Predicting Multiple Sclerosis Relapses Using Patient Exposure Trajectories

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2024

Pietro Bosoni[1,*], Mahin Vazifehdan[1], Daniele Pala[1], Eleonora Tavazzi[2], Roberto Bergamaschi[2], Riccardo Bellazzi[1] and Arianna Dagliati[1]

[1] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100 Pavia, Italy*
[2] *Multiple Sclerosis Centre, Department of Clinical Neurology, Neurological Institute C. Mondino, Via Mondino 2, 27100 Pavia, Italy*

Abstract

Multiple Sclerosis (MS) is a prevalent autoimmune neurodegenerative disease characterized by progressive nerve inflammation, leading to increasingly severe symptoms. Approximately 85% of MS cases exhibit a relapse-remitting pattern, where sudden symptom exacerbations (relapses) are followed by periods of improvement (remissions). Previous research has shown that MS progression is influenced by external factors such as weather and air pollution. In this paper, we present a Machine Learning-based approach to predict the timing of MS relapses based on environmental exposure. This work was conducted as part of the Intelligent Disease Progression Prediction (iDPP) CLEF 2024 Challenge, which focused on the impact of environmental exposure on MS progression using retrospective data. Specifically, we utilized two anonymized datasets from clinical institutions in Pavia and Turin, Italy, containing real patient data. We employed Topological Data Analysis to compute personal exposure trajectories and used two predictive approaches, one based on the application of Linear Regression, Random Forest, and Extreme Gradient Boosting models to the last follow-up data, and one based on the application of Mixed-Effects modeling on longitudinal data from the first to the last follow-up. Results suggest that integrating environmental variables yields valuable insights for predicting MS relapses, emphasizing the need for improved methods of calculating personal pollution exposure patterns to enhance the accuracy of MS progression predictions.

Keywords

Multiple Sclerosis, Relapse Prediction, Environmental Data, Personal Exposure Trajectory, Topological Data Analysis, Mixed-Effects Model, Machine Learning

1. Introduction

Multiple Sclerosis (MS) is a dysimmune, demyelinating, inflammatory and neurodegenerative disease affecting approximately three million individuals worldwide [1]. Although it can manifest at any age, its onset is typically during the late twenties to early thirties. MS is

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
* Corresponding author.
pietro.bosoni@unipv.it (P. Bosoni); mahin.vazifehdan01@universitadipavia.it (M. Vazifehdan); daniele.pala@unipv.it (D. Pala); eleonora.tavazzi@mondino.it (E. Tavazzi); roberto.bergamaschi@mondino.it (R. Bergamaschi); riccardo.bellazzi@unipv.it (R. Bellazzi); arianna.dagliati@unipv.it (A. Dagliati)
0000-0002-1431-6044 (P. Bosoni); 0009-0003-0506-306X (M. Vazifehdan); 0000-0001-5741-3349 (D. Pala); 0000-0002-1397-511X (D. Pala); 0000-0003-2580-3646 (E. Tavazzi); 0000-0002-1397-511X (R. Bergamaschi); 0000-0002-6974-9808 (R. Bellazzi); 0000-0002-5041-0409 (A. Dagliati)

characterized by the development of demyelinating lesions (plaques) in the central nervous system, associated with a widespread and subtle inflammatory and degenerative tissue damage. These pathological events can manifest with multifunctional symptoms, including fatigue, muscle weakness, numbness, coordination and balance issues, as well as cognitive impairment. The disease progression of MS often follows a non-linear trajectory, characterized by episodes of symptom exacerbation known as relapses, followed by complete or incomplete remission [2]. The Expanded Disability Status Scale (EDSS) is a method used to quantify disability and monitor changes in disability levels over time. The scale ranges from 0 to 10 in 0.5-unit increments, with higher scores indicating greater levels of disability [3].

While the exact cause of MS remains elusive, it is widely believed to result from a complex interplay of genetic predisposition and exposure to environmental and lifestyle-related factors [4,5]. Understanding the intricate mechanisms underlying disease onset and progression is crucial for devising effective prevention strategies, but it poses significant challenges due to the complexity of these interactions. Machine Learning (ML) and Artificial Intelligence (AI) offer promising avenues for addressing these challenges by analyzing large datasets to discern disease patterns, model causal relationships, and forecast outcomes.

