Baseline Machine Learning Approaches To Predict Multiple Sclerosis Disease Progression

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2023

Alessandro Guazzo[1]*, Isotta Trescato[1]*, Enrico Longato[1], Erica Tavazzi[1], Martina Vettoretti[1] and Barbara Di Camillo[1,3]

[1] *Department of Information Engineering, University of Padova, Padova, Italy*

[3] *Department of Comparative Biomedicine and Food Science, University of Padova, Padova, Italy*

* *These authors contributed equally*

Abstract

Multiple Sclerosis (MS) is a chronic disease that causes the disruption of the ability of the nervous system to transmit signals thus resulting in a progressive neurologic impairment. Unfortunately, MS onset and progression are extremely heterogeneous across patients that tend to have significantly different management and treatment needs. This heterogeneity in disease progression is among the main aspects of MS that hinder efforts to assess the efficacy of developmental treatments designed to delay disease progression, improve the patient's quality of life, and prolong survival. Consequently, the prediction of disease progression has vastly gained interest among researchers in recent years, with the main aims of deriving new relevant insights into disease mechanisms and manifestations and enabling better treatment development. However, this crucial point has not yet been sufficiently addressed mostly due to insufficient access to rich clinical datasets and effective methodologies. Developed in the context of the iDPP@CLEF 2023 challenge, this work aims at developing different machine-learning approaches to predict a worsening in patient disability caused by MS using a shared dataset provided by the challenge organisers. Results were modest (C-Index and AUROC ~ 0.6) and employing non-linear methods did not lead to a discernible advantage with respect to the well-known Cox proportional hazard model. Exploring alternative, more sophisticated, machine learning techniques or improving data pre-processing to obtain more relevant input features may help in augmenting model discrimination and obtaining satisfactory results.

Keywords

Cox Proportional Hazard Model, Survival Support Vector Machine, Random Survival Forest, Multiple Sclerosis

1. Introduction

Multiple Sclerosis (MS) is a chronic autoimmune disease that causes the demyelination of nerve cells. This damage leads to the disruption of the ability of affected parts of the nervous system to

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

guazzoales@dei.unipd.it (A. Guazzo); trescatois@dei.unipd.it (I. Trescato); enrico.longato@dei.unipd.it (E. Longato); erica.tavazzi@unipd.it (E. Tavazzi); martina.vettoretti@dei.unipd.it (M. Vettoretti); barbara.dicamillo@unipd.it (B. Di Camillo)
https://www.dei.unipd.it/~dicamill/ (B. Di Camillo)
0000-0001-5155-2567 (A. Guazzo); 0000-0003-0625-993X (I. Trescato); 0000-0001-5940-645X (E. Longato); 0000-0001-6188-6413 (E. Tavazzi); 0000-0002-5020-1818 (M. Vettoretti); 0000-0001-8415-4688 (B. Di Camillo)

transmit signals, resulting in a progressive neurologic impairment [1]. Typically, clinical practice is aimed at keeping the progression of the disability caused by the disease under control. As an obstacle, however, MS onset and progression are extremely heterogeneous across patients [2], which, as a result, exhibit significantly different management and treatment needs. This poses challenges for caregivers and clinicians who would benefit from the possibility of preemptively identifying the need for specific interventions or tailored therapeutic decisions that would have significant implications for the patient's quality of life. Therefore, developing automatic tools that could aid the clinicians' decision-making process to facilitate therapeutic choices would be of uttermost importance.

Specifically, this paper addresses the challenge proposed in the context of iDPP@CLEF 2023, in which participants were asked to rank subjects based on their risk of experiencing a worsening in the MS progression course. The challenge consisted of two Tasks. In Task 1, the prediction problem was cast as a survival analysis task where developed algorithms should reflect how early a patient experiences a worsening event. In Task 2, participants were asked to explicitly assign the cumulative probability of worsening at different prediction horizons, namely: 2, 4, 6, 8, and 10 years. The worsening event was defined according to clinical standards by considering Expanded Disability Status Scale (EDSS) [3] values over time. In particular, for each Task, two different clinically-relevant definitions of worsening were considered for as many sub-tasks (a and b), as detailed in Section 3. For each sub-task, participants were given a dataset containing 2.5 years of visits, with the occurrence (yes/no) of the worsening event and the time of the event. The data came from two Italian clinical institutions, one in Pavia and the other in Turin, and consisted of fully anonymised data from real-world MS patients.

To address the proposed problems, three different survival analysis approaches were used: the Cox proportional hazard model, the survival support vector machine (SSVM), and the random survival forest (RSF), refining the pipeline employed to an ALS patients cohort to address the previous iDPP@CLEF Challenge [4, 5]. Performance was evaluated via the metrics identified by the challenge organisers, i.e., C-index for Task 1, area under the receiver operating characteristic curve (AUROC) and O/E ratio for Task 2 [6, 7].

The paper is organized as follows: Section 2 introduces related works and the main methodological approaches implemented until now to address MS progression prediction. Section 3 describes our approach in terms of data preprocessing and adopting machine learning techniques. Section 4 discusses the findings and, finally, Section 5 draws conclusions and presents the outlook for future work.

