 2.528)
uhu-etsi-1_T2a_03	0.639 (0.437, 0.842)	0.786 (0.365, 1.208)
uhu-etsi-1_T2a_04	0.675 (0.471, 0.878)	0.794 (0.371, 1.218)
uhu-etsi-1_T2a_05	0.722 (0.534, 0.909)	0.816 (0.387, 1.246)
uwb_T2a_CGBSA	0.798 (0.649, 0.947)	1.467 (0.891, 2.042)
uwb_T2a_survGB	0.812 (0.654, 0.969)	1.644 (1.035, 2.254)
uwb_T2a_survGB_minVal	0.808 (0.648, 0.967)	1.658 (1.046, 2.270)
uwb_T2a_survRF	0.820 (0.672, 0.968)	1.458 (0.884, 2.032)
uwb_T2a_survRFmri	0.839 (0.699, 0.979)	1.447 (0.875, 2.019)

F.2.2 Sub-task b

Table 10 and onwards present the AUROC and OE ratio values for all submissions in Task 2, sub-task b.

Within the two-year time frame, the run denoted as CBMUniTO_T2b_coxnet (0.676) achieved the highest AUROC value. The best O/E ratio, equal to 1.019, is obtained by HULATUC3M_T2b_survRF, signifying a favourable balance between observed and expected events.

Table 11 showcases the performance with a four-year time window. In this case, sisinflab-aibio_T2b_GB2 achieves the highest AUROC score of 0.639. Regarding the O/E ratio, sisinflab-aibio_T2b_RF2 maintains the optimal balance between observed and expected events, with a value of 1.005.

Table 12 displays the performance over a six-year time span. CBMUniTO_T2b_coxnet attains the highest AUROC score of 0.635, while uhu-etsi-1_T2b_03 (0.985) demonstrates the best O/E ratio.

Table 13 provides the performance measures at eight years. CBMUniTO_T2b_cwgbsa achieves the highest AUROC value of 0.673. In terms of the O/E ratio, uhu-etsi-1_T2b_03 (1.001) achieves the most desirable balance between observed and expected events.

Table 14 reports the performance over the longest time span considered, i.e., ten years. In this scenario, CBMUniTO_T2b_cwgbsa (0.709) demonstrates the highest AUROC value among the submitted runs. The identifier with the highest O/E ratio is uhu-etsi-1_T2b_03 (1.054), indicating good calibration.

In Sub-task b, the identifier CBMUniTO_T2b_coxnet consistently achieves the highest AUROC values across multiple time windows, indicating its effectiveness in prediction. Additionally, CBMUniTO_T2b_cwgbsa demonstrates strong AUROC scores in eight- and ten-years' time spans.

Regarding the O/E ratio, the runs identified as HULATUC3M_T2b_survRF and uhu-etsi-1_T2b_03 exhibit a favourable balance between observed and expected events in the considered time windows.

Table 10. AUROC and OE ratio for all the submitted runs for task 2 subtask b, with a two-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2b_coxnet	0.676 (0.514, 0.838)	1.082 (0.467, 1.697)
CBMUniTO_T2b_cwgbsa	0.632 (0.477, 0.787)	1.101 (0.481, 1.721)
HULATUC3M_T2b_survRF	0.560 (0.329, 0.791)	1.019 (0.422, 1.615)
onto-med_T2b_0.2.1.0e-5.10000.100	0.604 (0.432, 0.776)	0.585 (0.133, 1.037)
onto-med_T2b_0.2.1.0e-5.10000.200	0.585 (0.433, 0.736)	0.547 (0.110, 0.985)
onto-med_T2b_0.2.1.0e-5.5000.100	0.569 (0.384, 0.754)	1.065 (0.455, 1.675)
onto-med_T2b_0.2.1.0e-5.5000.200	0.523 (0.329, 0.717)	1.035 (0.434, 1.636)
sbb_T2b_Cox	0.642 (0.397, 0.887)	1.098 (0.449, 1.748)
sbb_T2b_RSF	0.514 (0.281, 0.747)	0.966 (0.357, 1.576)
sbb_T2b_SSVM	0.547 (0.345, 0.750)	0.814 (0.255, 1.373)
sisinflab-aibio_T2b_GB1	0.462 (0.249, 0.675)	0.000 (-0.003, 0.003)
sisinflab-aibio_T2b_GB2	0.614 (0.442, 0.786)	0.000 (0.000, 0.000)
sisinflab-aibio_T2b_RF1	0.469 (0.265, 0.672)	0.018 (-0.062, 0.098)
sisinflab-aibio_T2b_RF2	0.535 (0.324, 0.746)	0.011 (-0.052, 0.075)
uhu-etsi-1_T2b_03	0.652 (0.488, 0.816)	1.475 (0.757, 2.193)
uhu-etsi-1_T2b_05	0.630 (0.450, 0.811)	1.328 (0.647, 2.009)
uhu-etsi-1_T2b_s02	0.644 (0.460, 0.827)	1.483 (0.764, 2.203)
uwb_T2b_CGBSA	0.514 (0.311, 0.717)	1.818 (1.021, 2.615)
uwb_T2b_survGB	0.569 (0.392, 0.747)	1.045 (0.441, 1.649)
uwb_T2b_survGB_minVal	0.606 (0.437, 0.776)	0.920 (0.353, 1.486)