This paper presents an ML-based approach aimed at evaluating whether exposure to different pollutants and weather factors can be useful in predicting MS relapses. The research was conducted as part of the Intelligent Disease Progression Prediction (iDPP) CLEF 2024 Challenge [6,7], which presented two evaluation tasks on forecasting ALS progression and one task on predicting MS relapse. Specifically, we focused on the MS task, predicting the week of the first relapse after the baseline considering environmental data and EDSS scores [3] based on a weekly granularity, given the status of the patient at the baseline.

2. Materials and Methods

The task outlined in this paper was performed using two retrospective datasets. After a comprehensive preprocessing phase to address missing data and align longitudinal measurements, these datasets were analyzed using a combination of ML algorithms. This section details the datasets utilized, the preprocessing procedures implemented, and the methodologies applied for the prediction task.

2.1. Datasets and Preprocessing

Two retrospective datasets were provided for training and testing purposes. Both datasets originate from two clinical institutions in Italy, one in Pavia and the other in Turin, and contain fully anonymized data on real MS patients. The training dataset comprises 834 observations from 199 patients, while the test dataset includes 290 observations from 81 patients, provided on two different dates according to the iDPP CLEF 2024 Challenge. These datasets contain both static (demographic and clinical) and longitudinal data on EDSS scores, together with functional systems (FS) sub-scores, and environmental exposure measurements, all times-indexed in weeks from baseline (the first available follow-up visit in the study period).

The preprocessing procedure was structured into three main phases. In the initial phase, static variables such as age at onset and age at baseline were discretized, using medians as cutoff points. Additionally, residence classification was refactored into two categories: rural areas and cities/towns. The second phase involved applying appropriate methods to handle missing data

in the datasets. Finally, the third phase aligned the longitudinal data by follow-ups and computed measures for personal exposure.

2.1.1. Missing Data Handling

For static data, variables with more than 7% missing data in the training set, such as ethnicity, diagnosis criteria, and details about diagnosis criteria, were excluded. For longitudinal data, various imputation methods were evaluated separately for FS scores and environmental features to determine the most effective approach for subsequent analyses. Imputation methods assessed for FS scores features included k-Nearest Neighbors (k-NN), K-Means, Multiple Imputation by Chained Equations (MICE) Forest, and MissForest. For environmental features, methods evaluated included Last Observation Carried Forward (LOCF), Linear Interpolation, k-NN, Hot Deck, and MICE Forest.

Complete cases from the training dataset were selected, and missing values were artificially introduced at a rate of 10% to simulate missing data scenarios. The performance of these imputation methods was assessed using the Root Mean Square Error (RMSE) metric.

2.1.2. Personal Exposure Computation

A second preprocessing step involved aligning longitudinal data based on the follow-ups temporal granularity. Specifically, we calculated the mean of each environmental factor's measurements between consecutive follow-ups for each patient (or for the four weeks preceding the first recorded follow-up), assigning this aggregated value as the cumulative exposure on the follow-up date. Subsequently, longitudinal data were standardized.

Next, we employed Topological Data Analysis (TDA) to compute multivariate personal exposure trajectories between consecutive follow-ups, using the mapper2D function from the TDA Mapper R-package [8], integrated into the TDA-PseudoTime code [9]. TDA offers an analytical method for complex data, identifying shape characteristics that remain robust despite rescaling distances, thus providing a qualitative description of the data. Each identified trajectory was utilized as a binary variable to indicate whether the patient falls within that specific exposure category or not.

To compute Topological Maps using TDA, we defined observation similarity by computing the cosine distance of aggregated environmental values assigned at each follow-up date. Filter functions, including single-value decomposition and L1-infinity centrality, were used to project data points into coordinate spaces describing data distribution. The resulting projections were then partitioned into overlapping bins, based on resolution parameters, and clustering was performed within each bin. We explored various resolution parameters using grid search, adjusting the number of bins, their overlap, and geometric scale (number of clusters per bin).

TDA analysis revealed a complex structure comprising multiple trajectories that delineate distinct exposure patterns across follow-up periods. We applied a minimum spanning tree filter to identify the shortest paths within the topology, weighting edges by temporal sequence of observations starting from the initial follow-up (t=0).