2. Related Work

Different approaches have been proposed in the literature to describe and predict the prognosis of MS patients. Most of the literature's models describe MS progression by considering as outcomes the evolution over time of EDSS values [8, 9], the occurrence of relapses [10, 11], the transition to different disease stages [12, 13], and/or a survival endpoint for death [14].

Depending on the specifics of the research question and the corresponding selected outcome, in the literature, prediction tasks are modelled via different machine learning frameworks, i.e., classification, regression, or survival analysis.

The most popular and successful regression techniques for forecasting MS prognosis include linear regression [15], random forests (RF) [8], convolutional neural networks [16], and recurrent neural networks [17]. Classification tasks, such as the prediction, at a given time, of the occurrence of a clinically relevant event, are mainly based on support vector machines (SVM) [18, 19], RF [13] and logistic regression [10, 11]; Cox proportional hazard model [14], and RSF are the most frequent approaches in the survival analysis setting.

The vast majority of literature models use as predictors a set of variables collected at a visit considered as baseline (e.g., the visit at the time of MS diagnosis, or the first visit of a clinical trial), however, studies that considered as predictors variables collected over a 1-2 year time window around a baseline visit showed promising results [11]. More recently, also variables derived from MRI images (i.e., the number and location of T1 or T2 lesions) have started to be considered as possible predictors of MS progression, showing good predictive power [20, 21]. Finally, it is worth mentioning that many literature models have been developed on top of specific clinical trial datasets instead of datasets collected during everyday clinical practice. This may be a potential limitation to the practical use of predictive models by clinicians as models developed specifically after a clinical trial may struggle to perform well in a different and broader context.

3. Methodology

This paragraph describes the experimental workflow using the nomenclature used in the iDPP challenge guidelines. The two tasks will be referred to as:

- **Task 1**: predicting risk of disease worsening in MS;
- **Task 2**: predicting cumulative probability of worsening in MS;

and the two subtasks as:

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

A common preprocessing and model-training pipeline was devised and applied to all subtask-specific datasets provided by the organisers.

The following sections describe: i) the data processing steps needed to obtain the final set of static and dynamic input variables (sections 3.1.1 and 3.1.2); the data filtering step aimed at excluding scarcely populated variables and subjects (section 3.1.3); the normalisation (section 3.1.5) and imputation (section 3.1.6) steps; the model-independent workflow for training, feature selection, and hyperparameter optimisation via bootstrap resampling (section 3.1.4).

3.1. Preprocessing

The data were made available in several .csv files, ad described in [6, 7], reporting static variables, and dynamic assessments, such as information about the relapses, the EDSS measurements, the evoked potentials, the MRI, and the MS course of the subjects.

3.1.1. Static Variables' Preprocessing

In the raw data, *sex* was reported as a string with two levels: "male" and "female". It was then mapped to a Boolean variable such that 0 = "male" and 1 = "female". Similarly, the variable *centre*, which records the clinical centre from which the data originated, was converted to a Boolean variable such that 0 = "Pavia" and 1 = "Turin".

For the residence area classification, there were three possible levels: "rural area," "towns," and "cities". The two dummy variables included in the models were *residence_rural_area*, and *residence_towns*. The variable *ethnicity* only assumed the value "Caucasian", therefore it was discarded from the predictors. While it is not used to build the models, it is essential to account for the fact that the built models are only trained on Caucasian subjects.

The column *other_symptoms* was discarded because of the low percentage of valorised rows (2%).

3.1.2. Dynamic Variables' Preprocessing

One of the dynamic variables available for iDPP Challenge was the EDSS, as measured and reported by the clinician, together with the corresponding date of annotation. To describe the score's evolution during the observation window provided, four variables were derived for each subject: the minimum EDSS value (*min_EDSS*), the maximum EDSS value (*max_EDSS*), the first EDSS value (*first_EDSS*), and the last EDSS value (*last_EDSS*).

The second dynamic dataset carried information related to the evoked potentials (EP): the type of EP tested, the location examined, the date, and a Boolean variable to account for alterations. In the proposed models, alterations in EP were accounted as Boolean variables, divided by location (*altered_auditory_EP*, *altered_somatosensory_EP,* and *altered_visual_EP*). For each subject, the variable has value = 1 if an alteration in that EP was ever recorded, and 0 otherwise.

MRI assessments were also made available together with the date of the visit and reported details on the examined area, T1 lesions presence, T1 gadolinium-enhanced lesions presence and number, new T2 lesions presence and number, and total T2 lesions presence and number. To embed this information in the static dataset, we mapped the original columns as follows:

- *brain_stem_T1_lesions_gadolinium*: Boolean, has value 1 if the subject had at least one T1 gadolinium-enhanced lesion in the Brain Stem area, 0 otherwise.
- *brain_stem_T2_lesions*: Boolean, has value 1 if the subject had at least one T2 lesion in the Brain Stem area, 0 otherwise.
- *max_brain_stem_T1_lesions_gadolinium*: is the maximum value registered in the column *number_of_lesions_T1_gadolinium* for *mri_area_label* equal to "Brain Stem".
- *max_brain_stem_T2_lesions*: is the maximum value registered in the column *number_of_total_lesions_T2* for *mri_area_label* equal to "Brain Stem".

