

# Time-to-Event Interpretable Machine Learning for Multiple Sclerosis Worsening Prediction: Results from iDPP@CLEF 2023

Angela Lombardi<sup>1,\*</sup>, Maria Luigia Natalia De Bonis<sup>1,†</sup>, Giuseppe Fasano<sup>1,†</sup>, Alessia Sportelli<sup>1,†</sup>, Tommaso Colafoglio<sup>1,2</sup>, Domenico Lofù<sup>1</sup>, Paolo Sorino<sup>1</sup>, Fedelucio Narducci<sup>1</sup>, Eugenio Di Sciascio<sup>1</sup> and Tommaso Di Noia<sup>1</sup>

<sup>1</sup>Department of Electrical and Information Engineering, Polytechnic of Bari, Bari, Italy

<sup>2</sup>Department of Computer, Automatic and Management Engineering (DIAG), Sapienza Università di Roma, Roma (Italy)

## Abstract

In this work, we present a framework for the interpretable analysis of machine learning algorithms to predict the Multiple Sclerosis worsening using the datasets provided by the iDPP@CLEF 2023 Challenge. The proposed framework is modular and allows to investigate the link between the provided static and dynamic features and the outcome to be predicted. Our findings show that better performance could be achieved by using Random Survival Forests together with temporal information about the clinical scores and a proposed feature related to the normalized frequency of patients' relapses.

## Keywords

Disease progression prediction, Time-to-event machine learning, Multiple Sclerosis

## 1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system (CNS) affecting millions worldwide. It causes a variety of neurological symptoms due to neurodegeneration, demyelination, and inflammation. MS has a complex and heterogeneous presentation, making diagnosis challenging. However, advancements in computational diagnostic tools and algorithms have significantly improved the accuracy and efficiency of MS diagnosis.

To assess and diagnose MS, many diagnostic tools are used. One of the most used is magnetic resonance imaging (MRI), which allows to see CNS abnormalities that are symptomatic of demyelination. Contrast agents based on gadolinium improve the detection of active lesions and aid in differentiating MS from other disorders. In addition, MRI can measure the temporal

---

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

\*Corresponding author.


†These authors contributed equally.

✉ [angela.lombardi@epoliba.it](mailto:angela.lombardi@epoliba.it) (A. Lombardi); [m.debonis4@studenti.poliba.it](mailto:m.debonis4@studenti.poliba.it) (M. L. N. De Bonis); [g.fasano8@studenti.poliba.it](mailto:g.fasano8@studenti.poliba.it) (G. Fasano); [a.sportelli3@studenti.poliba.it](mailto:a.sportelli3@studenti.poliba.it) (A. Sportelli); [tommaso.colafoglio@poliba.it](mailto:tommaso.colafoglio@poliba.it) (T. Colafoglio); [domenico.lofu@poliba.it](mailto:domenico.lofu@poliba.it) (D. Lofù); [paolo.sorino@poliba.it](mailto:paolo.sorino@poliba.it) (P. Sorino); [fedelucio.narducci@poliba.it](mailto:fedelucio.narducci@poliba.it) (F. Narducci); [tommaso.dinoia@poliba.it](mailto:tommaso.dinoia@poliba.it) (T. Di Noia)

🌐 <https://sisinflab.poliba.it/> (T. Di Noia)

🆔 0000-0000-0000-0001 (A. Lombardi)

© 2023 Copyright for this paper by its authors.  
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and spatial evolution of lesions, which is helpful for evaluating treatments and for monitoring diseases [1, 2]. Another crucial diagnostic tool for MS is the examination of cerebrospinal fluid (CSF). CSF is obtained via lumbar puncture and examined for the presence of oligoclonal bands and elevated levels of immunoglobulin G (IgG) and myelin basic protein. These CSF anomalies offer further evidence that inflammatory processes are present in the CNS. Electrophysiological techniques that measure the speed of nerve signal conduction in particular pathways include visual evoked potentials (VEP), brainstem auditory evoked potentials (BAEP), and somatosensory evoked potentials (SSEP) [3]. These tests may reveal demyelination and neuronal degeneration, supporting the diagnosis of MS.

In MS diagnosis and treatment, computational methods have become increasingly important. To improve the precision and effectiveness of diagnosis, prognosis, and treatment decision-making, clinical and imaging data are processed using machine learning (ML) and artificial intelligence (AI) approaches. These algorithms can precisely identify and measure MS lesions, aiding in the evaluation of disease burden and progression [4]. Moreover, ML and AI methods are becoming increasingly important for the prediction of disease course and treatment response [5]. By analyzing clinical and imaging data from a large number of patients, ML models can identify patterns and biomarkers that are associated with disease progression or treatment outcomes. This information can guide personalized treatment strategies and improve patient care.