In addition to incorporating them into a multivariate exposure trajectory, we computed additional aggregating features for each patient by calculating the mean and the slope coefficient of environmental factors across all follow-ups, aiming to explore the average influence of environmental factors and their temporal changes over time.

2.2. Models Development

The goal of the MS task was to predict the week of the first relapse after the baseline for each patient, given that all patients are guaranteed to experience at least one relapse post-baseline. Our idea was to compute an additional feature (delta) in the training set, representing the time difference in weeks between consecutive follow-ups and, for the last follow-up, the difference between the last follow-up and the first relapse. Therefore, models were trained to predict this delta, which, when added to the number of weeks from baseline to the previous follow-up, would represent the outcome. For models development, we adopted two strategies:

- ***Last-Observation*** taking a snapshot of clinical and exposure data at the last follow-up and training Linear Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGB) Models.
- ***Mixed-Effects*** using longitudinal data from the first to the last follow-up and applying Mixed Effects Models.

Furthermore, we considered three scenarios with different subsets of variables: ***First*** with TDA trajectories and static features (including demographic and clinical information); ***Second*** with TDA trajectories, static features, and environmental slope features; ***Third*** with TDA trajectories, static features, and environmental mean features. The optimal subset of features for the MS task was selected based on RMSE computed on the training set using the Last-Observation strategy.

2.2.1. Last-Observation Models Training

To optimize model hyperparameters and select the optimal feature subset, we split the training dataset into further training and test sets (80%-20%) using stratified sampling. A LR model was employed as a benchmark. Then, using a grid search, we fine-tuned two of the most used models for regression problems, such as RF and XGB. Specifically, the trainControl function from the caret R-package [10] was used to perform a 10-fold cross-validation as a resampling strategy, iteratively creating training and validation sets. Finally, RF and XGB models in their optimal configurations for each scenario, along with the LR benchmark model, were evaluated on the 20% test data to determine the best framework for predicting the week of the first relapse on the test set provided by the iDPP CLEF 2024 Challenge [6,7].

For RF, the parameter controlling the number of variables randomly sampled as candidates at each split of the algorithm was varied in the range [2-15]. For XGB, we varied several hyperparameters on a grid, including number of trees (200 to 1000 with a 50-step), maximum tree depth (2, 3, 4, 5), learning rate (0.025, 0.05, 0.1, 0.3), gamma for regularization (0, 1, 2, 3), column sampling (0.25, 0.5, 0.75, 1.0), and minimum leaf weight (1, 2, 3, 4).

2.2.2. Mixed-Effects Models Training

We used a linear Mixed-Effects model to analyze all repeated observations for each patient, employing the *lmer* function from the *lme4* R-package [11].

Initially, we trained a baseline model where delta was the response variable. Then, we explored different models configurations incorporating random slopes, random intercepts, and

both random intercepts and slopes, as described in Equation 1, Equation 2, and Equation 3, respectively.

1. Random Slope Only Model, which specifies a random slope for week_from_baseline without a random intercept:
   $$\text{delta} \sim \text{week\_from\_baseline} + (0 + \text{week\_from\_baseline} \mid \text{patient\_id}) \quad (1)$$
2. Random Intercept Only Model, which specifies a random intercept for patient_id without a random slope:
   $$\text{delta} \sim \text{week\_from\_baseline} + (1 \mid \text{patient\_id}) \quad (2)$$
3. Random Intercept and Slope Model, which includes both random intercepts and slopes for the week_from_baseline, allowing the model to account for individual differences in the baseline levels and the rates of change over time for each patient:
   $$\text{delta} \sim \text{week\_from\_baseline} + (1 + \text{week\_from\_baseline} \mid \text{patient\_id}) \quad (3)$$

The performance of these models was evaluated using metrics such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Log-Likelihood (logLik), and deviance. Following the feature selection strategy outlined in Section 2.2.1, we included the same fixed effects as covariates used in the Last-Observation models. Finally, to assess the models' ability to predict relapse week, we applied them to the first and last observations, predicting the relapse delta from the first and from the last follow-up.

3. Results

3.1. Imputation of Missing Longitudinal Data

Table 1 summarizes the parameter settings and performance of each imputation method applied to the FS scores and environmental features in the training dataset. The RMSE results indicate that k-NN and MICE Forest achieved the best performances among the imputation approaches for both types of features, with k-NN being particularly effective for environmental features. Consequently, we selected the k-NN method to impute the missing longitudinal data in both the training and test sets.

