Identification of Multiple Sclerosis Signals' Dependence on Patients' Medical Conditions Through Stochastic Perturbation of Features in Five Machine Learning Models

Spyros Lavdas[1,2(✉)], Dimitrios Sklavounos[2], Panagiotis Gkonis[3], Panagiotis Siaperas[4], and Nikolaos Bakas[2,5]

[1] Department of Computer Science, Neapolis University Pafos, 8042 Paphos, Cyprus
s.lavdas@nup.ac.cy
[2] Department of Computer Science, Athens Metropolitan College, Marousi, Greece
{dsklavounos,nbakas}@mitropolitiko.edu.gr
[3] Department of Digital Industry Technologies, National and Kapodistrian University of Athens, Dirfies Messapies, Greece
pgkonis@uoa.gr
[4] Institute of Occupational Science and Rehabilitation Metropolitan College, Athens Metropolitan College, Marousi, Greece
psiaperas@mitropolitiko.edu.gr
[5] National Infrastructures for Research and Technology – GRNET, Athens, Greece

Abstract. Multiple sclerosis (MS) is a disease that deteriorates the central human nervous system, which can potentially cause significant brain, spinal cord and visual problems. Based on recent studies, MS has affected 3 million people with a prevalence rate of 3.9%. To this end, a wealth of information about MS has been produced, which makes MS the ideal candidate for applying artificial intelligence (AI) techniques for early diagnosis through a machine learning (ML) exploration framework. Accordingly, the current work studies to what extent the nervous system has been degenerated by analyzing data from a recently published dataset. Such data has been derived by motor evoked potential (MEP) measurements conducted in each patient hospital visit. Therefore, five machine learning models have been trained with cross-validation, in order to obtain the best one with good generalization properties. We compare the accuracy of all models utilizing various metrics (maximum obtained accuracy is ∼96% with XGBoost model). Furthermore, we use sensitivity analysis in order to explain the dependence of the target variable on the input parameters statistically.

Keywords: multiple sclerosis · EDSS · artificial intelligence · gradient boosting · polynomial regression · hyperparameter tuning · random forests

M. Papadaki et al. (Eds.): EMCIS 2022, LNBIP 464, pp. 47–59, 2023.
https://doi.org/10.1007/978-3-031-30694-5_5

1 Introduction

Multiple sclerosis (MS) is a well-known autoimmune chronic disease which causes visuals, sensor and motor problems having as a direct consequence the deterioration of the functional status of the central nervous system (CNS). Besides, it is difficult enough to timely detect MS disease as there are no certain symptoms and physical findings that dictate its diagnosis. To this end, a multitude of medical approaches have been adopted in order to accurately assess and diagnose MS. To date, the most applicable medical treatment is magnetic resonance imaging (MRI) which is a non-invasive imaging technology that creates anatomical images.

1.1 Related Work

Most of the MRI research works for MS diagnosis are based on the implementation of machine learning (ML) or deep learning (DL) techniques in MRI scans in order to extract critical conclusions from brain images, [25,26]. Although such type of processes is accurate and robust, it has turned out to be time-consuming and susceptible to manual errors, [21]. Instead, ML techniques have been rapidly evolved as the most promising player in the arena of MS decision support systems during the last decade. Such type of techniques does not require any prior knowledge or experience related to MS from clinicians facilitating the most accurate and objective diagnosis.

In particular, the most widely ML and DL-employed techniques have incorporated multiple data sources as input parameters such as clinical data, MRI scans, optical coherence tomography (OCT) data and motor evoked potential (MEP) measurements, [2,18]. Some representative works will be presented in order to clarify the importance of ML models for MS decision support. To start with, [9]. In this paper the authors proposed a machine learning pipeline for clinical questionnaires analysis which aimed at detecting MS disease course. In particular, patient-reported outcomes (PRO) questionnaires were used in order to capture the self-perception of the MS disease. Besides, in a recent work [16], serum and CSF levels of forty-five cytokines were analyzed to identify MS diagnostic markers. Thus, cytokines were analyzed using multiplex immunoassay. Analysis of variance-based parameters and Pearson correlation coefficient scores were employed in order to utilize the appropriate input parameters for classification purposes. In the same context, [1], text mining methods were introduced in transcriptomic data analysis of multiple sclerosis disease for the first time. A complete predictive model was developed by taking into consideration consecutive transcriptomic data preprocessing procedures. Besides, the KmerFIDF method was utilized as a feature extraction method and linear discriminant analysis for dimensionality reduction. Additionally, in [17], a support vector machine (SVM) method with tenfold validation performed on specific properties of patients' blood, such as zinc, adiponectin, total radical-trapping antioxidant parameter and, sulfhydryl, in order to predict MS with high sensitivity, specificity, and accuracy.