Time-to-event machine learning models, also known as survival analysis models, are used to predict the time until an event of interest occurs, such as disease progression or death [6]. These models have been proven particularly valuable in the field of healthcare, including predicting disease progression in conditions like MS. In addition, they have shown greater performance than traditional parametric approaches for disease progression prediction tasks [7]. Parametric models assume specific distributions for the survival times, such as the Weibull or exponential distribution, which may not always accurately capture the complexities of real-world data. In contrast, machine learning models are capable of handling high-dimensional data, nonlinear relationships, and interactions between predictors. Censoring and time-varying predictors are common challenges in survival analysis. Machine learning survival models can handle censoring, i.e., the event of interest has not yet occurred at the end of the study. They can also incorporate time-varying predictors, allowing for dynamic predictions and accounting for changes in the predictors over time.

In this work, we exploit the MS dataset developed in the iDPP@CLEF 2023 challenge including demographic and clinical characteristics of about 1800 patients to explore the potential of different time-to-event ML approaches for MS disease progression prediction. The ultimate goal is to provide an interpretable framework that maximises the performance of the different prediction tasks of the challenge and at the same time highlights different aspects related to the interpretability of the results such as possible bias and the impact of clinical predictors on the achieved performance.

The paper is organized as follows: Section 2 introduces related works; Section 3 describes our approach; Section 4 explains our experimental setup; Section 5 discusses our main findings; finally, Section 6 draws some conclusions and outlooks for future work.

## 2. Related Work

Several studies have explored the application of machine learning techniques for predicting the worsening of multiple sclerosis (MS) and identifying patients at a higher risk of disease progression.

Zhao et al. [8] used different ML algorithms to predict an increase in  $EDSS \geq 1.5$  (worsening) or not (non-worsening) at up to 5 years after the baseline visit. They utilized a comprehensive set of clinical and imaging features, including demographic information, clinical scores, and magnetic resonance imaging (MRI) data. Their models achieved high accuracy in identifying patients who experienced worsening, demonstrating the potential of machine learning in risk stratification for MS progression.

Fiorini et al. [9] exploited different classifiers to analyze clinical data for the detection of MS courses and distinguish between progressive and benign patterns.

Montolio et al. [10] developed different ML models based on Retinal nerve fiber layer (RNFL) thickness and clinical data FOR MS diagnosis and MS disability course prediction founding new powerful biomarkers.

Although several models have been proposed that can discriminate between different levels of disease progression, time-to-event ML models have not yet been extensively explored to predict the course of disease in a continuous manner.

## 3. Methodology

In this work, we implemented a workflow for the following tasks of the challenge:

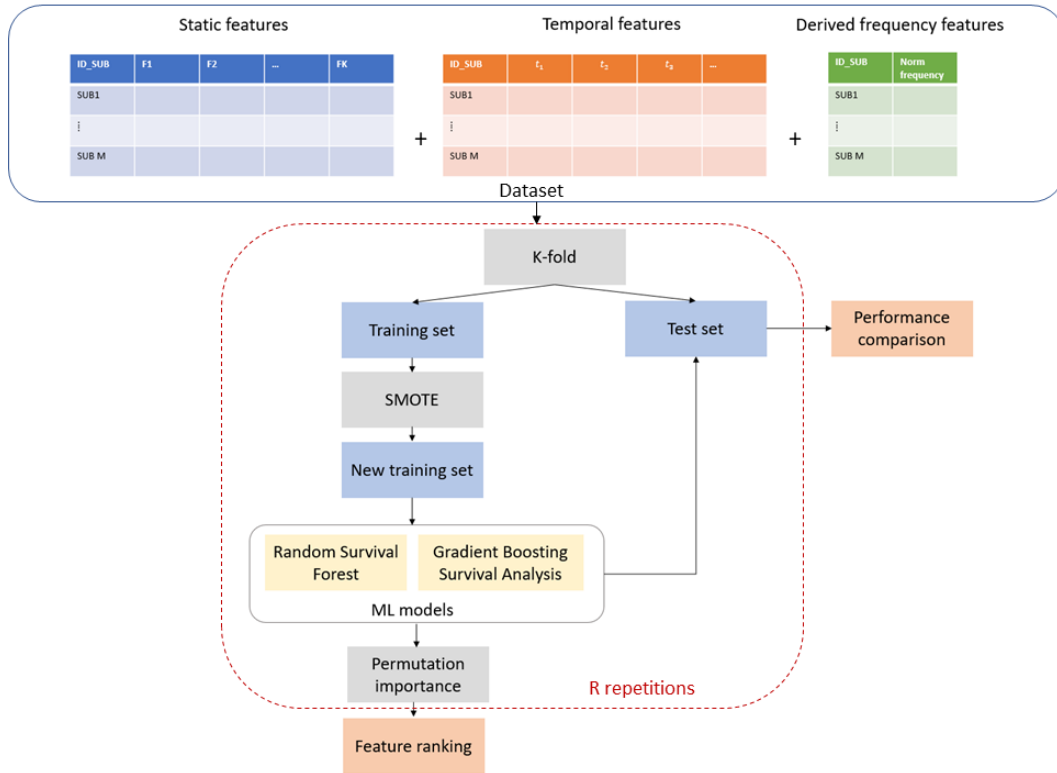
- Task 1a: predicting risk of disease worsening by using survival analysis, where worsening is defined based on EDSS Expanded Disability Status Scale (EDSS) with the threshold  $EDSS \geq 3$  at least twice within one-year interval;
- Task 1b: predicting risk of disease worsening, where worsening depends on the first recorded value accordingly to current clinical protocols;
- Task 2a: predicting the probability of worsening (MS): by explicitly assigning a cumulative probability of worsening at different time windows (e.g., between years 0 and 2, 0 and 4, 0 and 6, 0 and 8, 0 and 10) for subjects assigned to Task 1a;
- predicting a cumulative probability of worsening at different time windows (0-2; 0-4; 0-6; 0-8; 0-10) for subjects assigned to Task 1b.

