Predicting and Explaining Risk of Disease Worsening Using Temporal Features in Multiple Sclerosis

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

Tommaso Mario Buonocore[1], Pietro Bosoni[1], Giovanna Nicora[1], Mahin Vazifehdan[1], Riccardo Bellazzi[1], Enea Parimbelli[1], and Arianna Dagliati[1]

[1] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, Pavia, 27100*

Abstract

We present an evaluation study of the usage of two different post-hoc model agnostic XAI methods, namely SHAP and AraucanaXAI, to provide insights about the most predictive factors of worsening in MS patients, based on clinical observations carried out during a period of 2.5 years. We pre-processed the temporal features considering a Latent Class Mixed Modelling (LCMM) approach in order to discover and extract temporal trajectories as an additional informative feature. The different XAI approaches are compared according to four quantitative evaluation metrics consisting in identity, fidelity, separability and time to compute an explanation. Furthermore, a qualitative comparison of post-hoc generated explanations is carried out on specific scenarios where the ML model predicted the outcome incorrectly, in the effort to debug potentially problematic model behaviour. The combination of the qualitative and quantitative results forms the basis for a critical discussion of XAI methods properties and desiderata for healthcare applications at large, advocating for more meaningful and extensive XAI evaluation studies involving human experts.

Keywords [1]

Multiple sclerosis, neurological disease, degenerative disease, disease worsening, XAI, explainability, black-box, interpretable machine learning, predictive modelling, local explanation, surrogate model, evaluation, temporal features, temporal data mining

1. Introduction

1.1. AI to predict Multiple Sclerosis progression

Multiple sclerosis (MS) is a chronic, autoimmune disease of the central nervous system (CNS) that affects millions of individuals worldwide. It is characterized by the progressive destruction of myelin, the protective covering of nerve fibers, leading to impaired communication between the brain, spinal cord, and other parts of the body. MS exhibits a wide range of symptoms, including fatigue, muscle weakness, numbness, coordination and balance problems, and cognitive dysfunction. The etiology of MS remains elusive, with both genetic and environmental factors playing a role in its development.

The BrainTeaser project uses Artificial Intelligence to better understand MS, predict disease progression, and suggest interventions to slow it down. By harnessing AI's potential, models can be developed to predict outcomes for different patient groups, aiding in patient care and clinical trials. BrainTeaser aims to create an interpretable approach that analyzes temporal data and predicts the likelihood of adverse events. Detecting complications during disease progression is crucial for MS

[1]CLEF 2023 – Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: buonocore.tms@gmail.com (A. 1)
ORCID: 0000-0002-2887-088X (A. 1); 0000-0002-1431-6044 (A. 2); 0000-0001-7007-0862 (A. 3); 0009-0003-0506-306X (A. 4); 0000-0003-0679-828X (A. 5); 0000-0002-6974-9808 (A. 6); 0000-0002-5041-0409 (A. 7)


patients and clinicians, and including mechanisms that identify important clinical features helps predict and prevent adverse outcomes.

As indicated in Faggioli et al. [1], "worsening" is mainly defined as an increase in the Expanded Disability Status Scale (EDSS) depending on the baseline value: if baseline EDSS < 1, worsening event occurs when an increase of EDSS by 1.5 points is first observed; if 1 ≤ Baseline EDSS < 5.5, worsening event occurs when an increase of EDSS by 1 point is first observed; if baseline EDSS ≥ 5.5, worsening event occurs when an increase of EDSS by 0.5 points is first observed. In the present work we aim at building an AI/ML-based predictive model to predict a worsening event in MS patients, and in turn use such model to investigate the most predictive factors for such task.

1.2. Explainable AI

The growing interest in eXplainable AI (XAI) has led to the development of methods that provide both local and global explanations for black-box Machine Learning (ML) models [3]. While global explainability deals with a general understanding of the model classification behaviour, local explainability focuses on understanding individual predictions, which is particularly important in high-stakes applications like healthcare, and is as well put forward by regulations such as GDPR [4], the EU Artificial Intelligence Act [5], and the US Algorithmic Accountability Act.