Table 1
Performance of the implemented imputation methods. Methods with the lowest Root Mean Square Error (RMSE) for each feature type are in bold

Features type	Method	Parameters	RMSE
FS scores	K-Means	cluster = 3 Max Iteration = 300 Algorithm = Lloyd Forgy	0.032
FS scores	MissForest	Tree = 100 Max depth = 15	0.005
FS scores	MICE Forest	Iteration = 3 Algorithm = LightGBM	0.0045
FS scores	**k-NN**	Neighbor = 5 Weight = Inverse distance	**0.0041**

Environmental	LOCF		25.89
Environmental	Hot Deck	Neighbor = 1	23.47
Environmental	Linear Interpolation		16.26
Environmental	MICE Forest	Iteration = 3 Algorithm = LightGBM	12.97
Environmental	**k-NN**	Neighbor = 5 Weight = Inverse distance	**6.68**

3.2. Personal Exposure Trajectories

The TDA algorithm identified three main trajectories, designated based on the topological structure of the networks as Traj4 (26 patients), Traj4_2 (74 patients), and Traj4_3 (99 patients). We compared the patient groups identified by each trajectory in terms of exposure to Ozone (O_3), Sulfur Dioxide (SO_2), and Wind Speed over the weeks from baseline. Figure 1 shows that the smoothed conditional means of O_3 reveal a decreasing exposure for subjects in the green Traj4 trajectory. In contrast, subjects in the orange Traj4_2 and blue Traj4_3 trajectories maintain a more stable exposure over time.

Figure 2 illustrates that while all three trajectories begin with similar SO_2 values, the green Traj4 trajectory experiences a rapid decline followed by an increase, ultimately stabilizing at lower exposure levels compared to the more consistent patterns observed in the orange Traj4_2 and blue Traj4_3 trajectories. In Figure 3, all three trajectories exhibit a general decreasing trend in Wind Speed.

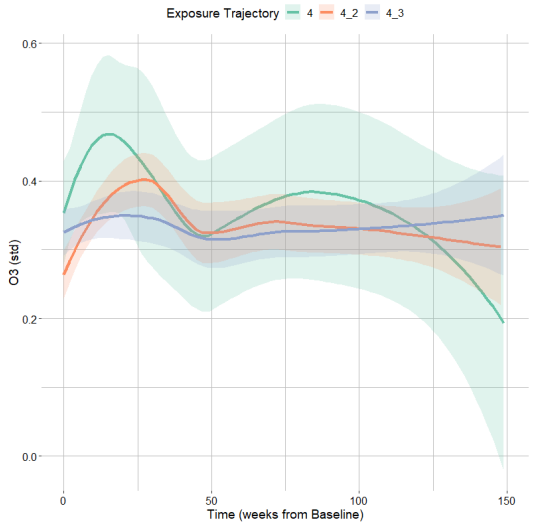

Figure 1: Ozone (O_3) standardized values in weeks from baseline for subjects belonging to different exposure trajectories.

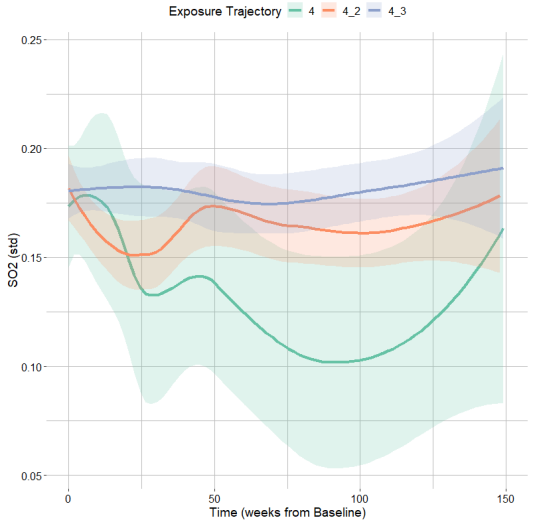

Figure 2: Sulfur Dioxide (SO_2) standardized values in weeks from baseline for subjects belonging to different exposure trajectories.

