


UNIVERSIDAD CARLOS III DE MADRID

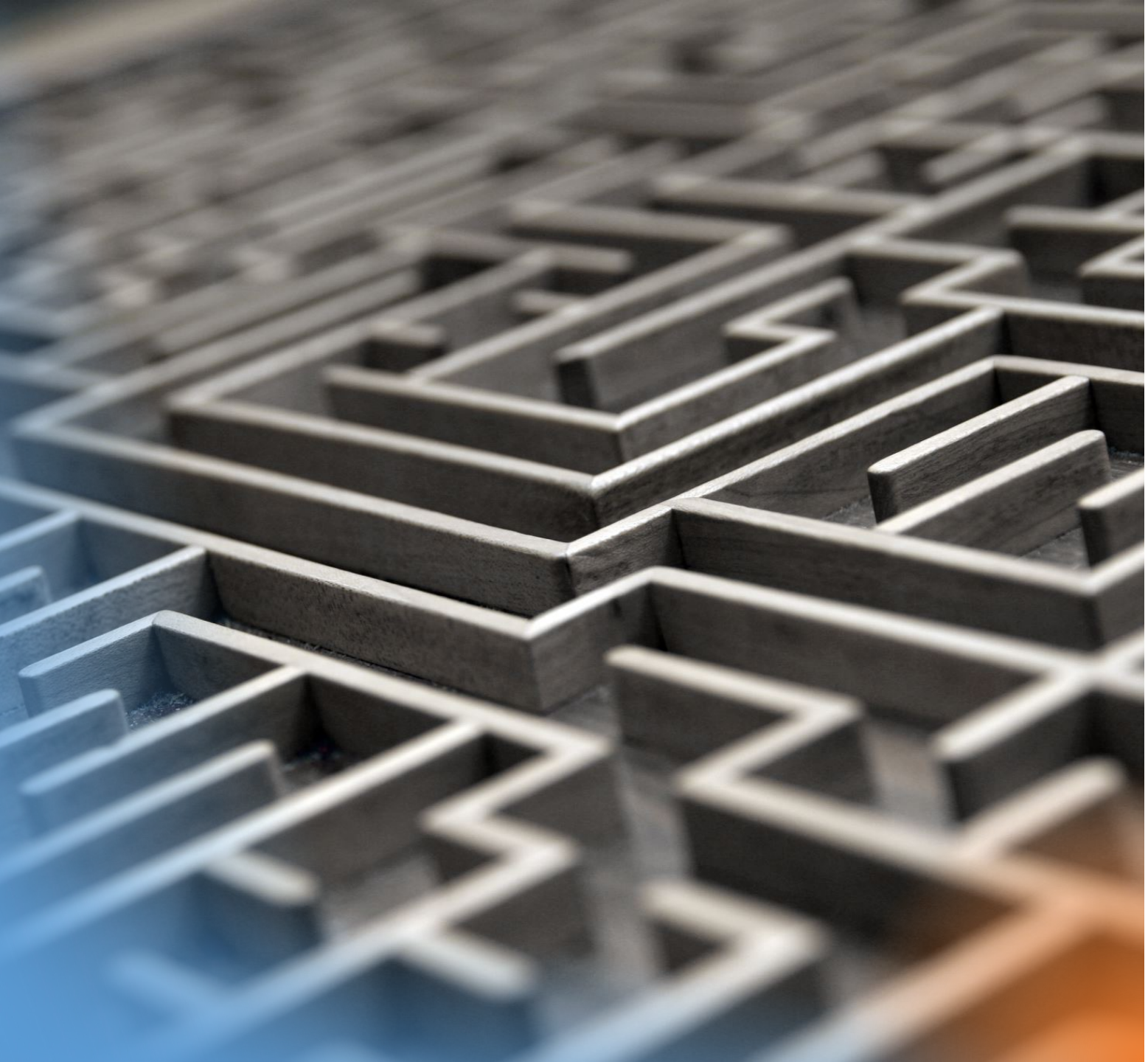
# HULAT@IDDP CLEF 2023: Intelligent Prediction of Disease Progression in Multiple Sclerosis Patients

Alberto Ramos González, Paloma Martínez, Israel González-Carrasco

Computer Science and Engineering Department, Universidad Carlos III de Madrid,
Av. Universidad, 20, 28915 Leganés,
Madrid, Spain