identifier	AUROC	O/E ratio
uwb_T2b_survRF	0.590 (0.410, 0.769)	2.292 (1.398, 3.187)
uwb_T2b_survRFmri	0.596 (0.421, 0.770)	2.257 (1.370, 3.145)

Table 11. AUROC and OE ratio for all the submitted runs for task 2 subtask b, with a four-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2b_coxnet	0.633 (0.486, 0.780)	0.858 (0.430, 1.286)
CBMUniTO_T2b_cwgbsa	0.626 (0.484, 0.768)	0.850 (0.424, 1.276)
HULATUC3M_T2b_survRF	0.507 (0.338, 0.675)	0.784 (0.375, 1.193)
onto-med_T2b_0.2.1.0e-5.10000.100	0.500 (0.342, 0.658)	0.464 (0.149, 0.778)
onto-med_T2b_0.2.1.0e-5.10000.200	0.494 (0.346, 0.643)	0.405 (0.111, 0.699)
onto-med_T2b_0.2.1.0e-5.5000.100	0.536 (0.368, 0.704)	0.782 (0.374, 1.191)
onto-med_T2b_0.2.1.0e-5.5000.200	0.531 (0.366, 0.696)	0.725 (0.332, 1.119)
sbb_T2b_Cox	0.567 (0.382, 0.752)	0.807 (0.380, 1.234)
sbb_T2b_RSF	0.529 (0.366, 0.692)	0.711 (0.310, 1.111)
sbb_T2b_SSVM	0.629 (0.468, 0.791)	0.520 (0.177, 0.863)
sisinflab-aibio_T2b_GB1	0.465 (0.299, 0.631)	$> 10^6$ ($> 10^6$, $> 10^6$)
sisinflab-aibio_T2b_GB2	0.639 (0.485, 0.793)	$> 10^5$ ($> 10^5$, $> 10^5$)
sisinflab-aibio_T2b_RF1	0.502 (0.333, 0.672)	0.637 (0.268, 1.006)
sisinflab-aibio_T2b_RF2	0.421 (0.262, 0.581)	1.005 (0.542, 1.469)
uhu-etsi-1_T2b_03	0.578 (0.428, 0.727)	1.064 (0.587, 1.540)
uhu-etsi-1_T2b_05	0.477 (0.314, 0.640)	0.971 (0.516, 1.426)
uhu-etsi-1_T2b_s02	0.590 (0.435, 0.745)	1.076 (0.597, 1.555)
uwb_T2b_CGBSA	0.580 (0.423, 0.737)	0.774 (0.367, 1.180)
uwb_T2b_survGB	0.597 (0.454, 0.741)	1.259 (0.741, 1.778)
uwb_T2b_survGB_minVal	0.612 (0.468, 0.756)	1.228 (0.716, 1.740)
uwb_T2b_survRF	0.552 (0.401, 0.704)	1.523 (0.953, 2.093)
uwb_T2b_survRFmri	0.561 (0.407, 0.715)	1.525 (0.955, 2.096)