There are also published works which are not strongly dependent on medical data, but their analysis has been built from raw data such as gait disturbances [17] or exhaled breath analysis [8]. In both of the former works, four classification algorithms were employed in total: i) Logistic Regression (LR), ii) XGBoost (XGB) iii) SVM and iv) artificial neural network (ANN) model in order to analyze the imported raw data for MS prediction and classification purposes. Noteworthy, classifications and predictions have been enhanced by including the parameter of expanded disability status scale (EDSS), [14], which is a method of quantifying disability in multiple sclerosis and tacking down the evolution of the disability. Namely, it holds values from 0 (healthy person) to 10 (death). In this context, all of the following indicative published works, [12,13,27] have adopted the EDSS parameter as the target of their classification ML techniques. To this end, a multitude of ML techniques has been utilized, such as Bayesian, random decision trees as well as simple logistic-linear regression.

Apart from the EDSS parameter, there is a specific additional type of data that improved the accuracy and sensitivity of ML models. Such type of data is derived from MEP measurements, namely conducted measurements which quantify the conductivity of the CNS. In [5,23], MEP measurements were carried out and were further analyzed by random forests and linear regression classifiers. To this end, the current study is based on MEP measurements and utilizes the EDSS as a target parameter for the employed ML algorithms. The required dataset regarding the MEP measurements has been derived from a recently published paper, [24].

It is worth noting that the evaluation of a patient's disability is a multi-parameter medical process which is prone to EDSS miscalculation due to manual errors, and it is time-consuming as well. Thus, the estimation of EDSS through an automated analysis of a nervous system signal pulse, as suggested by the current work, could accelerate all the procedures in terms of MS prognosis and medical treatment.

The rest of the current research work is organized as follows: In Sect. 2, the structure of the used dataset and the metadata regarding patients is presented. The key features of the five employed machine learning models are described in the next Sect. 3 while the analysis of the derived results is developed in Sect. 4. Finally, the main conclusions and the proposal for future work are given in the last section.

2 Description of Dataset

The dataset derived from the work of Yperman et al., [24], contains data regarding electrical signals which have propagated through the NS and detected from either hand or foot. In particular, the brain of each patient has been stimulated through a magnetic stimulator and an external trigger system leading to the creation of a signal pulse which propagates along the nervous system. Samples of the resulting signal are detected, are stored and exported into a file. To be more specific, 2000 time points in a time window of 100 ms determine the signal shape. In total, the dataset includes information about:

– metadata of patients (963 records) such as age, gender, time of hospital visit, type of machine that conducted the measurements, teams that carried out the MEP measurements, etc. For more information, see [24].
– MEP measurements (96290 records).
– EDSS values (7414 records) for specific patients.

As it has already been mentioned, the target of the ML techniques is the EDSS value for each patient at the specific time of MEP measurement. To achieve this, critical properties of the resulting signals should be taken into consideration in order to be used as input parameters to the machine learning models. The following table describes all the input parameters used for the development and analysis of the suggested prediction model.

Table 1. Description of the utilized input parameters.

Input parameters	**Description**
Age	Age of the patient
AnatomyAH	AnatomyAH = 1 corresponds to the MEP hand measurement, while AnatomyAH=0 corresponds to the MEP feet measurement
R	R=1 corresponds to right(hand or foot), while R=0 corresponds to left (hand or foot)
maximum peak	Maximum peak voltage of signal (mV)
FilterTrue	FilterTrue=1 corresponds that the machine has applied frequency filter to the raw MEP measurements, while FilterTrue=0 means that there was not any filter applied to the MEP measurement
Time minimum	The timepoint when the minimum peak occurs
half power max(min)	The width of the pulse when the maximum(minimum) peak occurs
Time maximum	The timepoint when the maximum peak occcured
time of first local-0.25(0.5,0.75) min	The timepoint when a local peak obtains voltage $\geq$ 0.25(0.5,0.75) of the minimum peak
mean	Voltage average of the pulse
minimum peak	Minimum peak voltage of signal (mV)
delta time of first local-0.25(0.5,0.75) min	Time difference between time of first local-0.25(0.5,0.75) and the time minimum
Male	Male=1 corresponds to male, while Male=0 corresponds to female
delta time max min	Time difference between the times that maximum and minimum occur
time of first local-0.25(0.5,0.75) max	The timepoint when a local peak obtains voltage $\leq$ 0.25(0.5,0.75) of the maximum peak
Energy	The energy of the signal pulse calculated by $\int_0^{100\ ms} \|U(t)\|^2 dt$
std deviation	Standard deviation
TeamA	TeamA=1 means that the group A conducted the MEP measurement, while TeamA=0 corresponds to team B
MachineA	MachineA=1 corresponds to the machine A that used to carry out the MEP measurement, while MachineA=0 corresponds to machine B
N min(max) local 0.25(0.5,0.75)	Number of minimum(maximum) locals when the peak voltage corresponds to voltages of $\geq(\leq)$0.25(0.5,0.75) of minimum(maximum) peaks

3 Machine Learning Models

The following models for approximating Features-Target relationship have been employed.