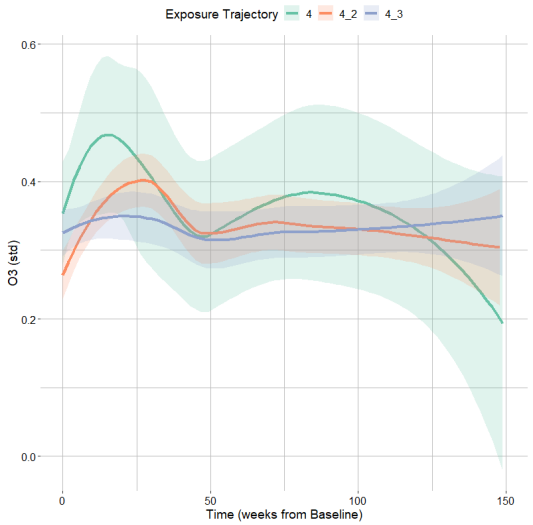

Figure 3: Wind Speed standardized values in weeks from baseline for subjects belonging to different exposure trajectories.

In our approach to computing TDA on the test dataset, we maintained consistency by using the same TDA training parameters, which resulted in the same number of clusters. Although this method might not be optimal, it allowed for a direct comparison between the findings from the training and test datasets. We identified clusters with similar clinical progressions by visually inspecting plots of the average EDSS scores over time within the three clusters (Figure 4 and Figure 5). Based on this comparison, we retained clusters that exhibited similar clinical trajectories, ensuring the relevance of the clusters identified during the training phase.

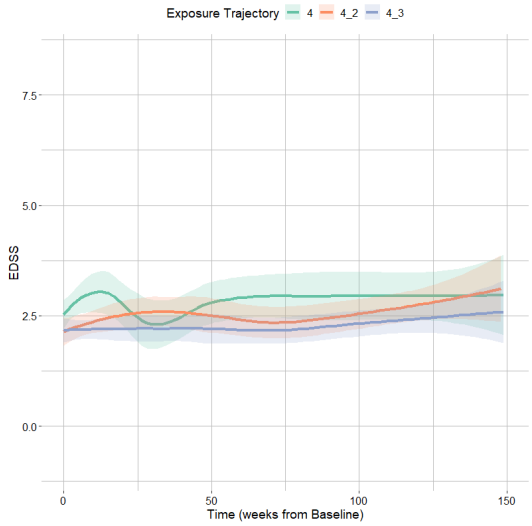

Figure 4: Expanded Disability Status Scale (EDSS) scores in weeks from baseline for subjects belonging to different exposure trajectories, considering the training set.

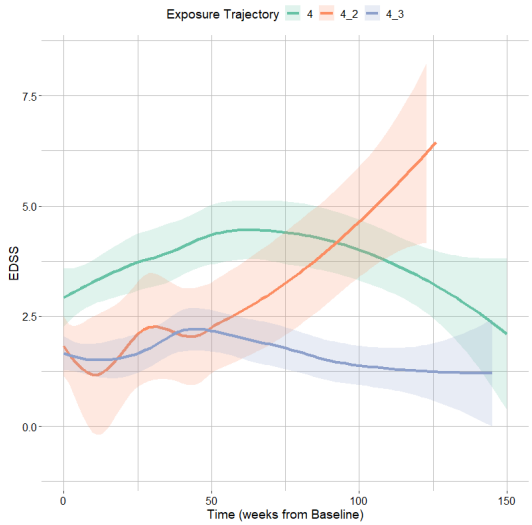

Figure 5: Expanded Disability Status Scale (EDSS) scores in weeks from baseline for subjects belonging to different exposure trajectories, considering the test set.

While this approach seems feasible, we recognize that the manual steps involved could be refined for enhanced accuracy. A more robust strategy would entail utilizing the same topology of the training set and subsequently assigning subjects to the original trajectories based on the Jaccard distance. This refined methodology offers a more systematic and objective approach, minimizing the potential biases inherent in manual cluster identification.

3.3. Last-Observation Model Selection

Table 2 presents the model performances using the Last-Observation strategy on the training set across three scenarios with different feature subsets. The best performance was achieved by

the RF model in the Second scenario, which includes TDA trajectories, static features, and environmental slope features. Overall, the RF model consistently outperformed the XGB and LR models across all three scenarios, with the optimal number of variables randomly sampled as candidates at each split (*mtry*) consistently being equal to 2.