- *cervical_spinal_cord_T1_lesions_gadolinium*: Boolean, has value 1 if the subject had at least one T1 gadolinium-enhanced lesion in the Cervical Spinal Cord area, 0 otherwise.
- *cervical_spinal_cord_T2_lesions*: Boolean, has value 1 if the subject had at least one T2 lesion in the Cervical Spinal Cord area, 0 otherwise.
- *max_cervical_spinal_cord_T1_lesions_gadolinium*: is the maximum value registered in the column *number_of_lesions_T1_gadolinium* for *mri_area_label* equal to "Cervical Spinal Cord'.
- *max_cervical_spinal_cord_T2_lesions*: is the maximum value registered in the column emphnumber_of_total_lesions_T2 for *mri_area_label* equal to "Cervical Spinal Cord".
- *spinal_cord_T1_lesions_gadolinium*: Boolean, has value 1 if the subject had at least one T1 gadolinium-enhanced lesion in the Spinal Cord area, 0 otherwise.
- *spinal_cord_T2_lesions*: Boolean, has value 1 if the subject had at least one T2 lesion in the Spinal Cord area, 0 otherwise.
- *max_spinal_cord_T1_lesions_gadolinium*: is the maximum value registered in the column *number_of_lesions_T1_gadolinium* for *mri_area_label* equal to "Spinal Cord".
- *max_spinal_cord_T2_lesions*: is the maximum value registered in the column emphnumber_of_total_lesions_T2 for *mri_area_label* equal to "Spinal Cord".

As for the MS type, the information was not reported for 230 subjects over 440. This information was thus discarded for the scarcity of records.

Another piece of dynamic information was related to the relapses experienced by each subject. In the dedicated sheet, for each subject was available the date on which a relapse started: in the proposed models, this information was accounted for by deriving, for each subject, the number of relapses that occurred during the proposed observation period (*n_relapses*).

3.1.3. Quality Control

Once all variables were converted and embedded in the static dataset, a data quality analysis was performed to verify that each variable had more than 70% valid values and that no subjects had more than 20% of missing values.

3.1.4. Bootstrap

A total of 100 bootstrap sets, each having the same size as the entire dataset, were created from the training set made available by the challenge organisers. These bootstrap sets were employed as internal training sets, containing approximately 63.2% of the subjects, with the possibility of subject repetition. The remaining 36.8% of subjects, called out-of-bag subjects, were used as the corresponding validation set.

To ensure an effective model training process, a stratification analysis was performed on each bootstrap set to evaluate the extent to which the considered variables were appropriately distributed between the internal training set and the respective validation set. In order to assess the adequacy of the stratification, a comparison of variable distributions was conducted for each training/test pair. Statistical tests were employed based on the type of variable: the Kruskal-Wallis test [22] was utilized for continuous variables, while the Chi-squared test [23] was employed for categorical and ordinal variables. The stratification of a variable was deemed satisfactory based on the outcome of the corresponding test. The percentage of well-stratified variables for each split was calculated using the following equation:

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

Finally, to ensure a thorough and balanced analysis of performance, while accounting for computational complexity, the k bootstrap sets with the highest percentage of well-stratified variables were selected. The number of final bootstrap sets k was determined as the minimum number of bootstrap sets such that each subject is included in the validation set at least five times.

3.1.5. Normalisation

Normalisation is helpful to avoid introducing bias related to the different dynamic ranges of each variable, and to promote consistency between the scale of the coefficients that might be estimated during model training. Here, min-max scaling was used. In practice, let x be the variable to be normalised, this method constrains x into the range [0, 1] according to the following equation.

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)} \tag{2}$$

The normalisation parameters were derived separately on the k selected internal training sets and applied to the respective k validation sets.

3.1.6. Imputation

Imputation of the remaining missing values in the preprocessed input variables was performed using the R package mice [24] with one imputation, 20 iterations, and classification and regression trees as imputation method (*m=1, maxit=20, method="cart"*). The imputation parameters were estimated separately on the k selected internal training sets and applied to the respective k validation sets. To check the robustness of the imputation process, we compared the distributions of each variable before and after imputation.

3.2. Model Training

Three survival analysis methods were considered, namely: Cox, SSVM, and RSF. They were chosen to represent a broad spectrum of methodological approaches including parametric (SSVM), semi-parametric (Cox), linear (Cox, SSVM), and nonlinear (RSF) models. The same training workflow, based on feature selection and hyperparameter optimisation was considered for all models.

The Cox model and the RSF give as output the survival function. The survival function is characterised by survival probabilities that can easily be inverted to obtain a risk score to be used to rank MS patients according to their risk of worsening ($risk(t) = 1 - S(t)$). Such probabilities can be used to address Task 1 by considering the risk at the maximum observed time ($t = 15$

years), and Task 2 by considering the risks at the requested times ($t \in (2, 4, 6, 8, 10)$ years). Instead, the SSVM can be used either as a ranker or a time regressor depending on how the risk ratio hyperparameter is set during model training. Here, the SSVM was initially trained as a time regressor, then, its predicted times were converted into worsening probabilities in the range [0-1] using Platt scaling [25] for each time window of interest (Task 1: 15 years, Task 2: 2, 4, 6, 8, 10 years).