The proposed workflow is reported in Figure 1. Each step is detailed in the following sections.

### 3.1. Dataset

The Challenge organizers provided the following data for the four tasks:

- static features about each patient with information on age, sex, and others related to the onset of the disease such as the presence of certain symptoms, age of onset and the medical centre;
- information about the relative start date of relapses;



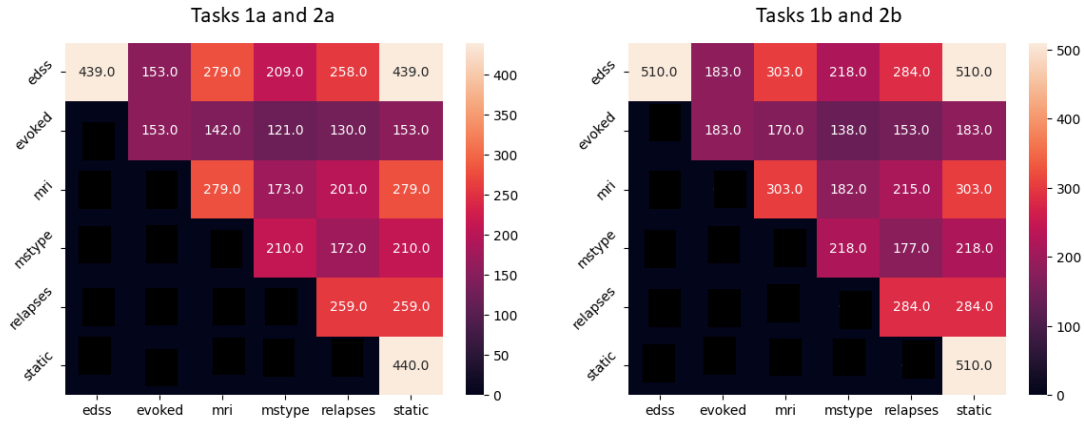
**Figure 1:** Proposed workflow for tasks 1a, 1b, 2a, 2b.

- tests on evoked potentials;
- information on the areas on which MRIs have been performed and the observed lesions;
- information about the MS course;
- the relative date when EDSS scores were measured, together with the EDSS scores evaluated by clinicians;
- the outcomes containing the patients' worsening occurrence, together with the time of occurrence.

More details on the datasets can be found in [11] and [12] Firstly, we compared the number of subjects with static information and outcomes with respect to all the other dynamic information, selecting for all tasks only the static features, relapses and MRI information (see Figure 2 for a visual comparison). In the second step, features on lesions extracted by using MRI images were also excluded due to the high number of missing data (NaN values).

In order to obtain the clinical history of the patient in terms of the time course of the EDSS scores, we used a landmarks-style approach [13], sampling the time instants of the variable "delta\_edss\_time0" on the training set and obtaining 16 temporal points for each patient. In the absence of data at a specific time instant, the  $EDSS = 0$  was entered.

Moreover, we transformed the information about the relapses for each patient  $i$ , into an additional relative frequency feature as:



**Figure 2:** Patient-related overlaps for each type of feature for all tasks.

$$variation\_relapses_i = \frac{n\_relapses}{(t_{end} - t_1)}, \quad (1)$$

where  $n\_relapses$  represents the number of relapses between the first recorded time ( $t_1$ ) and the last recorded time ( $t_{end}$ ).