Table 12. AUROC and OE ratio for all the submitted runs for task 2 subtask b, with a six-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2b_coxnet	0.635 (0.488, 0.782)	0.811 (0.443, 1.180)
CBMUniTO_T2b_cwgbsa	0.655 (0.512, 0.797)	0.809 (0.441, 1.176)
HULATUC3M_T2b_survRF	0.493 (0.333, 0.653)	0.726 (0.378, 1.074)
onto-med_T2b_0.2.1.0e-5.10000.100	0.508 (0.353, 0.663)	0.464 (0.185, 0.742)
onto-med_T2b_0.2.1.0e-5.10000.200	0.512 (0.358, 0.666)	0.416 (0.153, 0.680)
onto-med_T2b_0.2.1.0e-5.5000.100	0.533 (0.374, 0.692)	0.722 (0.375, 1.069)
onto-med_T2b_0.2.1.0e-5.5000.200	0.482 (0.325, 0.639)	0.660 (0.328, 0.992)
sbb_T2b_Cox	0.601 (0.428, 0.774)	0.782 (0.404, 1.160)
sbb_T2b_RSF	0.511 (0.337, 0.685)	0.695 (0.339, 1.052)
sbb_T2b_SSVM	0.560 (0.386, 0.733)	0.444 (0.159, 0.729)
sisinflab-aibio_T2b_GB1	0.456 (0.294, 0.618)	> 10^8 (> 10^8, > 10^8)
sisinflab-aibio_T2b_GB2	0.629 (0.479, 0.779)	> 10^6 (> 10^6, > 10^6)
sisinflab-aibio_T2b_RF1	0.445 (0.289, 0.600)	0.707 (0.363, 1.050)
sisinflab-aibio_T2b_RF2	0.619 (0.470, 0.767)	1.202 (0.754, 1.651)
uhu-etsi-1_T2b_03	0.566 (0.411, 0.722)	0.985 (0.579, 1.391)
uhu-etsi-1_T2b_05	0.561 (0.398, 0.725)	0.913 (0.522, 1.303)
uhu-etsi-1_T2b_s02	0.610 (0.456, 0.764)	1.008 (0.598, 1.419)
uwb_T2b_CGBSA	0.604 (0.452, 0.756)	1.515 (1.012, 2.017)
uwb_T2b_survGB	0.589 (0.440, 0.737)	1.363 (0.886, 1.840)
uwb_T2b_survGB_minVal	0.602 (0.451, 0.754)	1.375 (0.896, 1.854)
uwb_T2b_survRF	0.549 (0.398, 0.700)	1.351 (0.876, 1.826)
uwb_T2b_survRFmri	0.559 (0.407, 0.711)	1.364 (0.886, 1.841)

Table 13. AUROC and OE ratio for all the submitted runs for task 2 subtask b, with an eight-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2b_coxnet	0.651 (0.503, 0.800)	0.803 (0.465, 1.141)
CBMUniTO_T2b_cwgbsa	0.673 (0.530, 0.816)	0.802 (0.464, 1.140)
HULATUC3M_T2b_survRF	0.607 (0.452, 0.761)	0.746 (0.421, 1.072)
onto-med_T2b_0.2.1.0e-5.10000.100	0.517 (0.362, 0.673)	0.486 (0.223, 0.748)
onto-med_T2b_0.2.1.0e-5.10000.200	0.502 (0.347, 0.657)	0.442 (0.191, 0.692)
onto-med_T2b_0.2.1.0e-5.5000.100	0.594 (0.442, 0.746)	0.715 (0.396, 1.034)
onto-med_T2b_0.2.1.0e-5.5000.200	0.564 (0.412, 0.717)	0.655 (0.350, 0.960)
sbb_T2b_Cox	0.594 (0.419, 0.770)	0.735 (0.392, 1.079)
sbb_T2b_RSF	0.668 (0.496, 0.839)	0.672 (0.344, 1.000)
sbb_T2b_SSVM	0.474 (0.296, 0.651)	0.403 (0.149, 0.657)
sisinflab-aibio_T2b_GB1	0.487 (0.329, 0.645)	> 10^9 (> 10^9, > 10^9)
sisinflab-aibio_T2b_GB2	0.616 (0.463, 0.768)	> 10^8 (> 10^8, > 10^8)
sisinflab-aibio_T2b_RF1	0.471 (0.316, 0.627)	0.803 (0.465, 1.141)
sisinflab-aibio_T2b_RF2	0.600 (0.447, 0.754)	1.442 (0.989, 1.895)