The same training workflow was used for all considered models (and both sub-tasks). For each bootstrap set, an optimal model was obtained by performing hyperparameter optimisation and feature selection. Hyperparameters were optimized using a 5-fold cross-validation (CV) and a random search approach [26] over a given hyperparameter space specific for the considered methodology to be trained. A single hyperparameter was optimised for the Cox model: the strength of the L2 regularisation (α, 500 values randomly sampled from a log-uniform distribution with support $[10^{-5} - 10^{5}]$). For the SSVM a linear kernel was used and the rank ratio was set equal to 0 so as to obtain a time regression model. The hyperparameter space consisted of a single hyperparameter: the penalisation weight of the squared hinge loss objective function (β, randomly sampled from a log-uniform distribution with support $[10^{-3} - 10^{4}]$). Finally, the RSF's hyperparameter space consisted of two hyperparameters, namely the number of trees in each random forest (n_estimators, uniformly sampled in the interval $[40 - 500]$), and the maximum depth of each tree (uniformly sampled in the interval $[2 - 10]$). By default, the square root of the total number of features is evaluated at each node for splitting.

The optimal hyperparameters were chosen as those that maximised the average Harrell's concordance index (C-index) over the five validation folds. The optimal set of features was obtained contextually by using the forward recursive feature selection (FRFS) approach [27]. At each FRFS step, the best feature to be added was chosen as the one that led to the maximum average C-index over the 5 validation folds. The early stopping criterion for FRFS was a relative variation of the average C-index < 0.01%. For feature selection purposes, the optimal model (i.e., the one with the optimal hyperparameters and features set) at each FRFS iteration was applied also to the bootstrap validation set, and the corresponding FRFS-iteration-specific validation C-index was computed. These validation C-indices were used to determine the optimal number of features (N) to be included in the final model. Specifically, the median validation C-index obtained in the considered bootstrap sets, with its variability represented by the 25^{th} and 75^{th} percentiles, was evaluated as a function of the number of features considered within the FRFS, and the optimal number of features N was selected as the one that led to the maximum median C-index on the validation sets of the boots. In order to select the N features to be included in the final model, the average number of FRFS steps at which each variable was added to the set of predictors in all considered bootstrap sets was considered. If a variable was never selected in a given set, its associated number of steps was set to the maximum number of variables plus one. The final model was, then, trained on the whole training set using only the N variables with the lower average number of steps. The final model's hyperparameters (chosen from the same hyperparameters spaces considered in the previous step) were optimised again via 5-fold cross-validation and a random search approach. The optimal hyperparameters were chosen as those that led to the maximum cross-validation C-index on the 5 validations folds. Figure 1 summarises the model training workflow described in this section.

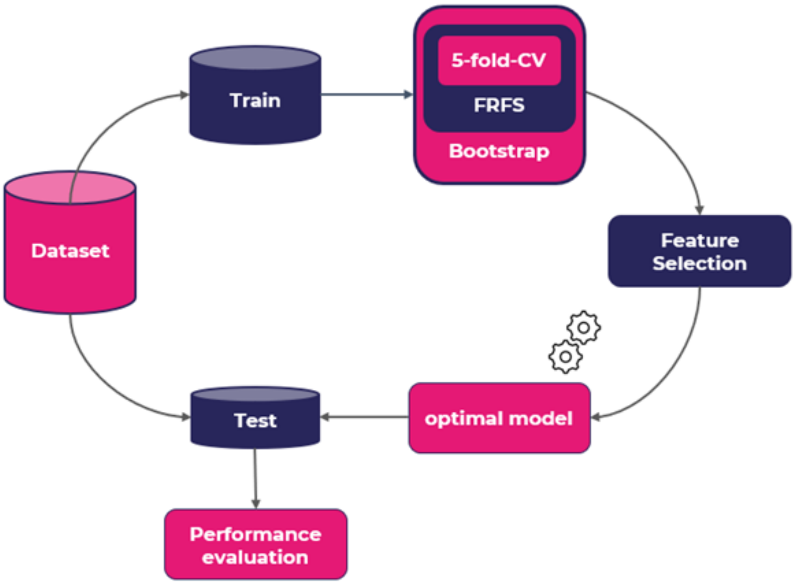

Figure 1: Model training workflow depiction. The training set is used to obtain K bootstrap to be used with FRFS and 5-fold CV to perform feature selection and hyperparameter optimisation. The optimal features are used to train a final model evaluated on the test set.

3.3. Performance evaluation

To assess the models' performance, the evaluation measures computed by iDPP challenge organisers were considered. For Task 1, C-index is used, while Task 2 runs are evaluated through the AUROC and O/E ratio.