Missing values for each variable were imputed using the median values. Moreover, the variable "centre" was excluded.

### 3.2. Validation scheme

We adopted a repeated k-fold cross-validation scheme, i.e., a resampling technique that combines k-fold cross-validation with repetition, as it presents several advantages such as: a robust estimate of model performance, better generalization, reduced bias and improved hyperparameter tuning [14, 15].

Moreover, within each validation round, we addressed the class imbalance between censored and not-censored events oversampling the minority class by means of the Synthetic Minority Oversampling TEchnique (SMOTE) [16].

### 3.3. Time-to-Event Machine Learning

We exploited time-to-event machine learning models to evaluate the risk of worsening for each patient as defined in the two tasks 1a and 2a. Each observation of the dataset is described by a set of features  $X = (x_1, \dots, x_n)$ , the time  $t$  when the event occurred, or the censoring time  $c > 0$ . By using an event indicator  $\delta \in \{0, 1\}$ , the observable time  $y$  of a right-censored sample is then defined as:

$$y = \min(t, c) = \begin{cases} t & \text{if } \delta = 1 \\ c & \text{if } \delta = 0 \end{cases} \quad (2)$$

Time-to-event data are modelled by using survival analysis. The key concepts and techniques in survival analysis include:

- Survival Function: it is denoted as  $S(t)$ , representing the probability that an individual survives beyond time  $t$ . It provides an estimate of the probability of event occurrence at each time point:

$$S(t) = Pr[T \geq t], \quad (3)$$

- Hazard Function: it is denoted as  $h(t)$ , and expresses the conditional probability that the event will occur within  $[t, t+dt)$ , given that it has not occurred before:

$$h(t) = \lim_{dt \rightarrow 0} \frac{Pr[t \leq T < t + dt | T \geq t]}{dt}, \quad (4)$$

and  $\int_0^t h(u) du$  is the cumulative hazard function;

- by subdividing the time axis into  $J$  parts, the risk score of a sample  $x$  could be assessed as:

$$r(x) = \sum_{j=1}^J H(t_j, x). \quad (5)$$

In this work, we adopted Random Survival Forests and Boosting Machines for time-to-event analysis.

- Random Survival Forests (RSF) are an extension of random forests specifically designed for survival analysis. They combine the principles of survival analysis with the concept of decision trees. Random Survival Forests provide a flexible and powerful approach for modeling the relationship between predictors and survival times. They build a collection of decision trees that partition the data into subsets based on predictor variables and survival times. At each node of the decision tree, a splitting criterion is used to determine the best predictor variable and threshold for splitting the data. Then the bagging is applied for resampling the original data with replacement to create multiple bootstrap samples. Each decision tree is built on a different bootstrap sample, and the final prediction is obtained by averaging the predictions of all trees. Moreover, only a random subset of predictor variables is considered at each split. This random feature selection adds an additional element of randomness to the model, reducing overfitting and improving generalization.
- Boosting Machines, such as Gradient Boosting (GB) Machines create a sequence of weak learners, which are combined to form a strong learner. Boosting starts with initializing predictions for each sample. Initially, all samples have equal weights. Boosting builds a series of weak models, usually decision trees, in an iterative manner. Each weak model is fitted to the data, and the weights of the samples are updated based on the model's performance. In each iteration, the weights of misclassified or poorly predicted samples are increased, while the weights of correctly predicted samples are decreased. This allows subsequent weak models to focus more on the difficult samples. At the end stage, the weak models are combined to form a strong learner. The final prediction is obtained by taking a weighted average of the predictions from all weak models.

### 3.4. Permutation feature importance

We computed the permutation feature importance [17] to assess the importance of features for both ML models. Permutation feature importance involves the following steps:

- Train a Model: first, a machine learning model is trained using the dataset;
- Calculate Baseline Performance: the model's performance metric is computed through cross-validation. This performance metric serves as the baseline or reference.
- Permute the Feature: the values of the feature of interest are randomly shuffled while keeping other features unchanged. This results in a dataset where the values of the feature no longer reflect their original relationship with the target variable.
- Evaluate Model Performance: the permuted dataset is passed through the trained model, and the performance metric is calculated again. The new performance metric reflects the model's performance when the feature's relationship with the target variable has been disrupted.
- Compute Feature Importance: the feature importance is computed by quantifying the drop in model performance caused by permuting the feature. The larger the drop in performance, the more important the feature is considered. Feature importance can be expressed as a difference, ratio, or percentage change between the baseline performance and the performance after permutation.

Here we adopted this technique to identify which features have the most impact on the model's performance and provide insights into the underlying relationships between the features and the target variable.