Hyperparameter tuning for the XGB model yielded consistent values across the three scenarios: number of trees = 2, maximum tree depth = 2, learning rate = 0.025, and column sampling = 0.25. The exceptions were the gamma parameter (0 in the First and Second scenarios, 3 in the Third scenario) and the minimum leaf weight (4 in the First scenario, 3 in the Second and Third scenarios). As expected, the LR benchmark model had the highest RMSE, except for the XGB model in the Third scenario.

Table 2
Model performances using the Last-Observation strategy across the three scenarios with different subsets of features (ordered by ascending RMSE)

Scenario	Model	RMSE
Second	RF	18.74
Third	RF	19.35
First	RF	19.72
First	XGB	20.17
Second	XGB	20.20
First	LR	22.11
Third	XGB	23.51
Third	LR	23.61
Second	LR	27.56

3.4. Mixed-Effects Models Selection

We evaluated three Mixed-Effects models using various metrics such as AIC, BIC, logLik, and deviance, as reported in Table 3. Based on the metrics provided, the Random Intercept and Slope Model was identified as the best choice.

Table 3
Mixed-Effects Models performances

Model	Parameters number	AIC	BIC	logLik	Deviance
Random Slope Only	4	9440.9	9459.8	-4716.4	9432.9
Random Intercept Only	4	11012.8	10993.9	5510.4	-11020.8
Random Intercept and Slope	6	2482.6	2454.3	1247.3	-2494.6

In the final Random Intercept and Slope Model, both fixed and random effects were considered to predict the response variable, delta. The model included data from 834 observations across 199 patients. The random effects showed significant variance in intercepts among patients (variance = 1324, standard deviation = 36.386) and minimal variance in slopes for week_from_baseline (variance = 3.958e-08, standard deviation = 0.000199), with a perfect

correlation (1.00) between intercept and slope. The residual variance was 31.22 (standard deviation = 5.587).

For the fixed effects, several predictors were found to be significant. Table 4 highlights the significant fixed effects in the final model, indicating the influence of various clinical and environmental factors, as well as exposure trajectories, on the delta values.

Table 4
Statistically significant fixed effects in the Random Intercept and Slope Model

Feature	Estimate	Standard Error	t-value	Probability(>|t|)
centre	3.127e+01	6.554e+00	4.770	2.18e-06 ***
residence_classification	1.699e+01	6.567e+00	2.587	0.009870 **
age_at_onset	-1.702e+01	6.349e+00	-2.681	0.007500 **
spinal_cord_symptom	2.331e+01	8.733e+00	2.670	0.007747 **
CO_slope	1.444e+02	5.511e+01	2.620	0.008968 **
humidity_slope	-1.297e+02	6.033e+01	-2.150	0.031841 *
sealevel_pressure_slope	8.559e+01	3.723e+01	2.299	0.021774 *
precipitation_sum_slope	1.507e+02	6.684e+01	2.255	0.024395 *
clusterTraj4_2	3.632e+01	8.857e+00	4.101	4.54e-05 ***
clusterTraj4_3	3.012e+01	8.514e+00	3.537	0.000428 ***

3.5. Relapse Prediction Performances

We trained a RF model with *mtry* = 2 using the second subset of features, which includes TDA trajectories, static features, and environmental slope features, to forecast the first relapse after baseline. This identical feature set was used to train the Mixed-Effect Model, applying it to both the First Observation (LMER_First_Obs) and the Last Observation (LMER_First_Obs). We evaluated these three models using the test set provided by the iDPP CLEF 2024 Challenge, with performance outcomes detailed in Table 5. The RF model consistently delivered superior results in both RMSE and MAE metrics.

Table 5
Performances of the selected Models on the test set

Model	RMSE	MAE
RF	41.52	22.49
LMER_First_Obs	48.07	28.05
LMER_Last_Obs	72.51	47.74

4. Discussion and Conclusions

MS can be debilitating and challenging to predict, influenced by numerous external factors that are often difficult to identify and quantify. MS progression is known to be partially related to the occurrence of relapses, with each relapse impacting on the patient's overall condition. Predicting relapses is essential for improving disease prognosis, but it is complicated by the variable nature of MS and the challenge of modeling the combined influence of external and biological factors. The iDPP CLEF 2024 Challenge addresses this issue by focusing on the impact

of exposure to pollutants on MS progression. Indeed, one of the tasks proposed in the Challenge aims to predict the week of the first relapse after baseline using environmental data and EDSS scores, representing a potential step for developing novel approaches to enhance MS patient care in relation to environmental exposures.