IDPP@CLEF 2023

**This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No GA101017598.**





# Agenda

- Introduction
- Challenges
- Proposed solution
- Results
- Conclusions and future directions

# INTRODUCTION



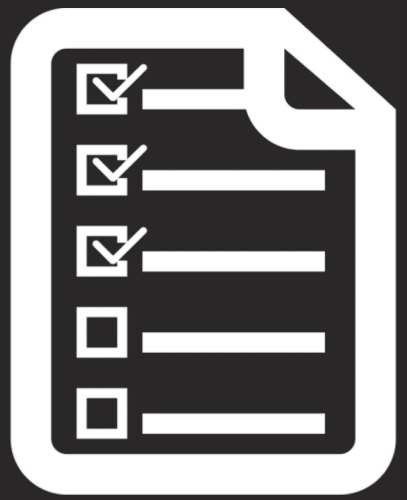
- Multiple Sclerosis (MS) is a chronic disease that causes progressive or alternating deterioration of the patient's neurological functions.
  - Patients alternate periods at the hospital with care at home.
  - This is a challenge due to the different manifestations of the disease, its progression, and the patient's quality of life.
  - **Doctors need support tools for improving detection and prediction**.
- Patient survival analysis is a statistical technique used to model and analyse the time to the event of interest.
  - Different survival analysis methods can be classified into statistical methods or methods based on Machine Learning.



# CHALLENGES

- Find improvements in the accuracy of diagnoses, treatments and prognoses.
- Identification of patterns and prediction of disease progression.
- Support systems for medical staff.
- Improve efficiency in medical care and reduce costs.
- **In no case should it replace medical staff.**

# TASKS



1

**Prediction of the risk of worsening in patients with MS.**

- 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), according to clinical standards.
- **Two subtasks (A and B) regarding different rules of EDSS values.**

2

**Prediction of the cumulative probability of MS worsening in different time windows.**

- Task 2 refines Task 1 by asking participants to explicitly assign the cumulative probability of worsening at different time windows, i.e., between years 0 and 2, 0 and 4, 0 and 6, 0 and 8, 0 and 10.
- **The same two sub-tasks are used as in task 1.**

# PROPOSED SOLUTION

(1) Exploration of the dataset

(2) Feature extraction

( 3) Training

(4) Evaluation

# (1) EXPLORATION OF THE DATASET

- 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, Pavia and Turin (Italy), and it contains data about real patients, fully anonymized.
  - Static data (basic patient information) + Dynamic data (containing information on relapses, EDSS scores, evoked potentials, MRIs, and MS course).
- Ground-Truth:
  - The **worsening-occurrence** is expressed as a Boolean variable with 0 for “not occurred” and 1 for “occurred”.
  - The **time-of-occurrence** is expressed as a relative delta with respect to initial time (time 0).

# (2) EXTRACTION OF CHARACTERISTICS

- Elimination of variables with missing information.
  - Mostly in dynamic data.
  - Patients with missing dynamic data should not be removed. Therefore, features with a high percentage (>40%) of missing data are removed.
  - In the survival analysis, timing is crucial: check for inconsistencies (EDSS-TIME WINDOW).
- Transformation of categorical variables into numerical values.
- Additional features have been introduced to facilitate the worsening prediction process.
  - 17 variables including Num_EDSS, Mean_EDSS, Baseline and Level_EDSS (for subtasks B).

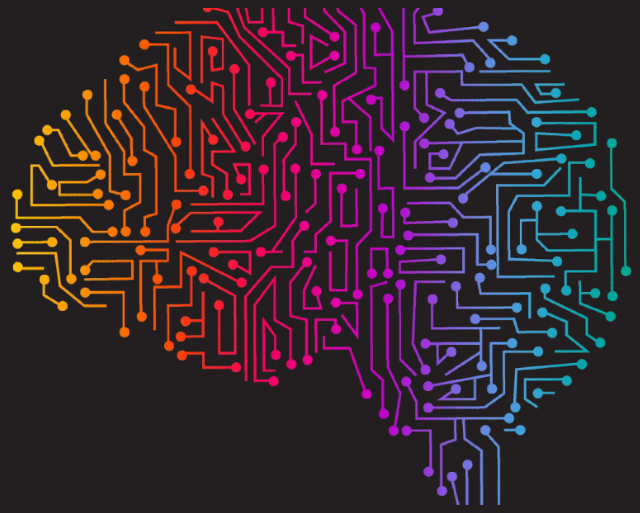
# (3) TRAINING

## Random Survival Forest (RSF): Machine-Learning approach.

- Introduced to extend RF for dealing with **right-censored data.**
- Implementation follows the same general principles as RF.
- Survival function of RSF is the KM (Kaplan-Meier) estimator and is calculated for the terminal nodes.