### 3.5. Performance evaluation

The following metrics were adopted to assess the performance during the training phase:

- the concordance index (C-index), i.e., a generalization of the area under the ROC curve (AUC) that can take into account censored data. It estimates the probability that the order of the predictions of a pair of comparable patients is consistent with the individual risk scores [18]. It can be computed as:

$$C - index = \frac{\sum_{i,j} 1_{T_j < T_i} \cdot 1_{r_j > r_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j < T_i} \cdot \delta_j}, \quad (6)$$

where  $r_i$  is the risk score for patient  $i$ ,  $1_{T_j < T_i} = 1$  if  $T_j < T_i$  else 0;  $1_{r_j > r_i} = 1$  if  $r_j > r_i$  else 0.

- The cumulative dynamic AUC. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) can be extended to survival data by defining sensitivity (true positive rate) and specificity (true negative rate) as time-dependent measures [19]. Cumulative cases are all individuals that experienced an event prior to or at time  $t$  ( $t_i \leq t$ ), whereas dynamic controls are those with  $t_i > t$ . The associated cumulative dynamic AUC quantifies how well a model can distinguish subjects who fail by a given time ( $t_i \leq t$ ) from subjects who fail after this time ( $t_i > t$ ).

**Table 1**

C-index of the ML models averaged across the repeated cross-validation rounds

Model	Static dataset A	Complete dataset A	Static dataset B	Complete dataset B
RSF	0.49	0.65	0.5	0.53
RSF SMOTE	0.47	0.58	0.48	0.51
GB	0.53	0.55	0.55	0.53
GB SMOTE	0.53	0.53	0.56	0.46

## 4. Experimental Setup

We organized the setup of the experiments to address the following research questions:

- RQ1 What is the impact of the static features on the model performance for the two datasets A and B?
- RQ2 What is the additional contribution of the dynamic features?
- RQ3 Is there a significant ranking of all the features for both definitions of MS worsening?
- RQ4 Does the imbalance of the types of events ('censored' and non-censored) affect the performance of the models?

Accordingly, 16 models were trained, i.e. two RSF models (without SMOTE and with SMOTE) and two GB models (without SMOTE and with SMOTE) for each dataset for both the partial dataset, including only static features and the total dataset, including also dynamic features. We have run all the experiments on Google Colab [20]. The ML models have been implemented by using the Python package scikit-survival-0.20.0 [21]. All submissions were completed with the "SisInflab-AIBio" team.

## 5. Results

### 5.1. Cross-validation results

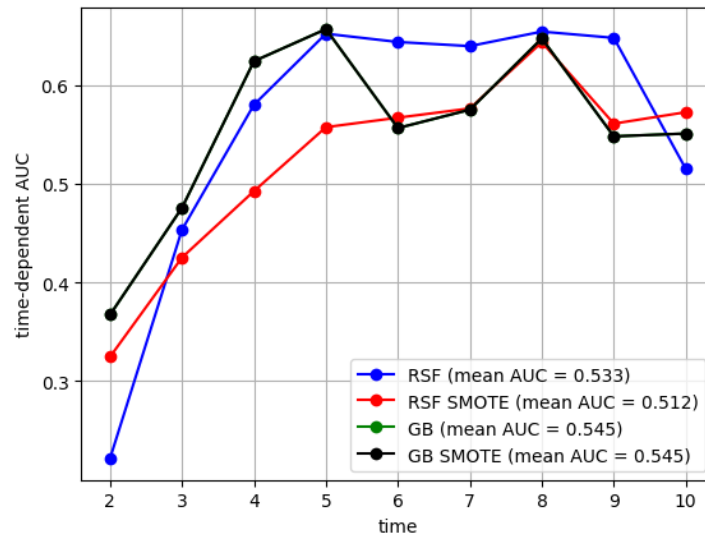
#### RQ1: impact of the static features on the performance

As shown in Figure 3 and Table 1, RSF models perform worse than GB models for dataset A. The same can be observed for dataset B (see Figure 4). It is worth noting, however, that the best performance is close to chance level, highlighting that, on their own, static features fail to predict the MS worsening and that the dynamic information should therefore be exploited.

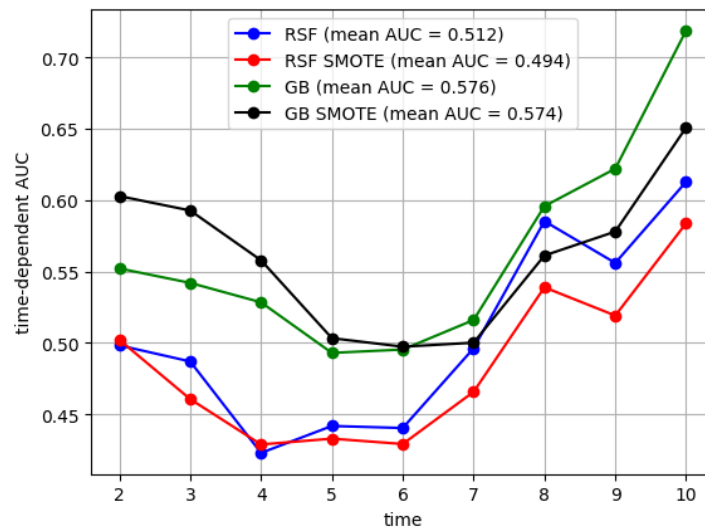
#### RQ2: the additional contribution of the dynamic features

As shown in Figure 5 and Table 1, for dataset A, RSF models perform better than GB models. Furthermore, these models perform better than all ML models trained using only static data. This finding is not observed for dataset B: as it turns out from Figure 6 only the RSF model performs slightly better than the GB models trained using the static dataset.





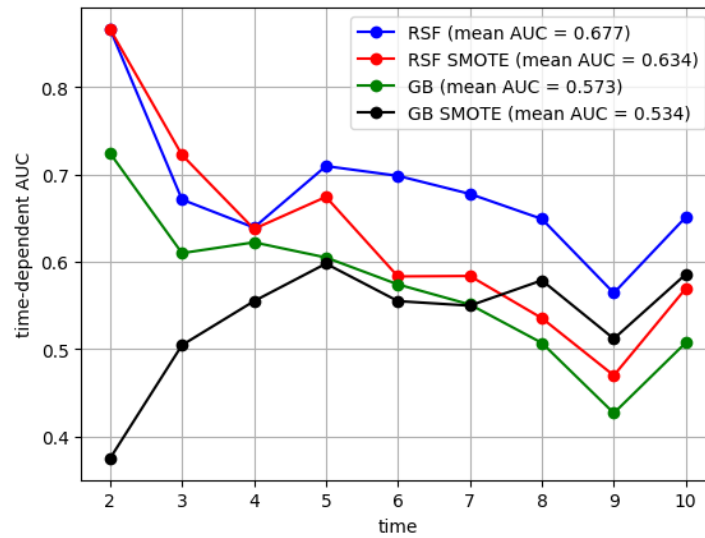
**Figure 3:** Temporal AUC for all the ML models for the static variables of dataset A averaged across the repeated cross-validation rounds.



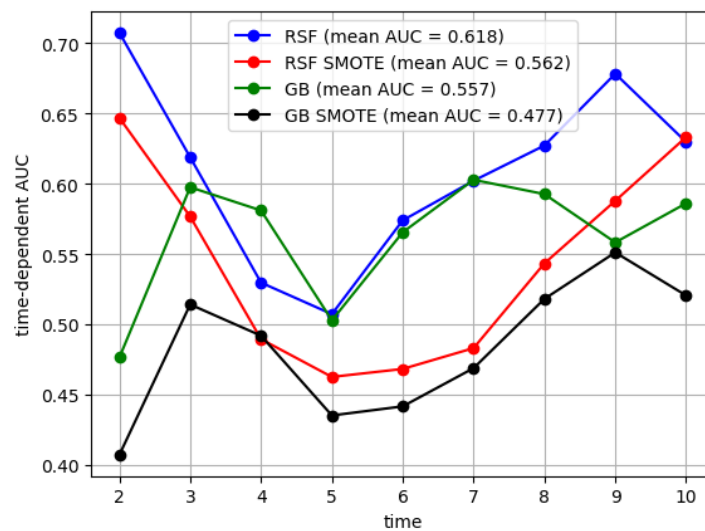
**Figure 4:** Temporal AUC for all the ML models for the static variables of dataset B averaged across the repeated cross-validation rounds.

### RQ3: ranking of the features for both tasks

The RSF models obtained without using the SMOTE technique were found to be the best for both datasets. We obtained the feature ranking for the two best models using the permutation feature importance. It can be observed in Figure 7 that for dataset A, most of the temporal features of the EDSS score appear as the most important together with the information on the onset of

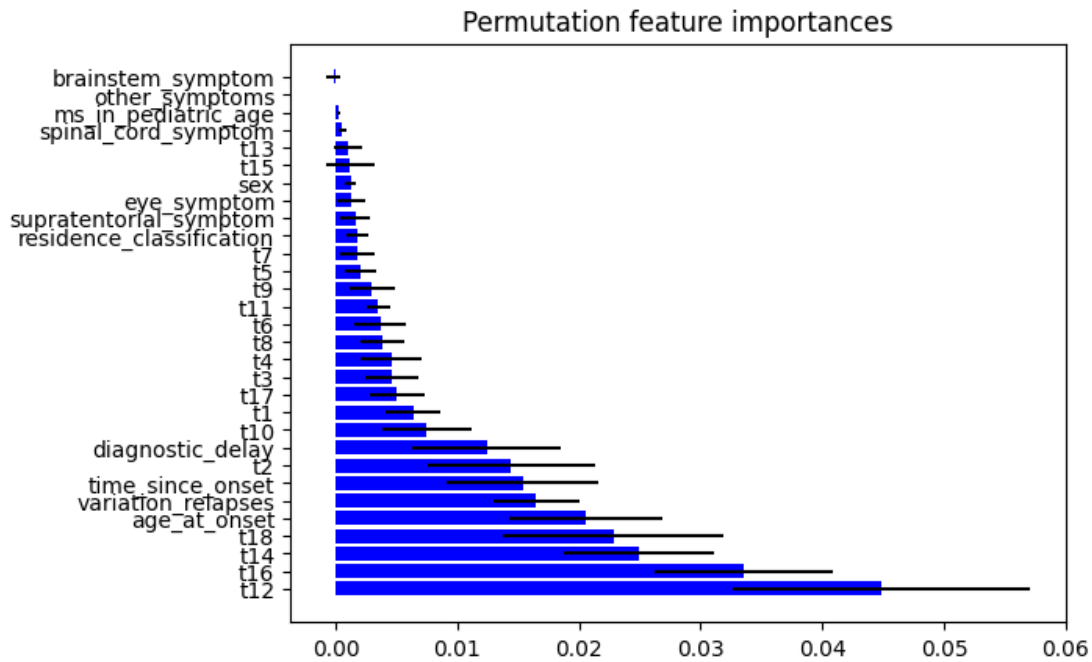


**Figure 5:** Temporal AUC for all the ML models for dataset A averaged across the repeated cross-validation rounds.



**Figure 6:** Temporal AUC for all the ML models for dataset B averaged across the repeated cross-validation rounds.

the disease and the feature *variation\_relapses* representing the relative frequency of relapses. For dataset B, which has a different definition of worsening, it can be observed in Figure 8 that only a few temporal points of the EDSS scores appear as significantly impacting performance, while information on disease onset and the feature *variation\_relapses* are confirmed relevant features.



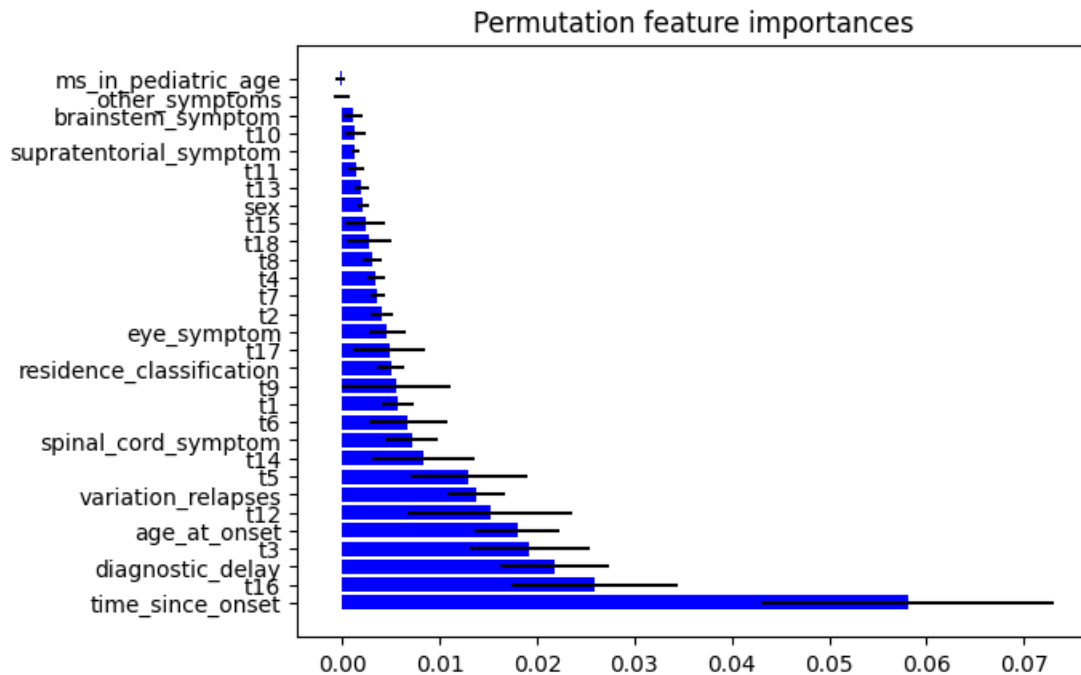
**Figure 7:** Permutation feature importances for the best model RSF for dataset A.