4. Results

The three described methods were employed to build two different models addressing the two subtasks of Task 1 and to derive the cumulative probabilities asked in Task 2, subtasks a and b, as described in section 3.2. This methodology resulted in 12 submitted runs (3 methods, 2 tasks, 2 subtasks).

4.1. Preprocessing

After the preprocessing of the data, as outlined in Section 3.1, a training set comprising 369 subjects out of 440 was obtained for subtask a. Additionally, a test set consisting of 88 subjects out of 110 was created for the same tasks and subtask. For subtask b, the training set contained 423 subjects out of the available 510, while the test set comprised 100 subjects out of the 128 available. All of the datasets employed include 32 variables.

Table 1
Task 1 metrics: C-index is reported with the 95% confidence intervals.

	Cox	RSF	SSVM
Subtask a	0.486 (0.486-0.799)	0.588 (0.438-0.738)	0.623 (0.506-0.739)
Subtask b	0.598 (0.460-0.737)	0.481 (0.382-0.580)	0.578 (0.457-0.698)

4.2. Task 1 Results

Table 1 shows the performance metrics of our submitted models for subtasks a and b of Task 1. In the first row, C-index with confidence intervals are reported for subtask a while in the second row those for subtask b. The different columns refer to the three employed methods (Cox, RSF, and SSVM respectively).

Models discrimination resulted rather poorly in all the different scenarios: the higher C-index is that of the SSVM model for subtask a, equal to 0.623, and that of the Cox model for subtask b, equal to 0.598.

Regarding the models trained to address subtask a, the Cox model retained 13 variables and the optimal α resulted to be 0.00579. The SSVM final model retained 4 variables, with β = 0.00855. The RSF kept 6 variables, with XX trees and a maximum depth of XX.

4.3. Task 2 Results

Task 2 runs were derived from Task 1 models as explained in Section 3.2. Sections 4.3.1 and 4.3.2 present and analyze the performance metrics for subtasks a and b, respectively. The tables 2 and 3 show the performance metrics for each time horizon considered (2, 4, 6, 8, and 10 years) for subtasks a and b, respectively. For each time horizon, the upper line of the table reports the AUROC values, while the lower line displays the O/E ratios. These metrics are reported for each considered method, namely Cox, RSF, and SSVM, which are presented in separate columns, together with their 95% confidence intervals.

4.3.1. Subtask a

For the 2-year time horizon, the Cox method shows the highest AUROC of 0.708, indicating better predictive performance compared to RSF (0.604) and SSVM (0.624). However, when considering the O/E ratio, all three methods have relatively similar performance, with Cox having a slightly higher value of 0.389 compared to RSF (0.385) and SSVM (0.358). Moving to the 4-year time horizon, the Cox method maintains its superiority in terms of AUROC (0.762), outperforming RSF (0.644) and SSVM (0.631). Similarly, the Cox and RSF methods obtain similar O/E ratios of 0.577 and 0.604, respectively, whereas the SSVM O/E ratio is equal to 0.405. Also for the 6-year time horizon, the Cox method reaches the highest AUROC (0.728) concerning RSF (0.638) and SSVM (0.643). Regarding the O/E ratio, Cox (0.539) and RSF (0.520) perform comparably, while SSVM (0.357) shows slightly worse performance. In the 8-year time horizon scenario, the SSVM method has C-index equal to 0.697, outperforming Cox (0.650) and RSF (0.556). At the same time, the O/E ratio favours Cox (0.547) and RSF (0.570), with SSVM (0.328)

Table 2
Task 2a metrics: AUROC and O/E Ratio are reported with the 95% confidence intervals.

		Cox	**RSF**	**SSVM**
2 years	AUROC	0.708 (0.491, 0.926)	0.604 (0.386, 0.822)	0.624 (0.461, 0.787)
	O/E ratio	0.389 (-0.043, 0.821)	0.385 (-0.045, 0.815)	0.358 (-0.057, 0.772)
4 years	AUROC	0.762 (0.576, 0.948)	0.644 (0.459, 0.830)	0.631 (0.466, 0.796)
	O/E ratio	0.577 (0.106, 1.047)	0.604 (0.123, 1.086)	0.405 (0.010, 0.799)
6 years	AUROC	0.728 (0.541, 0.916)	0.638 (0.445, 0.830)	0.643 (0.470, 0.816)
	O/E ratio	0.539 (0.124, 0.954)	0.520 (0.112, 0.929)	0.357 (0.019, 0.695)
8 years	AUROC	0.650 (0.454, 0.847)	0.556 (0.354, 0.759)	0.697 (0.530, 0.865)
	O/E ratio	0.547 (0.159, 0.934)	0.570 (0.174, 0.965)	0.328 (0.028, 0.627)
10 years	AUROC	0.608 (0.388, 0.828)	0.568 (0.342, 0.794)	0.659 (0.446, 0.872)
	O/E ratio	0.522 (0.168, 0.877)	0.491 (0.148, 0.835)	0.275 (0.018, 0.532)

Table 3
Task 2b metrics: AUROC and O/E Ratio are reported with the 95% confidence intervals.