identifier	AUROC	O/E ratio
uhu-etsi-1_T2b_03	0.496 (0.339, 0.652)	1.001 (0.624, 1.379)
uhu-etsi-1_T2b_05	0.494 (0.337, 0.651)	0.951 (0.583, 1.318)
uhu-etsi-1_T2b_s02	0.567 (0.413, 0.721)	1.010 (0.631, 1.389)
uwb_T2b_CGBSA	0.627 (0.477, 0.777)	1.295 (0.866, 1.724)
uwb_T2b_survGB	0.580 (0.427, 0.733)	1.404 (0.957, 1.850)
uwb_T2b_survGB_minVal	0.587 (0.433, 0.742)	1.430 (0.979, 1.880)
uwb_T2b_survRF	0.522 (0.367, 0.678)	1.304 (0.873, 1.734)
uwb_T2b_survRFmri	0.525 (0.369, 0.681)	1.302 (0.872, 1.732)

Table 14. AUROC and OE ratio for all the submitted runs for task 2 subtask b, with a ten-year time window. We report the measure as well as the 95% confidence interval.

identifier	AUROC	O/E ratio
CBMUniTO_T2b_coxnet	0.686 (0.526, 0.847)	0.845 (0.516, 1.174)
CBMUniTO_T2b_cwgbsa	0.709 (0.556, 0.862)	0.850 (0.520, 1.180)
HULATUC3M_T2b_survRF	0.579 (0.420, 0.737)	0.774 (0.460, 1.089)
onto-med_T2b_0.2.1.0e-5.10000.100	0.488 (0.322, 0.654)	0.501 (0.248, 0.755)
onto-med_T2b_0.2.1.0e-5.10000.200	0.523 (0.358, 0.687)	0.461 (0.218, 0.704)
onto-med_T2b_0.2.1.0e-5.5000.100	0.556 (0.397, 0.715)	0.694 (0.396, 0.992)
onto-med_T2b_0.2.1.0e-5.5000.200	0.455 (0.291, 0.618)	0.643 (0.356, 0.930)
sbb_T2b_Cox	0.622 (0.441, 0.803)	0.785 (0.451, 1.120)
sbb_T2b_RSF	0.646 (0.470, 0.821)	0.705 (0.388, 1.021)
sbb_T2b_SSVM	0.541 (0.356, 0.725)	0.397 (0.159, 0.635)
sisinflab-aibio_T2b_GB1	0.482 (0.319, 0.645)	$> 10^{10}$ ($> 10^{10}$, $> 10^{10}$)
sisinflab-aibio_T2b_GB2	0.527 (0.366, 0.689)	$> 10^{8}$ ($> 10^{8}$, $> 10^{8}$)
sisinflab-aibio_T2b_RF1	0.520 (0.358, 0.681)	0.847 (0.518, 1.177)
sisinflab-aibio_T2b_RF2	0.555 (0.393, 0.716)	1.625 (1.169, 2.082)
uhu-etsi-1_T2b_03	0.533 (0.373, 0.694)	1.054 (0.687, 1.422)
uhu-etsi-1_T2b_05	0.541 (0.379, 0.703)	1.045 (0.679, 1.411)
uhu-etsi-1_T2b_s02	0.609 (0.451, 0.767)	1.088 (0.715, 1.462)
uwb_T2b_CGBSA	0.628 (0.463, 0.793)	1.166 (0.780, 1.553)
uwb_T2b_survGB	0.594 (0.430, 0.758)	1.454 (1.022, 1.885)
uwb_T2b_survGB_minVal	0.626 (0.465, 0.787)	1.489 (1.052, 1.926)
uwb_T2b_survRF	0.506 (0.340, 0.672)	1.313 (0.903, 1.724)
uwb_T2b_survRFmri	0.491 (0.324, 0.658)	1.316 (0.905, 1.726)

F.3 Task 3: Position Papers on Impact of Exposition to Pollutants (ALS)

Figure 11 shows the C-index and 95% confidence intervals achieved on Task 3 sub-task a by the submitted runs and for the random classifier (last row). As observed by Karray (2023) and Branco, Soares, et al. (2023) 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 Karray 2023) and corresponds to the classifier ensemble.