1. Linear Regression as a baseline model for the next ones.
2. Polynomial Regression. As it is necessary to select from a vast pool of potential nonlinear features along with their number, the ITSO [4] as well as PROS [19] Optimization Algorithms for Feature Selection have been adopted, which have also been found experimentally vastly efficient.
3. Gradient Boosting [7,10,22] with hyperparameter tuning. Particularly, the grid search method with cross-validation has been utilized.
4. Random Forests [6], as implemented in [20]
5. Artificial Neural Networks [3].

For each model, the following computations are carried out:

- The accuracy among the Prediction and Target-variable for the Train and Test Sets.
- Error analysis: Residual Errors vs Target diagrams, Probability Density Functions, Cumulative Density Functions, for the Train and Test Sets as well.

The former approach is beneficial for detecting specific patterns occurring in the prediction and, hence, enhancing the generalization capability and reliability of the model.

4 Results

The application of the aforementioned ML techniques produced a vast amount of constructive results which shed light on the MS disease. It should be highlighted, that input parameter with high discrepancies or abrupt deviations of the signal voltage magnitude in the time domain has been removed from the calculations. In the same, context, all the spikes occurred at the beginning of the detected pulse have been deleted. Figure 1 shows the pairwise correlations among the input parameters. Particularly, we reform the correlation matrix, such that the non-zero values are concentrated around the diagonal of the matrix. This is performed with the CuthillMcKee Method [11,15]. This way, the clusters of associated features are computationally identified.

4.1 Performance of ML Models

After an exhaustive search for the hyperparameters of the studied ML models, we identified that RF and XGB exhibited the best performance (Figs. 2, 3). The error metrics are depicted in Fig. 4. Although the final models have adequate accuracy, some errors occur, especially for lower values of EDSS. Hence we run a data adequacy check (Fig. 5), where it is demonstrated that the accuracy is increased with the number of training samples.

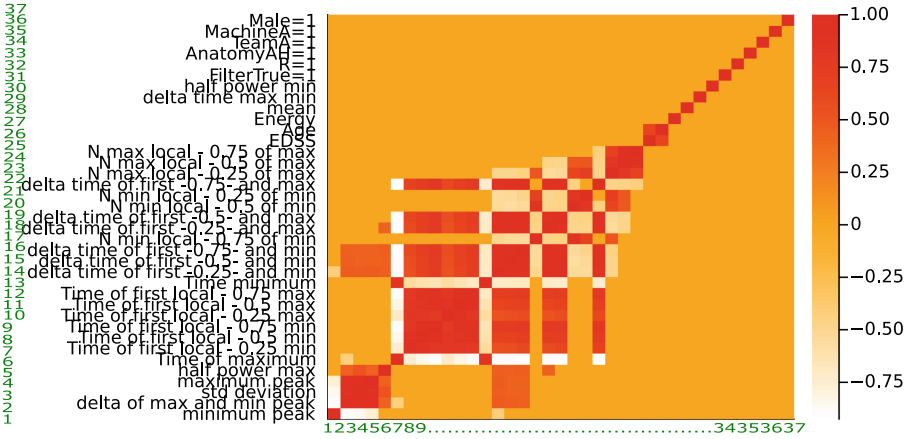

Fig. 1. CuthillMcKee representation of the input parameters, see Table 1. Note that the x-axis label is exactly the same as that of y-axis. For the sake of readability, the x-axis labels are described with numbers which correspond to specific input parameters, as shown on the left side of the graph.