## Elastic Net Cox: Semi-parametric method.

- For problems with a dataset that integrates a large number of features:
  - LASSO (Least Absolute Shrinkage and Selection Operator) for subset selection of discriminating features.
  - Ridge Penalty for ensuring feature selection by adding a weight to improve the stability of LASSO selection.
  - α sets the equilibrium between the LASSO and Ridge model (by cross-validation).
  - GridSearchCV for hyperparameters.

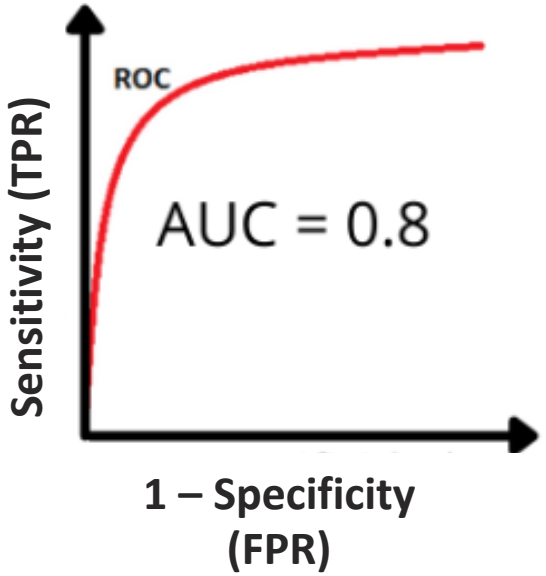


# (4) EVALUATION OF THE MODELS.
## TASK 1

- **C-Index (Harrell's C-index):**
  - Measure to evaluate the model and how well it fits the dataset when making risk predictions.
  - Very common when using data that may be censored, and it is necessary to evaluate risk models in a survival analysis.

# (4) EVALUATION OF THE MODELS. TASK 2

- **ROC curve:** Metric for evaluating the performance of models in solving the classification problem.
  - The ROC curve shows the relationship between the model's True Positive Rate (TPR) and the False Positive Rate (FPR).
    - TPR describes the rate at which the classifier predicts observations that are "positive" as "positive".
    - FPR describes the rate at which the classifier predicts "positive" observations that are actually "negative".
  - Perfect classifiers have a TPR of 1 and an FPR of 0.




- **AUROC (area under the ROC curve):** Metric for assessing the performance of models.
  - Values close to 1 determine a good efficiency, while close to 0.5 means that the model's performance is similar to chance.