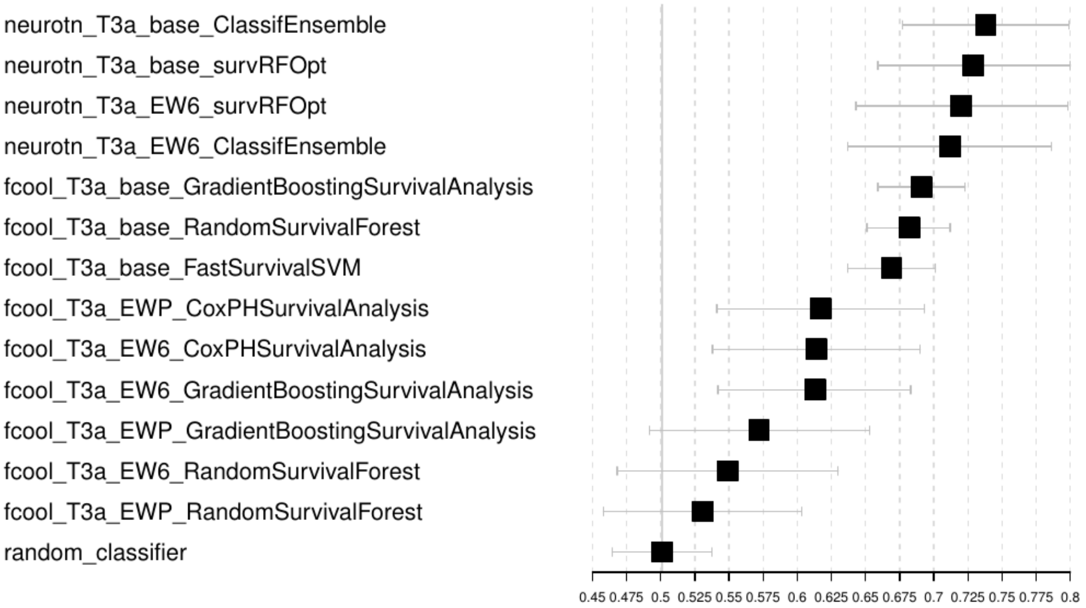


Figure 11. C-index (with 95% confidence interval) achieved by runs submitted to Task 3a.

Figure 12 shows the C-index and 95% confidence intervals achieved on Task 3 sub-task b by the submitted runs and for the random classifier (last row). In this sub-task only runs including environmental data (runs tagged with EWP and EW6) of FCOOL (Branco, Soares, et al. 2023) tend to perform worse than their counterpart that does not rely on the environmental data. Instead, the best-performing approach is provided by the NeuroTN team (Karray 2023) and corresponds to a survival random forest trained on EW6 data.