		Cox	**RSF**	**SSVM**
2 years	AUROC	0.642 (0.397, 0.887)	0.514 (0.281, 0.747)	0.547 (0.345, 0.750)
	O/E ratio	1.098 (0.449, 1.748)	0.966 (0.357, 1.576)	0.814 (0.255, 1.373)
4 years	AUROC	0.567 (0.382, 0.752)	0.529 (0.366, 0.692)	0.629 (0.468, 0.791)
	O/E ratio	0.807 (0.380, 1.234)	0.711 (0.310, 1.111)	0.520 (0.177, 0.863)
6 years	AUROC	0.601 (0.428, 0.774)	0.511 (0.337, 0.685)	0.560 (0.386, 0.733)
	O/E ratio	0.782 (0.404, 1.160)	0.695 (0.339, 1.052)	0.444 (0.159, 0.729)
8 years	AUROC	0.594 (0.419, 0.770)	0.668 (0.496, 0.839)	0.474 (0.296, 0.651)
	O/E ratio	0.735 (0.392, 1.079)	0.672 (0.344, 1.000)	0.403 (0.149, 0.657)
10 years	AUROC	0.622 (0.441, 0.803)	0.646 (0.470, 0.821)	0.541 (0.356, 0.725)
	O/E ratio	0.785 (0.451, 1.120)	0.705 (0.388, 1.021)	0.397 (0.159, 0.635)

having poorer performance. Finally, for the 10-year time horizon, SSVM (0.659) and Cox (0.608) demonstrate higher AUROC values compared to RSF (0.568). Similar to previous time horizons, Cox (0.522) and SSVM (0.275) achieve higher O/E ratios than RSF (0.491).

In summary, for different time horizons, the Cox method consistently exhibits competitive performance in terms of AUROC, outperforming RSF and SSVM in most cases. However, when considering the O/E ratio, Cox and RSF models often obtain comparable performance, while the SSVM model in most cases has a lower score.

4.3.2. Subtask b

For the 2-year time horizon, the Cox method achieves the highest AUROC (0.642), followed by SSVM (0.547) and RSF (0.514). In terms of O/E ratio, Cox (1.098) has the highest value, followed by RSF (0.966) and SSVM (0.814). For the 4-year time horizon, the SSVM model reaches AUROC equal to 0.629, outperforming the Cox model (0.807) and RSF (0.529). As for the O/E ratio, Cox (0.807) resulted in the best score, followed by RSF (0.711) and SSVM (0.520). Examining the 6-year time horizon, the AUROC is higher for the Cox (0.601), followed by SSVM (0.560) and RSF

(0.511). Analyzing the O/E ratio, Cox also achieves the highest value of 0.782, outperforming RSF (0.695) and SSVM (0.444). Considering the 8-year time horizon, the RSF method achieves the highest AUROC (0.668), followed by Cox (0.594) and SSVM (0.474), but inspecting the O/E ratios, Cox (0.735) performs the best, followed by RSF (0.672) and SSVM (0.403). Lastly, at the 10-year time horizon, the results are similar, with the RSF method having again the highest AUROC (0.646), followed by Cox (0.622) and SSVM (0.541), and Cox achieving the highest O/E ratio (0.785), followed by RSF (0.705) and SSVM (0.397).

For this subtask, the best-performing method varies depending on the time horizon when looking at the AUROC values, but the Cox method consistently outperforms the RSF and the SSVM in O/E ratios.

5. Conclusions and Future Work

This study aimed to address the first two tasks of the iDPP Challenge, focusing on predicting two different events of interest (subtasks a and b) associated with disease progression in a cohort of MS patients within a survival framework. Three methods were employed, namely Cox, Random Survival Forests, and Survival Support Vector Machines.

In Task 1, the results revealed poor discrimination abilities across all models. However, the Cox model achieved the best performance, retaining 13 variables after feature selection. In Task 2, the overall performance was modest, with the Cox model reaching again the better scores, particularly in terms of the E/O ratio, across various time horizons and subtasks. While not explicitly mentioned by the challenge, consideration was given to the clinical context and interpretability. In this regard, the Cox model is effective, by providing coefficient estimates for identified prognostic factors. Nonetheless, striking a balance between interpretability and predictive performance remains crucial when selecting the optimal model for a specific task.

The obtained results are consistent with the findings in the literature. Chalckou et al. [10] reported a C-index of 0.6, which aligns with our findings. Similarly, Ye et al. achieved a C-index of 0.67 on the training set and 0.59 on the test set [14]. However, some authors have achieved better results. For instance, Pisani et al. utilized random survival forests to predict the time to secondary progressive MS, considering therapy, and achieved an overall accuracy of 88%, along with a specificity of 87% and sensitivity of 92% [12].

The limited performance could potentially be attributed to the process of feature selection. In future developments, a revision of the selection made from the pool of available variables will be undertaken. Recursive Feature Elimination will also be tested as an alternative to the forward approach, aiming to mitigate any potential issues associated with early stopping.

Our findings indicate that employing non-linear or complex methods does not lead to a discernible advantage. Nonetheless, exploring alternative machine learning techniques may help in augmenting model discrimination in survival analysis.