Among the post-hoc local explainability methods, SHAP (SHapley Additive exPlanations) is one of the most commonly used. SHAP utilizes a game theoretic approach to calculate the importance of each feature in a prediction task, decomposing the final predicted probability by assigning partial, additive contributions to each feature [6]. However, explaining a model's behaviour through feature importance may be less intuitive for users with limited background in machine learning, such as clinicians. For this reason, in our previous work we introduced AraucanaXAI [7], [8], a post-hoc local explainer which uses decision tree surrogate models to provide more easily understandable local explanations, as illustrated in Figure 1.

We believe that decision trees are more suitable for delivering local explanations because their structure can be easily converted in a chain of if-then rules that are easily comprehensible to a wider range of users. In this work, we aim to compare, combine and evaluate SHAP and AraucanaXAI on real-world clinical data from MS patients collected in the context of the Brainteaser project. Despite the relative abundance of new XAI methods proposed in the literature, in fact, studies have rarely performed a quantitative comparison among different approaches, and even less have evaluated their performance using real world clinical data [9], [10].

2. Materials and Methods

2.1. Prediction Task

The iDPP CLEF 2023 [1],[2] has offered two evaluation tasks focused on predicting the progression of MS, and one position paper task on the impact of air pollution exposure on the progression of Amyotrophic Lateral Sclerosis (ALS)[2]. We chose to adopt different XAI approaches to predict the risk of worsening (Task 1a) for each MS patient, i.e., how early a subject experiences the worsening event. In this task, a patient's condition has worsened if the patient crosses the threshold EDSS ≥ 3 at least twice within a one-year interval.

2.2. Datasets and Preprocessing

Two fully anonymized datasets of 440 training patients and 110 test patients, collected from two clinical institutions (in Pavia and in Turin, Italy) were provided for the challenge. Both training and test sets included static and dynamic data, with information on EDSS, relapses, evoked potentials, Magnetic

[2] http://brainteaser.dei.unipd.it/challenges/idpp2023/

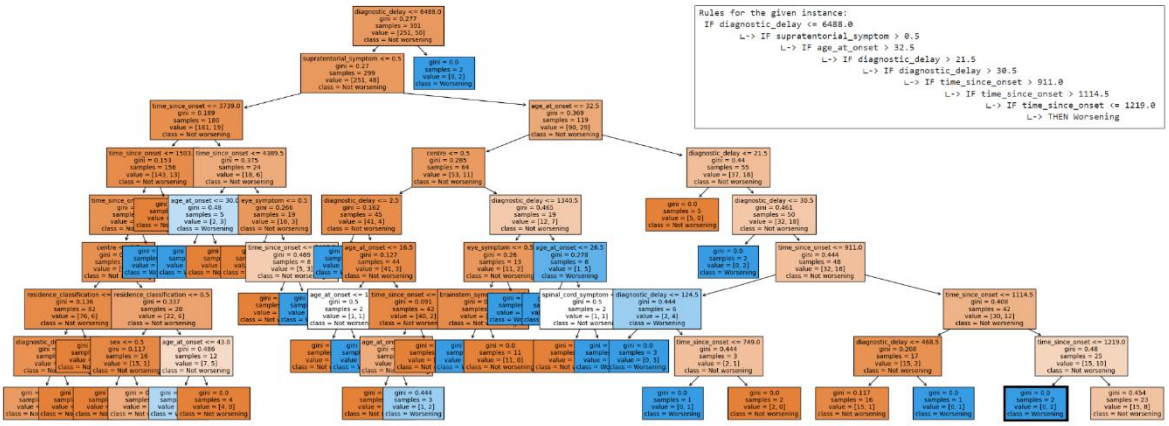

Figure 1: Example of unpruned AraucanaXAI decision tree explaining why the XGBoost model predicts the worsening of a target multiple sclerosis patient through a set of hierarchical decision rules.

Resonance Imaging (MRI), and MS course. As a first preprocessing step, we excluded 37 EDSS observations where the EDSS score measured by the clinician was missing, resulting in one patient drop; we also removed patients with missing values, which are distributed over ethnicity, residence classification, and diagnostic delay, obtaining a final training set with 401 patients.

2.3. Temporal EDSS Trajectories Analysis

As a second step, we adopted the Latent Class Mixed Modeling (LCMM) approach in order to discover and extract temporal EDSS trajectories as an additional feature. LCMM simplifies heterogeneous populations into more homogeneous clusters or classes based on mixed-effects, which are used to account for the likely correlation of repeated measurements and include a random intercept for each individual [11]. We fitted the latent class model using the "hlme" function of the R package "lcmm"[3], based on the framework described by Lennon et al. [11]. We trained the model with up to ten possible latent classes, using a linear link function and considering intercept and time ("delta_edss_time0") as fixed, random and mixture components.

Given the computational complexity, we decided to include in the model only two static features like patients' sex and age at onset as additional fixed effects over classes. A grid search with 50 departures and 10 iterations was used to set the initial values for the model parameters. For models selection, we considered both the Akaike Information Criterion (AIC), and we explored the distribution of the observations in the classes, excluding models with fewer observations in one of the latent classes.

2.4. ML and XAI Methods

XGBoost [12] was selected as the primary modeling technique for this study. XGBoost, short for Extreme Gradient Boosting, is an optimized implementation of gradient boosting that incorporates advanced features such as regularization, parallel computation, handling missing values, and early stopping. The hyperparameters (learning rate, gamma, maximum depth, column subsampling, and number of estimators) have been optimized through a random search with cross-validation, maximizing the area under the ROC curve. The risk of worsening for each patient has been calculated as the probability of the positive class (i.e., worsening). The algorithm was implemented using the "xgboost" Python library[4].

[3] https://www.rdocumentation.org/packages/lcmm/versions/2.0.2/topics/hlme
[4] https://xgboost.readthedocs.io/en/stable/python/index.html

For the explainability part, we employed two post-hoc, model-agnostic local XAI methods based on different paradigms: SHAP[5] and AraucanaXAI[6] (ARAU) and their open-source Python implementations available through the pip package manager for both approaches.

2.5. Evaluation Metrics for Predictive and Explanation Performance

The model's predictive performance in terms of risk of worsening has been evaluated using the Harrell's Concordance Index (C-Index), while the F1, precision, recall and ROC-AUC metrics have been used to evaluate the model's performance for the binary classification problem, selecting the optimal cut-off point through the Youden's index.

XAI approaches are evaluated and compared in terms of a set of metrics defined in previous research on XAI in healthcare [13]:

- Identity: if there are two identical instances, they must have the same explanations. Since our real-word dataset does not include any duplicated instance, we randomly sampled 20 examples from the test set and duplicated them to compute identity.
- Fidelity: concordance of the predictions between the XAI surrogate model and the original ML model
- Separability: if there are two dissimilar instances, they must have dissimilar explanations
- Time: average time required by the XAI method to output an explanation across the entire test set, expressed in milliseconds (ms)

3. Results

3.1. LCMM

We chose the LCMM model with five latent classes since it showed the lowest AIC value. As shown in Figure 2, these five trajectory groups included a stable-high trajectory (in red, 166 subjects), a stable medium trajectory (in khaki, 195 subjects), an increasing trajectory (in green, 23 subjects), a fast-decreasing trajectory (in blue, 24 subjects), and a slow-decreasing trajectory (in magenta, 31 subjects).

Clinical characteristics of subjects belonging to the different latent classes are reported in Table 1. A categorical characteristic like sex was compared using chi-square tests followed by Holm correction, while continuous variables like age at onset, time since onset, and EDSS scores at the baseline were compared using ANOVA with Tukey's post hoc test. Particularly, subjects in the stable-high trajectory differed significantly (p-value < 0.05) from the others in terms of time since onset (3200 days on average) and EDSS scores at the baseline (2.1 on average), showing higher values compared to all the other groups except for the fast-decreasing trajectory.

3.2. Predictive Performance

The results of our evaluation are reported in Table 2 for both the XGBoost (XGB) model and the XGBoost combined with LCMM (XGB+LCMM) model. Results refer to the test set provided by the challenge, consisting of 110 patients. Both models perform poorly on the evaluation metrics considered in the study.

[5] https://github.com/slundberg/shap
[6] https://github.com/bmi-labmedinfo/araucana-xai

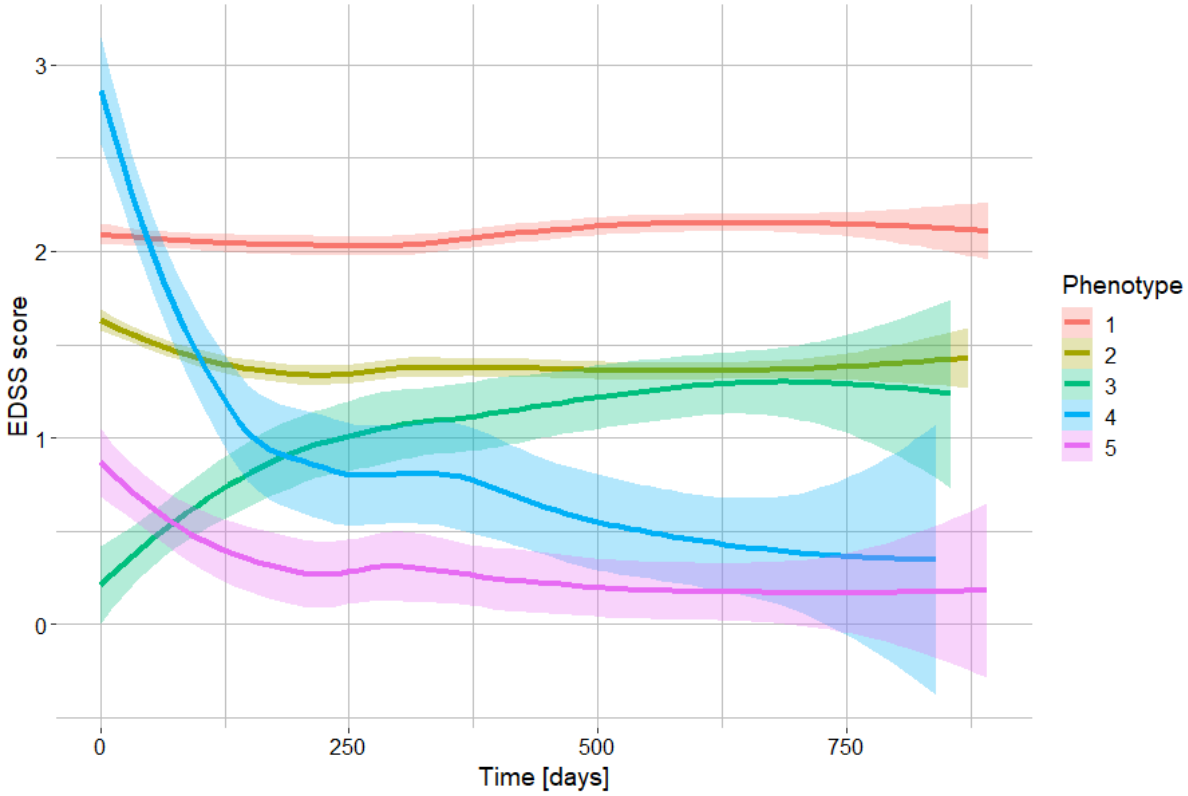

Figure 2: Latent class mixed modeling trajectories (with confidence intervals) of EDSS score on the training set.

Table 1
Clinical characteristics of subjects in the latent classes, presented as frequency (percentage), and mean ± standard deviation

	Overall	Phenotype				
		1	2	3	4	5
Sex, female	304 (69.2%)	114 (68.7%)	140 (71.8%)	14 (60.9%)	15 (62.5%)	21 (67.7%)
Age at onset (years)	30.6 ± 9.4	31.0 ± 9.6	31.2 ± 9.6	29.9 ± 9.9	26.9 ± 6.4	28.8 ± 8.8
Time since onset (days)	2520 ± 2450	3200 ± 3090	2340 ± 1870	1520 ± 1480	1220 ± 1160	1830 ± 2100
Baseline EDSS	1.8 ± 0.9	2.1 ± 0.6	1.7 ± 0.6	0.2 ± 0.5	3.0 ± 1.2	1.0 ± 0.8

Table 2
Predictive performance comparison between XGB and XGB+LCMM models

Model	C-Index (CI)	AUC-ROC	Precision	Recall	F1	Balanced Accuracy
XGB	0.46 (0.31-0.61)	0.45	0.11	0.06	0.08	0.48
XGB+LCMM	0.54 (0.38-0.70)	0.52	0.12	0.12	0.12	0.48

3.3. XAI Explanation Performance

Evaluation in terms of XAI shows that both the SHAP and AraucanaXAI methods achieved perfect fidelity and identity scores for the XGB and XGB+LCMM models. Results reported in Table 3 indicate that the explanations provided by both methods accurately represent the local behavior of the models.

The separability scores were also perfect, indicating clear distinctions between different instances. In terms of computational efficiency, both SHAP and AraucanaXAI demonstrated low time requirements per instance, with a clear advantage for SHAP averaging 0.47 milliseconds and AraucanaXAI averaging 11.61 milliseconds for the XGB model. For the XGB+LCMM model, SHAP had an average time of 0.96 milliseconds per instance, while AraucanaXAI had an average time of 10.45 milliseconds per instance. This implies the absence of significant additional computational cost when pairing the model's predictions with an explanation.

Table 3
Predictive performance comparison between XGB and XGB+LCMM models

		Fidelity		Identity		Separability		Time (ms/instance)	
		SHAP	ARAU	SHAP	ARAU	SHAP	ARAU	SHAP	ARAU
Model	XGB	1	1	1	1	.0024	.0028	0.47	11.61
	XGB+LCMM	1	1	1	1	.0000	.0002	0.96	10.45

4. Discussion

We adopted the LCMM approach to identify latent classes with different EDSS trajectories over time, extracting an additional informative feature that can be used to better predict the MS worsening. LCMM models are increasingly reported in the human epidemiology literature; indeed, they offer some advantages compared to using "one-off" exposure determinations, such as better informing aetiological associations by deeply phenotyping certain at-risk subpopulations. Furthermore, the trajectory approach allows a better understanding of the causes of between-individual variation in certain features by analysing the trajectory as an outcome rather than exposure. We chose a LCMM model with five latent classes, obtaining five well-separated trajectories.

Overall, the quantitative evaluation of the two XAI approaches over the test set indicates that both SHAP and AraucanaXAI successfully provided faithful and interpretable local explanations for the models, with SHAP being more computationally efficient. However, the underlying models have been proven to be clearly inadequate for MS worsening prediction. Results indicate that both the models underperform in terms of predictive accuracy, although the integration of LCMM successfully improved the baseline model performance. The Harrel's concordance index for both models is below the desired threshold, indicating limited ability to correctly rank the order of event probabilities. Additionally, the AUC-ROC values were around 0.5, suggesting poor discrimination between the classes. The precision, recall, and F1 scores also indicate low performance, with limited ability to correctly identify positive instances and achieve a balance between precision and recall. These findings indeed suggest the need for further investigation of the models' predictive capabilities, which can be analysed through the local explainers, to be used as a tool for model inspection and debugging. In particular, in the following we perform a more qualitatively-oriented analysis of the generated explanations for wrongly classified examples, employing both SHAP and AraucanaXAI, in the effort of understanding possible underlying reasons behind models' unsatisfactory performance across the board.

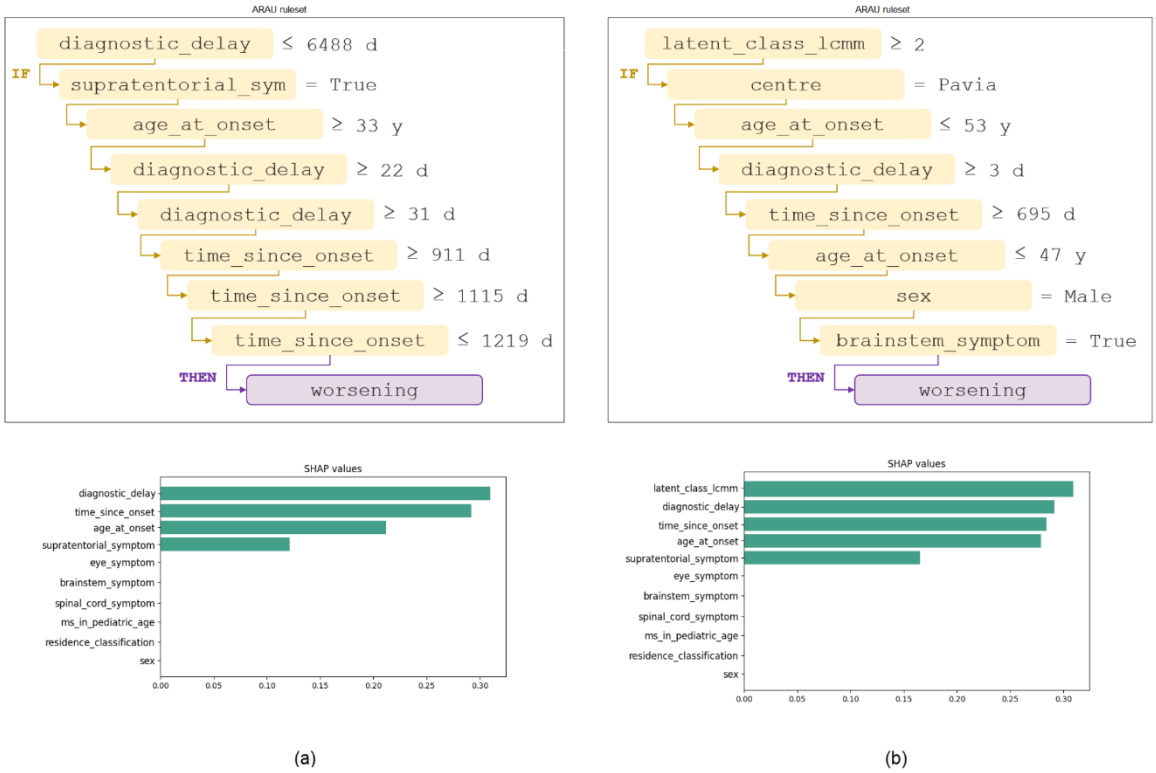

Figure 3: AraucanaXAI (top) and SHAP (bottom) explanations of a target false positive instance with the XGB model (a) and the XGB+LCMM model (b).

Table 4
Predictive performance comparison between XGB and XGB+LCMM models

Feature	Average Ranking (SHAP/ARAU)	SHAP Impact	ARAU Average Value
Diagnostic Delay	1st (2nd/2nd)	Positive	> 3 months
LCMM	2nd (4th/1st)	Positive	Class 1
Time Since Onset	2nd (1st/4th)	Positive	> 15 months
Age At Onset	4th (3rd/5th)	Negative	< 29 years
Supratentorial Symptoms	5th (8th/3rd)	Positive	FALSE

Let's consider the subset of 14 patients at high risk of worsening that have been misclassified by XGB+LCMM as negative (i.e., the false negatives). In Table 4, patients' features are ranked based on their relevance according to SHAP (feature importance) and AraucanaXAI (distance from the root of the first appearance of the attribute in the tree splits). For each feature, we use SHAP to describe the impact of the feature, where a positive impact suggests that the variable pushes towards the positive class (MS worsening), while a negative impact suggests the opposite (no worsening). In the same way, we use AraucanaXAI to describe the most common decision rule associated with each feature. By combining SHAP and AraucanaXAI explanatory capabilities, we can identify a narrative describing what fools the model in predicting a lower worsening risk for positive patients. In particular, the XGB+LCMM model tends to be overconfident in associating longer latencies (diagnostic delay, time since onset) and younger ages at onset to a higher risk of worsening. Similar conclusions can be drawn for stable-high EDSS trajectories (i.e., Class 1) in LCMM combined with absence of supratentorial symptoms.

XAI methods can be employed in this sense also to evaluate qualitatively the differences between similar models, like XGB and XGB+LCMM. In Figure 3, we see the comparison of SHAP and AraucanaXAI explanations for the same false positive instance in both the models. In this case, adding the latent class trajectory of EDSS has a significant impact in AraucanaXAI on both the hierarchy of the rules and their split values, suggesting a drastic change in the predictive model decision boundaries that is not evident by inspecting only the SHAP values. On the other hand, it is worth highlighting how both the explanation strategies agree on giving LCMM a prominent role in the model's decision process (i.e. most important feature in the ranking is indeed the LCMM latent class).

5. Conclusion

Despite the lack of success in accomplishing the task of predicting the worsening of MS patients, testing the classifiers provided a meaningful opportunity to delve deeper into the underlying reasons for their underperformance through the implementation of explainable artificial intelligence (XAI) techniques, aimed at model inspection. By leveraging SHAP and AraucanaXAI, we were able to gain a more comprehensive understanding of the shortcomings and limitations of our classifiers through feature importance and navigable decision trees, shedding light on the factors that may have hindered the models' ability to effectively tackle the prediction task at hand. This insightful analysis therefore facilitated a deeper exploration of the intricacies and complexities involved in the classification problem, although we acknowledge that the lack of medical domain experts' involvement prevents us from drawing well-grounded conclusions about the clinical soundness of the explanations generated by the XAI methods. To this end, we emphasize the need for extensive evaluation studies of XAI in healthcare incorporating not only the model developers and clinical researchers' perspective, but also the perspectives of patients, caregivers, ethics and legal experts, and other relevant stakeholders involved in AI-supported medical decision-making.

6. Acknowledgements

T.M.B., G.N., and have not been funded to carry out the research described in this article. P.B, E.P., M.V., R.B, A.D. have received funding from the BrainTeaser project under H2020 GA 101017598 for their active role in the consortium.

7. References

[1] G. Faggioli et al., 'Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge', in *CLEF 2023 Working Notes*, M. Aliannejadi, G. Faggioli, N. Ferro, and M. Vlachos, Eds., CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073., 2023.

[2] G. Faggioli et al., 'Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023', in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, A. Aliannejadi, M. Vlachos, G. Faggioli, and N. Ferro, Eds., Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2023.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, 'A Survey of Methods for Explaining Black Box Models', *ACM Comput. Surv. CSUR*, vol. 51, no. 5, p. 93:1-93:42, Aug. 2018, doi: 10.1145/3236009.

[4] B. Goodman and S. Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"', *AI Mag.*, vol. 38, no. 3, Art. no. 3, Oct. 2017, doi: 10.1609/aimag.v38i3.2741.

[5] M. Kop, 'EU Artificial Intelligence Act: The European Approach to AI', Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3930959, Sep. 2021.

[6] S. M. Lundberg and S.-I. Lee, ‘A Unified Approach to Interpreting Model Predictions’, in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4765–4774.

[7] E. Parimbelli, T. M. Buonocore, G. Nicora, W. Michalowski, S. Wilk, and R. Bellazzi, ‘Why did AI get this one wrong? — Tree-based explanations of machine learning model predictions’, *Artif. Intell. Med.*, vol. 135, p. 102471, Jan. 2023, doi: 10.1016/j.artmed.2022.102471.

[8] T. M. Buonocore, N. Giovanna, and P. Enea, ‘Araucana XAI Software’. Sep. 2022. doi: 10.5281/zenodo.1234.

[9] S. N. Payrovnaziri *et al.*, ‘Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review’, *J. Am. Med. Inform. Assoc.*, vol. 27, no. 7, pp. 1173–1185, Jul. 2020, doi: 10.1093/jamia/ocaa053.

[10] T. M. Buonocore, G. Nicora, A. Dagliati, and E. Parimbelli, ‘Evaluation of XAI on ALS 6-months mortality prediction’, *Proc. Work. Notes CLEF 2022 - Conf. Labs Eval. Forum*, vol. 3180, pp. 1228–1235.

[11] H. Lennon *et al.*, ‘Framework to construct and interpret latent class trajectory modelling’, *BMJ Open*, vol. 8, no. 7, p. e020683, Jul. 2018, doi: 10.1136/bmjopen-2017-020683.

[12] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: ACM, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[13] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, ‘Interpretability in healthcare: A comparative study of local machine learning interpretability techniques’, *Comput. Intell.*, vol. 37, no. 4, pp. 1633–1650, 2021, doi: 10.1111/coin.12410.