In this paper, we analyzed two retrospective datasets to predict the week of first relapse after the baseline based on environmental exposure, employing two strategies with different feature subsets. Under the Last-Observation strategy, the RF model yielded the best results when considering TDA trajectories, static features, and environmental slope features at the last follow-up. Its predictions were comparable to those of Mixed Effects Model applied to the first follow-up. While the prediction results can be improved, they are promising in terms of the significance of the variables used. The Mixed-Effects Model results indicate that several environmental variables' coefficients are statistically significant predictors of relapses, highlighting the importance of integrating environmental exposures into ML models to accurately characterize MS progression.

The difficulty in obtaining accurate predictions likely stems from the complex nature of MS, which involves numerous interacting factors and significant individual variability, as well as the approximations required to compute pollution exposure. Environmental monitoring often relies on sensors located in specific areas or satellite data aggregated over time and space, missing short-term or localized exposure peaks. Enhancing prediction accuracy requires more precise personal exposure computation, which can be achieved by collecting data with finer granularity and portable sensors.

Acknowledgements

This research was supported by the BRAINTEASER (Bringing Artificial Intelligence home for a better care of amyotrophic lateral sclerosis and multiple sclerosis) project, funded by the European Union's Horizon 2020 research and innovation programme (Grant Agreement number: 101017598).

References

1. Atlas of MS, 2024. URL: https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms.
2. Benedict RH, Pol J, Yasin F, Hojnacki D, Kolb C, Eckert S, Tacca B, Drake A, Wojcik C, Morrow SA, Jakimovski D, Fuchs TA, Dwyer MG, Zivadinov R, Weinstock-Guttman B. Recovery of cognitive function after relapse in multiple sclerosis. Mult Scler. 2021 Jan 1;27(1):71–8.
3. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology. 1983 Nov;33(11):1444–52.
4. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2017 Jan;13(1):25–36.
5. Bergamaschi R, Cortese A, Pichiecchio A, Berzolari FG, Borrelli P, Mallucci G, Bollati V, Romani A, Nosari G, Villa S, Montomoli C. Air pollution is associated to the multiple sclerosis inflammatory activity as measured by brain MRI. Mult Scler. 2018 Oct 1;24(12):1578–84.
6. Birolo G, Bosoni P, Faggioli G, Aidos H, Bergamaschi R, Cavalla P, Chiò A, Dagliati A, de Carvalho M, Di Nunzio GM, Fariselli P, García Dominguez JM, Gromicho M, Guazzo A,

Longato E, Madeira SC, Manera U, Marchesin S, Menotti L, Silvello G, Tavazzi E, Tavazzi E, Trescato I, Vettoretti M, Di Camillo B, Ferro N. Overview of iDPP@CLEF 2024: The Intelligent Disease Progression Prediction Challenge. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024. CEUR-WS.org; 2024. (CEUR Workshop Proceedings).
7. Birolo G, Bosoni P, Faggioli G, Aidos H, Bergamaschi R, Cavalla P, Chiò A, Dagliati A, de Carvalho M, Di Nunzio GM, Fariselli P, García Dominguez JM, Gromicho M, Guazzo A, Longato E, Madeira SC, Manera U, Marchesin S, Menotti L, Silvello G, Tavazzi E, Tavazzi E, Trescato I, Vettoretti M, Di Camillo B, Ferro N. {Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9th to 12th, 2024. Springer; 2024. (Lecture Notes in Computer Science).
8. TDA Mapper, 2024. URL: https://github.com/paultpearson/TDAmapper.
9. TDA PseudoTime, 2024. URL: https://github.com/aridag/TDA_PSEUDOTIME.
10. Caret: Classification and Regression Training, 2024. URL: https://cran.r-project.org/web/packages/caret/.
11. lme4: Linear Mixed-Effects Models using Eigen and S4, 2024. URL: https://cran.r-project.org/web/packages/lme4/index.html.