References

[1] A. Compston, A. Coles, Multiple sclerosis, The Lancet 372 (2008) 1502–1517. URL: https://www.sciencedirect.com/science/article/pii/S0140673608616207. doi:https://doi.

org/10.1016/S0140-6736(08)61620-7.

[2] H. L. Weiner, The challenge of multiple sclerosis: How do we cure a chronic heterogeneous disease?, Annals of Neurology 65 (2009) 239–248. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.21640. doi:https://doi.org/10.1002/ana.21640. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.21640.

[3] J. F. Kurtzke, Rating neurologic impairment in multiple sclerosis, Neurology 33 (1983) 1444–1444. URL: https://n.neurology.org/content/33/11/1444. doi:10.1212/WNL.33.11.1444. arXiv:https://n.neurology.org/content/33/11/1444.full.pdf.

[4] I. Trescato, A. Guazzo, E. Longato, E. Hazizaj, Baseline machine learning approaches to predict amyotrophic lateral sclerosis disease progression notebook for the idpp lab on intelligent disease progression prediction at clef 2022, in: Proceedings of the CLEF 2022 Conference and Labs of the Evaluation Forum, Bologna, Italy, 2022, pp. 5–8.

[5] A. Guazzo, I. Trescato, E. Longato, E. Hazizaj, D. Dosso, G. Faggioli, G. M. Di Nunzio, G. Silvello, M. Vettoretti, E. Tavazzi, C. Roversi, P. Fariselli, S. C. Madeira, M. de Carvalho, M. Gromicho, A. Chiò, U. Manera, A. Dagliati, G. Birolo, H. Aidos, B. Di Camillo, N. Ferro, Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[6] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, A. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2023.

[7] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), CLEF 2023 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073., 2023.

[8] C. Cordani, M. H. de la Cruz, A. Meani, P. Valsasina, F. Esposito, E. Pagani, M. Filippi, M. A. Rocca, Mri correlates of clinical disability and hand-motor performance in multiple sclerosis phenotypes, Multiple Sclerosis Journal 27 (2021) 1205–1221. URL: https://doi.org/10.1177/1352458520958356. doi:10.1177/1352458520958356. arXiv:https://doi.org/10.1177/1352458520958356, pMID: 32924846.

[9] R. A. Marrie, Q. Tan, O. Ekuma, J. J. Marriott, Development and internal validation of a disability algorithm for multiple sclerosis in administrative data, Frontiers in Neurology 12 (2021). URL: https://www.frontiersin.org/articles/10.3389/fneur.2021.754144. doi:10.3389/fneur.2021.754144.

[10] K. Chalkou, E. Steyerberg, M. Egger, A. Manca, F. Pellegrini, G. Salanti,

A two-stage prediction model for heterogeneous effects of treatments, Statistics in Medicine 40 (2021) 4362–4375. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9034. doi:https://doi.org/10.1002/sim.9034. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9034.

[11] Y. Ahuja, N. Kim, L. Liang, T. Cai, K. Dahal, T. Seyok, C. Lin, S. Finan, K. Liao, G. Savovoa, T. Chitnis, T. Cai, Z. Xia, Leveraging electronic health records data to predict multiple sclerosis disease activity, Annals of Clinical and Translational Neurology 8 (2021) 800–810. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acn3.51324. doi:https://doi.org/10.1002/acn3.51324. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acn3.51324.

[12] A. I. Pisani, A. Scalfari, F. Crescenzo, C. Romualdi, M. Calabrese, A novel prognostic score to assess the risk of progression in relapsing-remitting multiple sclerosis patients, European Journal of Neurology 28 (2021) 2503–2512. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ene.14859. doi:https://doi.org/10.1111/ene.14859. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ene.14859.

[13] M. A. Rocca, P. Valsasina, A. Meani, E. Pagani, C. Cordani, C. Cervellin, M. Filippi, Network damage predicts clinical worsening in multiple sclerosis, Neurology - Neuroimmunology Neuroinflammation 8 (2021). URL: https://nn.neurology.org/content/8/4/e1006. doi:10.1212/NXI.0000000000001006. arXiv:https://nn.neurology.org/content/8/4/e1006.full.pdf.

[14] F. Ye, X. Wu, T. Wang, J. Liang, J. Li, Y. Dai, K. Lan, W. Sheng, Identification of immune-associated gene signature and immune cell infiltration related to overall survival in progressive multiple sclerosis, Multiple Sclerosis and Related Disorders 55 (2021) 103188. URL: https://www.sciencedirect.com/science/article/pii/S2211034821004557. doi:https://doi.org/10.1016/j.msard.2021.103188.

[15] Y. Zhao, B. C. Healy, D. Rotstein, C. R. G. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, T. Chitnis, Exploration of machine learning techniques in predicting multiple sclerosis disease course, PLOS ONE 12 (2017) 1–13. URL: https://doi.org/10.1371/journal.pone.0174866. doi:10.1371/journal.pone.0174866.