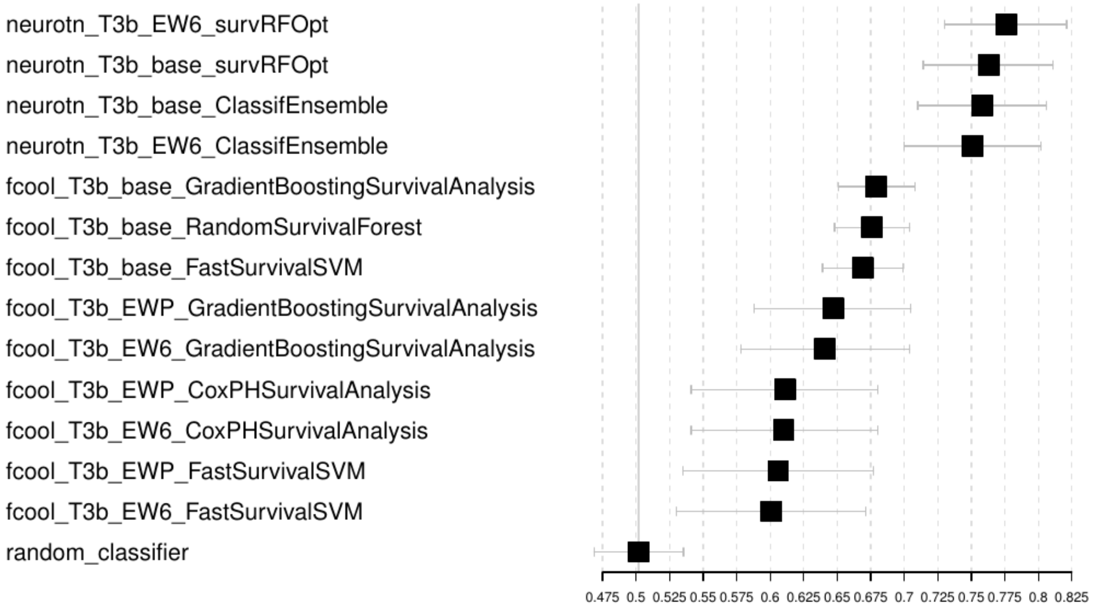


Figure 12. C-index (with 95% confidence interval) achieved by runs submitted to Task 3b.

Similarly, to sub-task a runs including environmental data (EWP, EW6) submitted by both participating teams (FCOOL, NeuroTN) tend to perform worse than their counterpart that does not rely on the environmental data. The best-performing approach is once more provided by the NeuroTN team (Karray 2023) and corresponds to a survival random forest (Figure 13).

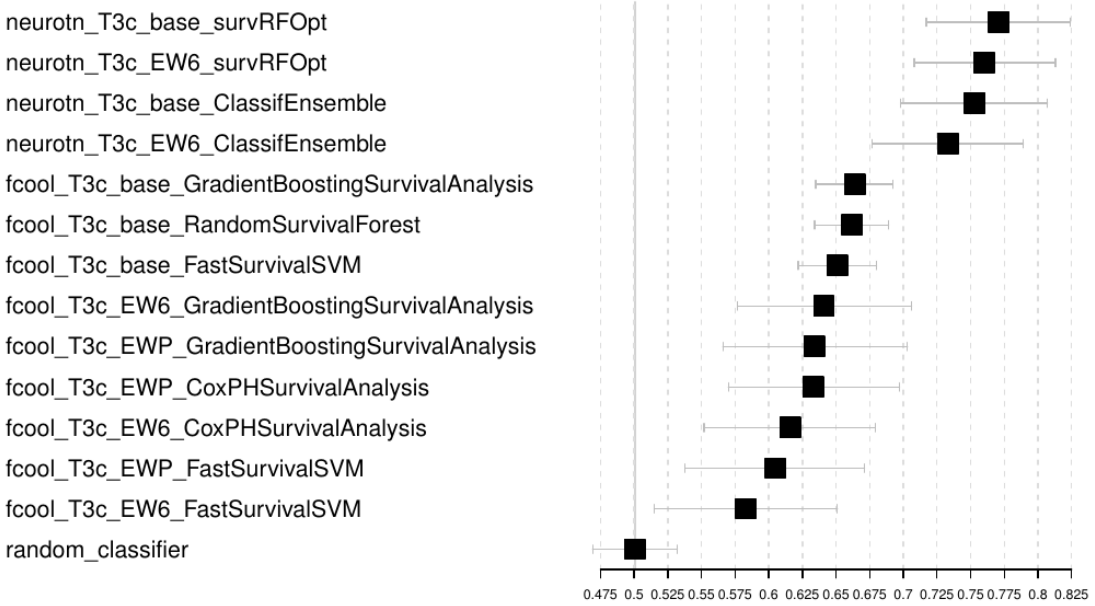


Figure 13. C-index (with 95% confidence interval) achieved by runs submitted to Task 3c.