#### RQ4 effect of class imbalance on the performance

As shown in Table 1 and Figures 5 and 6, no improvements in the performance of ML models occur using SMOTE, showing that this technique is not suitable for balancing the classes of censored and not-censored events.

## 6. Conclusions and Future Work

In this paper, we presented a framework for the interpretable analysis of ML algorithms to predict the MS worsening using the datasets provided by the iDPP@CLEF 2023 Challenge. The proposed framework is modular and allows to investigate the link between the provided static and dynamic features and the outcome to be predicted. Our findings show that for a more complex and condition-dependent definition of worsening (tasks 1b and 2b) significantly lower results are obtained than those obtained with a simpler definition of worsening (tasks 1a and 2a). In our work, only two ML methods with a landmark approach were considered, thus different algorithms such as those based on deep neural networks that can automatically model time series of different lengths will be exhaustively explored in future work. In addition, we excluded information about the patients' centre, which could instead play a key role in predicting the course of the disease. Future developments will involve the use of site harmonization algorithms prior to time-to-event analysis to remove potential bias related to this variable.



**Figure 8:** Permutation feature importances for the best model RSF for dataset B.

## References

- [1] S. Llufrui, Y. Blanco, E. Martinez-Heras, J. Casanova-Molla, I. Gabilondo, M. Sepulveda, C. Falcon, J. Berenguer, N. Bargallo, P. Villoslada, et al., Influence of corpus callosum damage on cognition and physical disability in multiple sclerosis: a multimodal study, *PloS one* 7 (2012) e37167.
- [2] M. Filippi, M. A. Rocca, O. Ciccarelli, N. De Stefano, N. Evangelou, L. Kappos, A. Rovira, J. Sastre-Garriga, M. Tintorè, J. L. Frederiksen, et al., Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines, *The Lancet Neurology* 15 (2016) 292–303.
- [3] G. Di Maggio, R. Santangelo, S. Guerrieri, M. Bianco, L. Ferrari, S. Medagliani, M. Rodegher, B. Colombo, L. Muiola, R. Chieffo, et al., Optical coherence tomography and visual evoked potentials: which is more sensitive in multiple sclerosis?, *Multiple Sclerosis Journal* 20 (2014) 1342–1347.
- [4] F. Moazami, A. Lefevre-Utile, C. Papaloukas, V. Soumelis, Machine learning approaches in study of multiple sclerosis disease through magnetic resonance images, *Frontiers in immunology* 12 (2021) 700582.
- [5] M. F. Pinto, H. Oliveira, S. Batista, L. Cruz, M. Pinto, I. Correia, P. Martins, C. Teixeira, Prediction of disease progression and outcomes in multiple sclerosis with machine learning, *Scientific reports* 10 (2020) 1–13.
- [6] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, *Random survival forests* (2008).

- [7] D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez, P. J. O'Connor, Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting, *Journal of biomedical informatics* 61 (2016) 119–131.
- [8] Y. Zhao, T. Wang, R. Bove, B. Cree, R. Henry, H. Lokhande, M. Polgar-Turcsanyi, M. Anderson, R. Bakshi, H. L. Weiner, et al., Ensemble learning predicts multiple sclerosis disease course in the summit study, *NPJ digital medicine* 3 (2020) 135.
- [9] S. Fiorini, A. Verri, A. Tacchino, M. Ponzio, G. Bricchetto, A. Barla, A machine learning pipeline for multiple sclerosis course detection from clinical scales and patient reported outcomes, in: 2015 37th Annual International Conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2015, pp. 4443–4446.
- [10] 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.
- [11] 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 Domínguez, 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. Tsirikla, 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.
- [12] 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 Domínguez, 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.
- [13] H. van Houwelingen, H. Putter, *Dynamic prediction in clinical survival analysis*, CRC Press, 2011.
- [14] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: *Ijcai*, volume 14, Montreal, Canada, 1995, pp. 1137–1145.
- [15] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, *BMC bioinformatics* 7 (2006) 1–8.
- [16] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* 16 (2002) 321–357.
- [17] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, *Bioinformatics* 26 (2010) 1340–1347.
- [18] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, L.-J. Wei, On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data, *Statistics in medicine* 30 (2011) 1105–1117.

- [19] H. Hung, C.-T. Chiang, Estimation methods for time-dependent auc models with survival data, *Canadian Journal of Statistics* 38 (2010) 8–26.
- [20] E. Bisong, E. Bisong, Google colaboratory, Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners (2019) 59–64.
- [21] S. Pölsterl, scikit-survival: A library for time-to-event analysis built on top of scikit-learn, *The Journal of Machine Learning Research* 21 (2020) 8747–8752.