# RESULTS. TASK 1

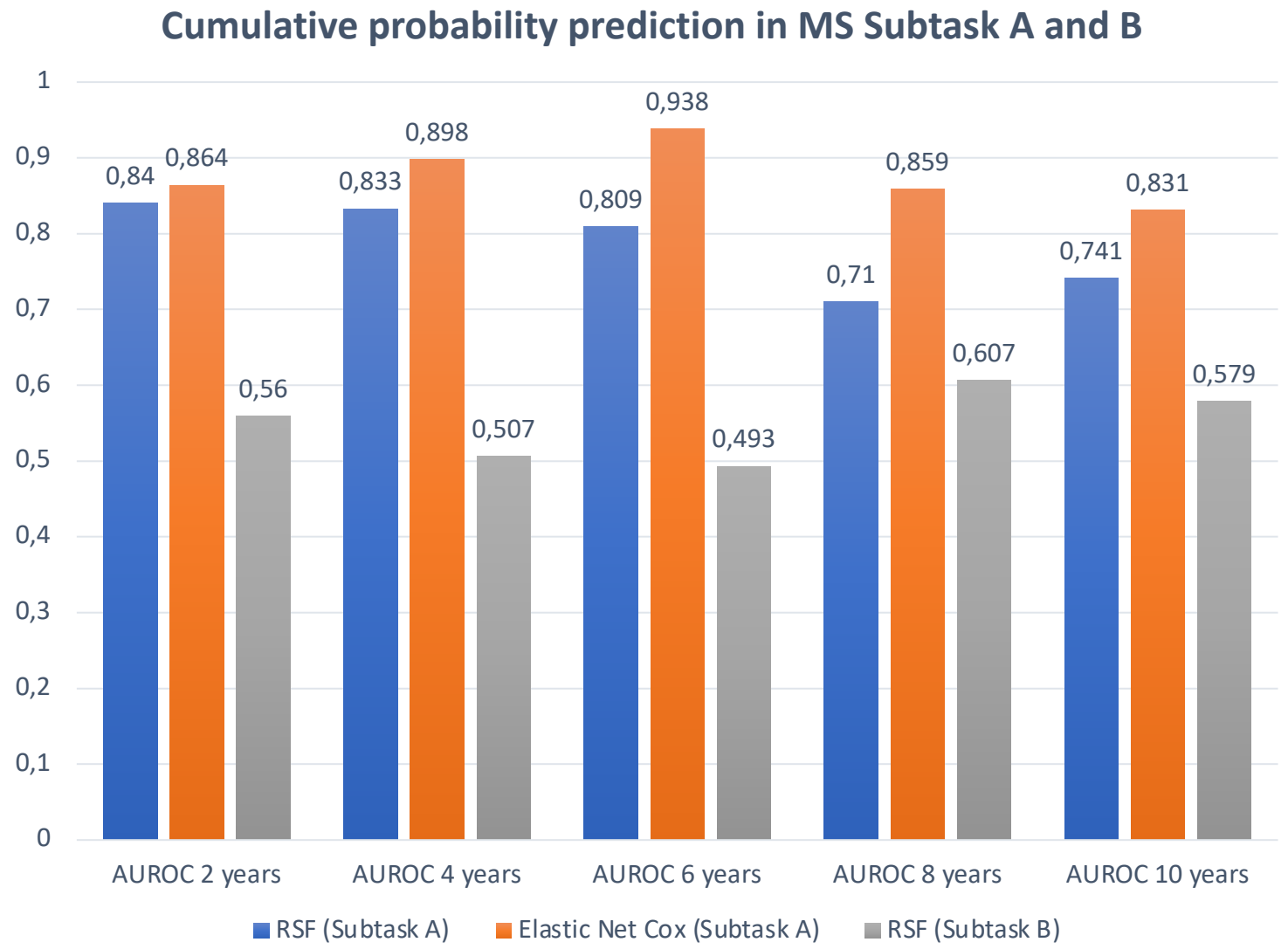
- Results obtained in Subtask A are significantly higher than in Subtask B.
- In future:
  - Test performance of the Elastic Net Cox in subtask B.
  - Obtain an improved dataset for better results (missing values, noise, etc.).
    - E.g., Summarize data, average data over different time windows or assign mean values for missing values in each characteristic.

| Subtask | Model | C-Index |
|---|---|---|
| A | RSF | 0,766 |
| **A** | **Elastic Net Cox** | **0,774** |
| B | RSF | 0,508 |

*Predicting the risk of MS worsening*

# RESULTS. TASK 2

- Best results with Elastic Net Cox in Subtask A.
  - Consistent over time.
- In future:
  - Test performance of the Elastic Net Cox model in subtask B.
  - Obtain an improved dataset for better results (missing values, noise, etc.).



# CONCLUSIONS

- Predict the risk of MS worsening (RSF and Elastic Net Cox models).
- Predict the cumulative probability of MS worsening in different time windows (RSF model).
- Advances in machine learning have demonstrated their potential in survival analysis, like predicting disease progression and identifying new features and patterns in MS.
- The results obtained have been better in subtasks 1A and 2A than in subtasks 1B and 2B, so other approaches can be proposed.
  - Summarize values or include other techniques for dealing with missing or omitted data.

# FUTURE WORK

- Work with larger datasets (include features removed due to missing information).
- Study different techniques to reduce the loss of patients due to lack of information.
- Stratifications could be performed in the population of patients with MS to find new patterns and to improve the understanding and diagnosis of the disease.
- Include more models for solving the problems (such as Elastic Net Cox for subtasks B).

# CONCLUSIONS

Contribution:

Alberto Ramos, Paloma Martínez and Israel González-Carrasco. **HULAT@IDDP CLEF 2023: Intelligent Prediction of Disease Progression in Multiple Sclerosis Patients**. CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece.