Evaluation of XAI on ALS 6-months mortality prediction

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2022

Tommaso Mario Buonocore[a], Giovanna Nicora[a,b], Arianna Dagliati[a], Enea Parimbelli[a,c]

[a] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, Pavia, 27100, Italy*
[b] *enGenome Srl, Via Ferrata 5, Pavia, 27100, Italy*
[c] *Telfer school of Management, University of Ottawa, 55 Laurier Ave. E, Ottawa, ON K1N 6N5, Canada*

Abstract

In the article we present a comparative evaluation study of three different model agnostic XAI methods, namely SHAP, LIME and AraucanaXAI. The prediction task considered consists in predicting mortality for ALS patients based on observations carried out during a period of 6-months. 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 correctly predicted the outcome, vs when it predicted it incorrectly. The combination of the results of the qualitative and quantitative evaluations carried out in the experiment form the basis for a critical discussion of XAI methods properties and desiderata for healthcare applications, advocating for more inclusive and extensive XAI evaluation studies involving human experts.

Keywords

CLEF[1], ALS, neurological disease, disease progression, XAI, explainability, black-box, interpretable machine learning, predictive modeling, local explanation, surrogate model, evaluation

1. Introduction

1.1 ALS

Amyotrophic Lateral Sclerosis (ALS) is a chronic disease characterized by progressive impairment of neurological functions. It causes the paralysis of all voluntary muscles, usually leading to death or respirator-dependence within 3-5 years from onset. The incidence of ALS in Europe and in populations of European descent is 2.6 cases for 100 000 people per year and the prevalence is of 7 cases per 100000 people [1], [2]. ALS progression rate and pattern can be highly variable, progressively impairing the ability to move, communicate, swallow, and breathe. The life expectancy is shorter than 3 years for half of the patients, with only 10% surviving for more than 10 years. Currently, there are still no effective treatments that can halt or reverse ALS progression.

The heterogeneity in ALS clinical progression and survival makes predicting individual disease outcomes and clinical events very challenging [3], [4]. Nevertheless, predicting ALS progression can improve interventions, clinical practice and patient management. Few predictive models of the ALS progression have been assessed [5], [6], identifying predictive variables such as site and age at onset, diagnostic criteria, diagnostic delay, forced vital capacity and progression rate. Even though these models are able to predict individuals' survival or intervention endpoints, there is a lack of tools able to model the entire disease progression over time, considering heterogeneous and longitudinal variables and their relationships. Some preliminary work has been done developing a model based on Dynamic Bayesian Networks [7].

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: buonocore.tms@gmail.com (A. 1); enea.parimbelli@unipv.it (A. 4);
ORCID: 0000-0002-2887-088X (A. 1); 0000-0001-7007-0862 (A. 2); 0000-0002-5041-0409 (A. 3); 0000-0003-0679-828X (A. 4)


1.2 BrainTeaser

The BrainTeaser project aims at leveraging Artificial Intelligence for various aims, including better describing ALS mechanisms, predict disease progression in a probabilistic, time dependent fashion; and suggest interventions that can delay the progression of the disease. By further exploiting the potential of AI, models able to catch and predict disease progression and outcomes over time for stratified cohorts could be used both in the clinical setting to support patient care, or in clinical trials designed to generate in silico patients mimicking sub-populations of subjects with different characteristics. BrainTeaser aims include defining an interpretable end-to-end approach that interpolates temporal data and predicts the probability of an adverse event. Detecting complications during disease progression is of paramount importance for ALS patients and clinicians, and extending predictive models with mechanisms that elucidate which clinical features are important is key to predict and prevent such adverse outcomes.

1.3 XAI

The resurgence of the research trend known as eXplainable AI (XAI) has recently led to the development of methods that complement black-box Machine Learning (ML) models with local and global explanations [8]. While global explainability deals with a general understanding of the model classification behavior, local explainability approaches provide insight about the internals of the prediction for a single instance. Local XAI is extremely important in high stakes applications, such as health care, where patients and stakeholders may wish to understand why a particular prediction was made. Furthermore the "right for an explanation" is put forward by regulations, such as the European Union's GDPR2 and [9], the very recent EU Artificial Intelligence Act [10] as well as the U.S. government's Algorithmic Accountability Act of 2019, and the U.S. Department of Defense's Ethical Principles for Artificial Intelligence.

Among the several methodologies proposed for post-hoc local explainability, the most frequently used are SHAP and LIME. SHAP [11] is a game theoretic approach that delivers local explainability in form of "feature importance" in a single prediction. In particular, given a single instance with its attributes and the predicted probability by the model, SHAP computes Shapley values for each attribute, which represent the contribution of that attribute to the final predicted probability. Therefore, SHAP decomposes the final predicted probability of any machine learning model by assigning a contribution to each feature. On the other hand, LIME locally approximates the behavior of a complex black-box model to that of a simpler and explanable-by-design linear classifier [12]. In case of linear classifiers, such as logistic regression, the local explanation will be the beta regression coefficients of the linear model trained on a neighborhood of the instances whose prediction must be explained. Following the assumption of LIME, other LIME-based approaches have been proposed over the years [13], [14]. Both SHAP and LIME provide local explanations in terms of local "feature importance". This representation of an explanation may be obscure for a potential user with little background on machine learning analysis, such as clinicians. Motivated by this, we have previously developed an approach, named AraucanaXAI, whose surrogate explainable model consist in decision trees [14] (https://github.com/detsutut/AraucanaXAI). We believe that decision trees may be more suitable for delivering local explanations, since their structure is easily converted in a chain of if-then rules comprehensible to a wider range of users.

Despite the abundance of new XAI methods proposed in the literature, studies have rarely performed a comparison among different approaches, and even less have evaluated their performance using real-world clinical data [15]. For this reason, the objective of this research is to perform a comparison, and critical appraisal of SHAP, LIME and AraucanaXAI, on real-world data about ALS patients, coming from a cohort made available by the BrainTeaser project, on the task of predicting mortality based on a set of routinely collected clinical variables.

2. Materials and Methods

2.1 Prediction task

The iDPP CLEF 2022 offers two evaluation tasks focused on predicting the progression of Amyotrophic Lateral Sclerosis [16], [17]. The first task aims at ranking risk of impairment due to ALS, while in the second task participants are required to predict time to a certain impairment-relevant event (i.e., NIV, PEG or death). To evaluate the different XAI approaches on a medical problem, we chose to train different ML classifiers to predict 6-months mortality.

2.2 Datasets and preprocessing

Data from 1756 training patients and 494 test patients were available to develop and evaluate the models. We dropped the "alive" feature from the dataset since it is highly correlated with the outcome (discordance in only 2% of the cases)[2]. The covariates provided in the datasets are split into static (e.g. routinely collected demographic and clinical information, as well as specific ALS-related data) and time-dependent variables (i.e. spirometry and ALSFRS-R [18] questionnaire results). Since the observation time for time-dependent variables is restricted to 6-months, the number of time-points only amounted to a single spirometry test, and minimum 1 to maximum 3 questionnaire administrations. For this reason, we used the spirometry result as-is, while we adopted two possible simple approaches for the ALSFRS-R data consisting in a) using just the most recent observation, or b) calculating a slope (i.e. $slope=score(time_{latest}) - score(time_0)$) normalized by the time difference Δt between $time_0$ and $time_{latest}$ (i.e., $slope/\Delta t$). Our experiments showed that the alternative strategies a) or b) did not impact the predictive performance of the algorithms, and the role of time-dependent variables is marginal in the specific prediction task chosen for our study. Also, from the ALSFRS-R results we only kept the original q_n score (i.e., 1-4 answer to the single item of the questionnaire) while discarding the cumulative (section and total) scores to avoid co-linearity and high-correlation among features. No further preprocessing of input variables was performed, apart from dropping features with more than 90% of missing values. Additional detailed information about the dataset and the codebook explaining each feature can be found at *http://brainteaser.dei.unipd.it/challenges/idpp2022/*. Table 1 summaries the number of instances in training and test sets, stratified by outcome, to highlight the imbalance of the dataset towards the occurrence of death.

Table 1. Training and test datasets class distribution

	Occurrence of death	**Non-occurrence of death**
Training Set	1486 (85%)	270 (15%)
Test set	417 (84%)	77 (16%)

2.3 ML and XAI methods

We trained a set of 4 well-known classifiers to predict death occurrence: Gradient Boosting (using XGB implementation), Random Forest, Logistic Regression and Multilayer perceptron. All the models were implemented through the scikit-learn python package and code is available at *https://github.com/detsutut/AraucanaXAI/blob/master/araucana_CLEF_pipeline.ipynb*.
For the XAI methods we have focused our attention on three different methods for post-hoc, model-agnostic, local explainability, selecting SHAP, LIME and AraucanaXAI (ARAU in the following) as they are the ones that provide open-source, Python implementations that are readily available for use by researchers, available through the pip package manager.

[2] Notably, we spotted this significant flaw in the preprocessing when generating explanations using AraucanaXAI, where the first binary split of the explainer tree was consistently alive:0 vs alive:1. That allowed us to re-examine the feature selection performed during preprocessing and to fix the overlooked flaw. This is an example of how explanation facilities can also be employed for model debugging.

2.2 Evaluation metrics

Figure 1 details the experimental workflow followed in this study. After training the four predictive models, they are used to generate a prediction for each instance in the test set, and explanations with SHAP, LIME and AraucanaXAI are obtained. Then, XAI approaches are evaluated and compared in terms of a set of metrics defined in previous research on XAI in healthcare [19]:

- *Identity*: if there are two identical instances, they must have the same explanations
- *Fidelity*: concordance of the predictions between the XAI surrogate model and the original ML model
- *Separability*: if there are 2 dissimilar instances, they must have dissimilar explanations
- *Time*: average time required by the XAI method to output an explanation across the entire test set

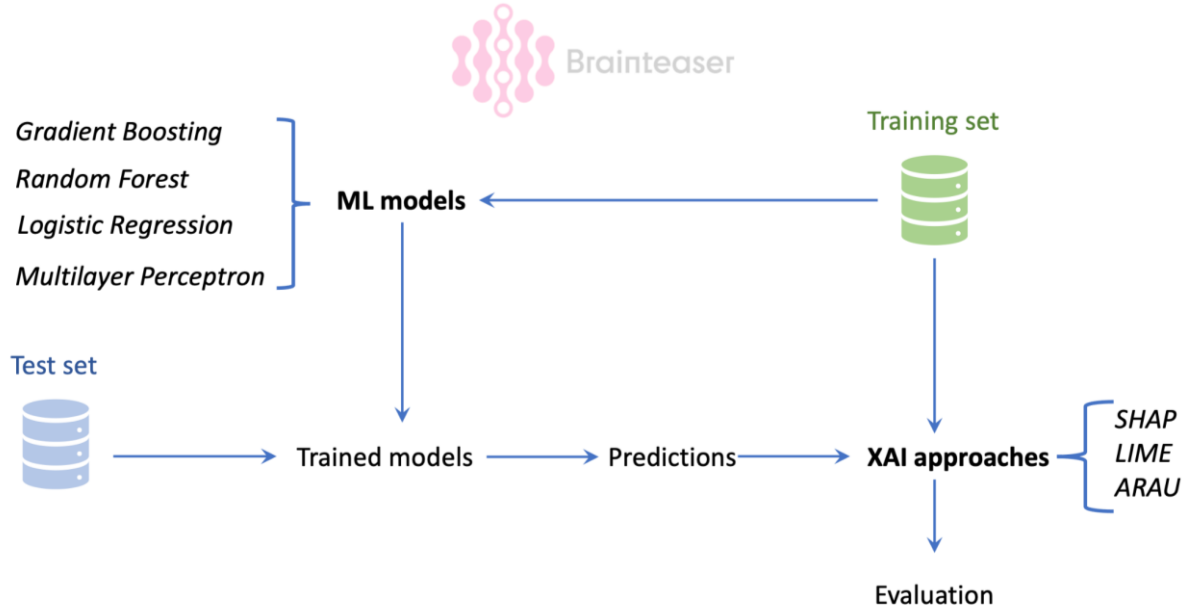

Figure 1. XAI experimental workflow: Four different ML models are trained and for each test instance the prediction is explained with three different XAI approaches (SHAP, LIME and AraucanaXAI). The training set is also employed by the XAI methods to produce the explanations.

3. Results

3.1 Predictive performance

Since the problem is imbalanced (see Table 1), we evaluated the performance of the different ML models in terms of balanced accuracy, which is the mean between sensitivity (i.e., the ability to identify positive cases, in this case the outcome "death") and specificity (i.e., the ability to correctly predict negative cases). All the models have low balanced accuracy, between 51% (Random Forest) and 55% (MLP). It is worth to note that we did not fine-tune the parameters (i.e., defaults hyperparameters with no further optimization) or the classification thresholds to try to achieve better classification performance since our aim is not to build a top performing predictive model but rather to evaluate how XAI can support the understanding of the model's behavior. To this end, the instances where the predictive model predicted the "wrong" class as most probable are indeed the most interesting to analyze (see discussion section).

3.2 XAI explanations performance

Table 2 reports the results of the evaluation of the three different XAI methods when applied to each of the predictive model trained at model development stage. Identity could not be calculated since there are no two identical instances in the test set. In line with previously reported results on an experiment on a different dataset [14], results on fidelity show a perfect score for SHAP and ARAU on a slight edge (it was more noticeable in [14]) compared to LIME. Separability is also consistently low across

all XAI methods, with LIME excelling with a perfect score, while ARAU is the worst-performing in terms of separability, albeit with still a very low number of identical explanations originating from different examples. Finally, SHAP is by far the fastest algorithm, with ARAU an order of magnitude slower, and LIME being the slowest (~>100x slower than SHAP).

Table 2. Evaluation of different XAI methods on a test dataset of 494 instances.

	Identity*	Fidelity (higher is better)			Separability (lower is better)			Time (s) (lower is better)		
		SHAP	LIME	ARAU	SHAP	LIME	ARAU	SHAP	LIME	ARAU
GB	N.A. *no 2 identical instances are present in the provided test set*	1	0.99	1	0.0005	0	0.03	0.01	484	13.5
RF		1	0.99	1	0	0	8.2E-6	7.7	539	48.4
LR		1	0.99	1	0	0	0.08	4.6	480	9.6
MLP		1	0.99	1	0	0	0.001	4.8	484	12.3

4. Discussion

The quantitative evaluation of the three different XAI methods did not reveal definitive superior performance of one of the approaches, albeit SHAP seems to be the better overall performing algorithm. However, the explainability evaluation metrics are not all that is needed to thoroughly assess the multifaceted construct of what constitutes a "good" explanation in XAI in healthcare. None of these measures indeed considers the fact that explanations, and especially local explanations, are ultimately directed, and should be optimized for, a human expert interacting with the AI system. A human-in-the-loop approach has been advocated as essential for the proper evaluation of XAI for healthcare [20] and concepts like causability (i.e., the integration of "usability" factors and causal reasoning to support explainability) are increasingly recognized as a central component of the quality of generated XAI explanations [21].

A more qualitatively-oriented analysis of the generated explanations is thus required, moving beyond the standard ML evaluation frameworks consisting of overall performance measures like AUC, sensitivity and specificity, brier score, AUPRC, to dig into the detailed behavior of different XAI methods when dealing with specific, and more relevant, situations: e.g. arguably an explanation would be much more important when the AI-based predictive model and the clinical expert do not agree on the prediction for a specific patient, and the clinician is called to question his expert opinion vs the AI, or rather decide that the AI algorithm simply predicted the wrong outcome. In the supplementary material[3] figures 1 to 4, we report an example of such comparative evaluation of SHAP, LIME and ARAU on four different cases of single-instance local explanations where the model succeeded (suppl. fig. 1 (pred:0, true class:0), suppl. fig. 2 (pred:1, true class:1)) or otherwise failed (suppl. fig. 3 (pred:0, true class: 1), suppl. fig. 4 (pred:1, true class: 0)) to predict the correct mortality outcome for an instance drawn from the test dataset.

Interestingly enough, analyzing such examples, we can highlight some more differences among the XAI methods that the evaluation metrics presented in table 2 did not highlight. These include the fact that a good agreement on the "most important features" that led to the ML model to the correct prediction (suppl fig 1 and 2) exists between SHAP and ARAU, albeit the two methods rely on fundamentally different underlying methodologies (game theory and feature contribution for SHAP vs recursive

[3]https://docs.google.com/document/d/e/2PACX-1vQEL-YeGtsjRxmvgYRUKsdwpumHUpv6sO5yNDWQMMHWPJnKX__7s57zBbjAoGEUwA/pub

partitioning of the data space for ARAU), while LIME which is theoretically more similar to ARAU (logistic linear surrogate model vs tree-based surrogate model) presents an explanation where the most important features are ethnicity and onset_bulbar, while the most important features identified by both SHAP and ARAU like onset_date, age_onset and the score of the 7th item of the ALSFRS-R questionnaire appear only later in the list. For what concerns incorrect predictions (suppl. fig. 3 and 4) it is also interesting to note how ARAU seems to have higher stability of the list of the most important features driving the classification (with onset_date, age_onset and q7 consistently appearing among the first binary splits of the explanation tree) when compared to SHAP's explanation which suggests a decreased importance of onset_date and age_onset in favor of an increased contribution to the incorrect classification of different questionnaire items (e.g. q3, q5, q6 and q9, see suppl fig 4) and demographic variables like sex, previously deemed unimportant in the cases of correct prediction. LIME also exhibits a similar stability of importance of features in explanations to ARAU, with ethnicity, onset_bulbar and q7 still scoring high importance in the incorrect prediction scenarios.

Another difference that emerges between ARAU, SHAP and LIME is that, exploiting the passthrough parameters related to the control of pruning of the surrogate explainer tree model (see ARAU's implementation documentation and usage examples, including available hyperparameters, here: https://github.com/detsutut/AraucanaXAI#readme) it is possible to control the complexity of the generated explanation in terms of number of nodes in the explainer tree. Methods that essentially rely on presentation of feature importance like SHAP or LIME do not offer such fine-grained control to the final user, possibly harming the overall effectiveness and usability of the explanation especially in scenarios with a significant number of covariates in the predictive model contributing to the final prediction (i.e. also explainable-by-design models like a logistic regression or a decision tree can quickly become hard to grasp if the dimensionality of the feature space is significantly large. E.g. compare the complexity of an explanation constituted by a single binary-split rule, or a regression with 3 covariates, with one with hundreds of splits or hundreds of odds-ratios or beta coefficients).

Our study of evaluation of XAI methods in the task of ALS mortality prediction has some limitations that we acknowledge: First of all, we did not optimize the ML models developed to achieve the best possible performance, which in turn affects the proportion of correct and incorrect predictions made by the models and ultimately the XAI explanations generated. Secondly, we overlooked more refined temporal data mining approaches to deal with the time-dependent covariates, which might have been proven useful to predictive performance, especially in a disease like ALS which progresses over time, if tackled with a more refined approach than the simple ones we attempted (see the overarching goals of the BrainTeaser project at: https://brainteaser.health). Finally, we did not include any clinical expert in our evaluation of XAI approaches, which significantly limits the qualitative evaluation, usability, and clinical soundness of the explanations generated by the XAI methods analyzed.

To this end, we advocate that further attention should be drawn by extensive evaluation studies[4] of XAI in healthcare involving clinical end users as well as a range of stakeholders that reaches beyond AI developers and clinical researchers, to include patients, their informal caregivers, ethics and legal experts and other relevant players that may be involved in AI-supported clinical decision-making.

5. Acknowledgements

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

[4] E.g., we are planning one of such studies, embracing open-science and participatory approaches, which you are welcome to express your interest to join here: https://osf.io/a8ezn/

6. References

[1] M. A. van Es *et al.*, 'Amyotrophic lateral sclerosis', *Lancet*, vol. 390, no. 10107, pp. 2084–2098, Nov. 2017, doi: 10.1016/S0140-6736(17)31287-4.

[2] M. H. B. Huisman *et al.*, 'Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology', *J Neurol Neurosurg Psychiatry*, vol. 82, no. 10, pp. 1165–1170, Oct. 2011, doi: 10.1136/jnnp.2011.244939.

[3] E. Beghi *et al.*, 'The epidemiology and treatment of ALS: focus on the heterogeneity of the disease and critical appraisal of therapeutic trials', *Amyotroph Lateral Scler*, vol. 12, no. 1, pp. 1–10, Jan. 2011, doi: 10.3109/17482968.2010.502940.

[4] S. Byrne *et al.*, 'Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study', *Lancet Neurol*, vol. 11, no. 3, pp. 232–240, Mar. 2012, doi: 10.1016/S1474-4422(12)70014-5.

[5] R. Kueffner *et al.*, 'Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach', *Sci Rep*, vol. 9, no. 1, Art. no. 1, Jan. 2019, doi: 10.1038/s41598-018-36873-4.

[6] H.-J. Westeneng *et al.*, 'Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model', *Lancet Neurol*, vol. 17, no. 5, pp. 423–433, May 2018, doi: 10.1016/S1474-4422(18)30089-9.

[7] A. Zandonà, R. Vasta, A. Chiò, and B. Di Camillo, 'A Dynamic Bayesian Network model for the simulation of Amyotrophic Lateral Sclerosis progression', *BMC Bioinformatics*, vol. 20, no. 4, p. 118, Apr. 2019, doi: 10.1186/s12859-019-2692-x.

[8] 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.*, vol. 51, no. 5, p. 93:1-93:42, Aug. 2018, doi: 10.1145/3236009.

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

[10] M. Kop, 'EU Artificial Intelligence Act: The European Approach to AI', Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3930959, Sep. 2021. Accessed: Dec. 19, 2021. [Online]. Available: https://papers.ssrn.com/abstract=3930959

[11] 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. Accessed: Jun. 29, 2019. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', *arXiv:1602.04938 [cs, stat]*, Feb. 2016, Accessed: Jul. 09, 2018. [Online]. Available: http://arxiv.org/abs/1602.04938

[13] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, and P. Bruza, 'LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models', *Decis. Support Syst.*, vol. 150, no. C, Nov. 2021, doi: 10.1016/j.dss.2021.113561.

[14] E. Parimbelli, G. Nicora, S. Wilk, W. Michalowski, and R. Bellazzi, 'Tree-based local explanations of machine learning model predictions, AraucanaXAI', *arXiv preprint arXiv:2110.08272*, 2021.

[15] S. N. Payrovnaziri *et al.*, 'Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review', *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1173–1185, Jul. 2020, doi: 10.1093/jamia/ocaa053.

[16] A. Guazzo *et al.*, 'Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022', in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*, 2022.

[17] A. Guazzo *et al.*, 'Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge', in *CLEF 2022 Working Notes*, 2022.

[18] J. M. Cedarbaum *et al.*, 'The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function', *Journal of the Neurological Sciences*, vol. 169, no. 1, pp. 13–21, Oct. 1999, doi: 10.1016/S0022-510X(99)00210-5.

[19] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, ‘Interpretability in healthcare: A comparative study of local machine learning interpretability techniques’, *Computational Intelligence*, vol. 37, no. 4, pp. 1633–1650, 2021, doi: 10.1111/coin.12410.
[20] A. Holzinger, ‘Interactive machine learning for health informatics: when do we need the human-in-the-loop?’, *Brain Inf.*, vol. 3, no. 2, pp. 119–131, Jun. 2016, doi: 10.1007/s40708-016-0042-6.
[21] A. Holzinger and H. Müller, ‘Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI’, *Computer*, vol. 54, no. 10, pp. 78–86, Oct. 2021, doi: 10.1109/MC.2021.3092610.