[16] P. Roca, A. Attye, L. Colas, A. Tucholka, P. Rubini, S. Cackowski, J. Ding, J.-F. Budzik, F. Renard, S. Doyle, E. Barbier, I. Bousaid, R. Casey, S. Vukusic, N. Lassau, S. Verclytte, F. Cotton, B. Brochet, R. Casey, F. Cotton, J. De Sèze, P. Douek, F. Guillemin, D. Laplaud, C. Lebrun-Frenay, L. Mansuy, T. Moreau, J. Olaiz, J. Pelletier, C. Rigaud-Bully, B. Stankoff, S. Vukusic, R. Marignier, M. Debouverie, G. Edan, J. Ciron, A. Ruet, N. Collongues, C. Lubetzki, P. Vermersch, P. Labauge, G. Defer, M. Cohen, A. Fromont, S. Wiertlewsky, E. Berger, P. Clavelou, B. Audoin, C. Giannesini, O. Gout, E. Thouvenot, O. Heinzlef, A. Al-Khedr, B. Bourre, O. Casez, P. Cabre, A. Montcuquet, A. Créange, J.-P. Camdessanché, J. Faure, A. Maurousset, I. Patry, K. Hankiewicz, C. Pottier, N. Maubeuge, C. Labeyrie, C. Nifle, R. Ameli, R. Anxionnat, A. Attye, E. Bannier, C. Barillot, D. Ben Salem, M.-P. Boncoeur-Martel, F. Bonneville, C. Boutet, J.-C. Brisset, F. Cervenanski, B. Claise, O. Commowick, J.-M. Constans, P. Dardel, H. Desal, V. Dousset, F. Durand-Dubief, J.-C. Ferre, E. Gerardin, T. Glattard, S. Grand, T. Grenier, R. Guillevin, C. Guttmann, A. Krainik, S. Kremer, S. Lion, N. Menjot de Champfleur, L. Mondot, O. Outteryck, N. Pyatigorskaya, J.-P. Pruvo, S. Rabaste, J.-P. Ranjeva, J.-A. Roch, J. Sadik, D. Sappey-Marinier, J. Savatovsky, J.-Y. Tanguy,

A. Tourbah, T. Tourdias, Artificial intelligence to predict clinical disability in patients with multiple sclerosis using flair mri, Diagnostic and Interventional Imaging 101 (2020) 795–802. URL: https://www.sciencedirect.com/science/article/pii/S2211568420301558. doi:https://doi.org/10.1016/j.diii.2020.05.009.

[17] Y. Zhao, M. Berretta, T. Wang, T. Chitnis, Gru-df: A temporal model with dynamic imputation for missing target values in longitudinal patient data, in: 2020 IEEE International Conference on Healthcare Informatics (ICHI), 2020, pp. 1–7. doi:10.1109/ICHI48887.2020.9374359.

[18] Y. Peng, Y. Zheng, Z. Tan, J. Liu, Y. Xiang, H. Liu, L. Dai, Y. Xie, J. Wang, C. Zeng, Y. Li, Prediction of unenhanced lesion evolution in multiple sclerosis using radiomics-based models: a machine learning approach, Multiple Sclerosis and Related Disorders 53 (2021) 102989. URL: https://www.sciencedirect.com/science/article/pii/S221103482100256X. doi:https://doi.org/10.1016/j.msard.2021.102989.

[19] A. Montolío, A. Martín-Gallego, J. Cegoñino, E. Orduna, E. Vilades, E. Garcia-Martin, A. P. del Palomar, Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography, Computers in Biology and Medicine 133 (2021) 104416. URL: https://www.sciencedirect.com/science/article/pii/S0010482521002109. doi:https://doi.org/10.1016/j.compbiomed.2021.104416.

[20] L. Razzolini, E. Portaccio, M. L. Stromillo, B. Goretti, C. Niccolai, L. Pastò, I. Righini, E. Prestipino, M. Battaglini, A. Giorgio, N. De Stefano, M. P. Amato, The dilemma of benign multiple sclerosis: Can we predict the risk of losing the "benign status"? a 12-year follow-up study, Multiple Sclerosis and Related Disorders 26 (2018) 71–73. URL: https://www.sciencedirect.com/science/article/pii/S2211034818302864. doi:https://doi.org/10.1016/j.msard.2018.08.011.

[21] A. Kuceyeski, E. Monohan, E. Morris, K. Fujimoto, W. Vargas, S. Gauthier, Baseline biomarkers of connectome disruption and atrophy predict future processing speed in early multiple sclerosis, NeuroImage: Clinical 19 (2018) 417–424. URL: https://www.sciencedirect.com/science/article/pii/S2213158218301499. doi:https://doi.org/10.1016/j.nicl.2018.05.003.

[22] P. E. McKight, J. Najab, Kruskal-wallis test, The corsini encyclopedia of psychology (2010) 1–1.

[23] R. J. Tallarida, R. B. Murray, R. J. Tallarida, R. B. Murray, Chi-square test, Manual of pharmacologic calculations: with computer programs (1987) 140–142.

[24] S. Van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, Journal of statistical software 45 (2011) 1–67.

[25] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 625–632.

[26] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

[27] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.