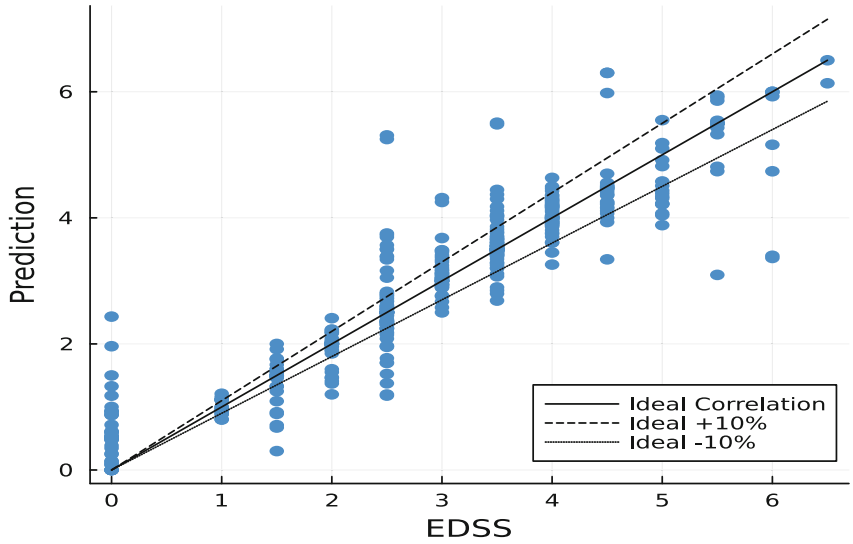

Fig. 2. Target vs predicted EDSS derived from linear model.

4.2 Sensitivity Analysis

Finally, we run a sensitivity analysis of the input parameters with respect to the target, by keeping all features constant to a particular value (median, 25% and 75% quantiles), and change the studied feature within its given values in the initial dataset. Accordingly, we predict using each one of the models. Figures 6

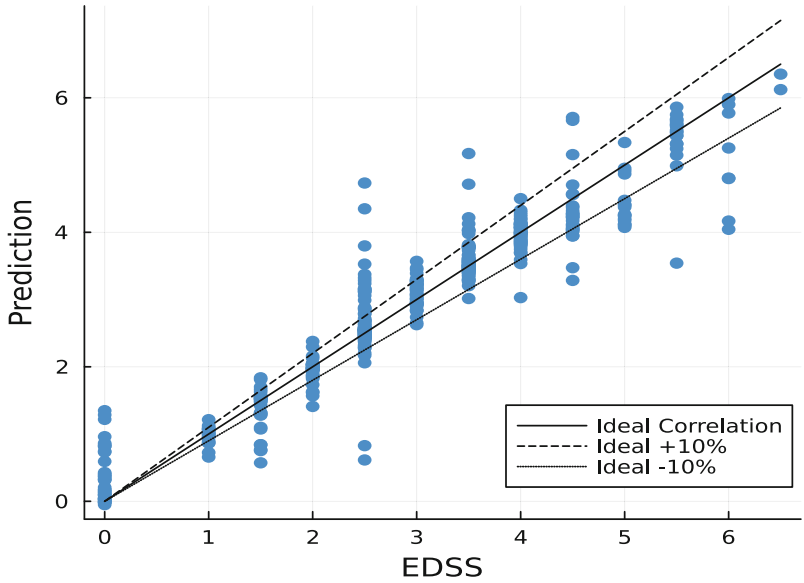

Fig. 3. Target vs predicted EDSS derived from XGBoost.

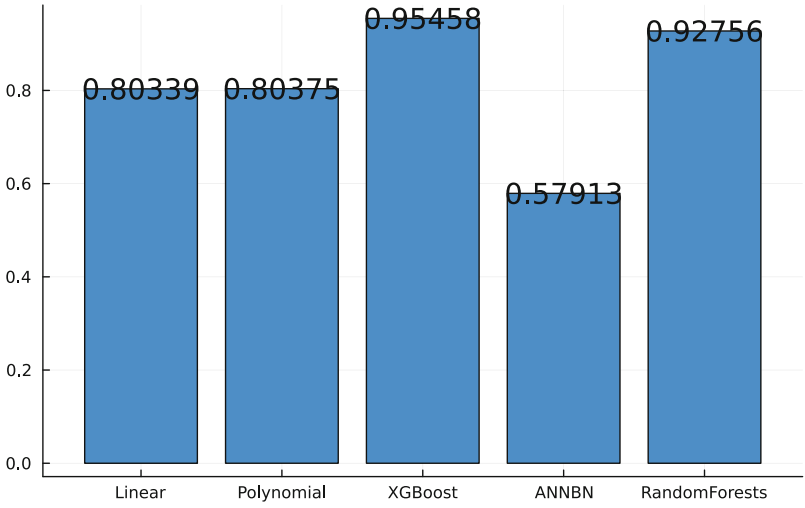

Fig. 4. Error metrics for 5 machine learning methods implemented in the current analysis.

and 7 illustrate that the Anatomy AH, when it is equal to the unit, the EDSS results in lower values. Furthermore, Figs. 8 and 10 depict an increasing pattern of EDSS with respect to age.

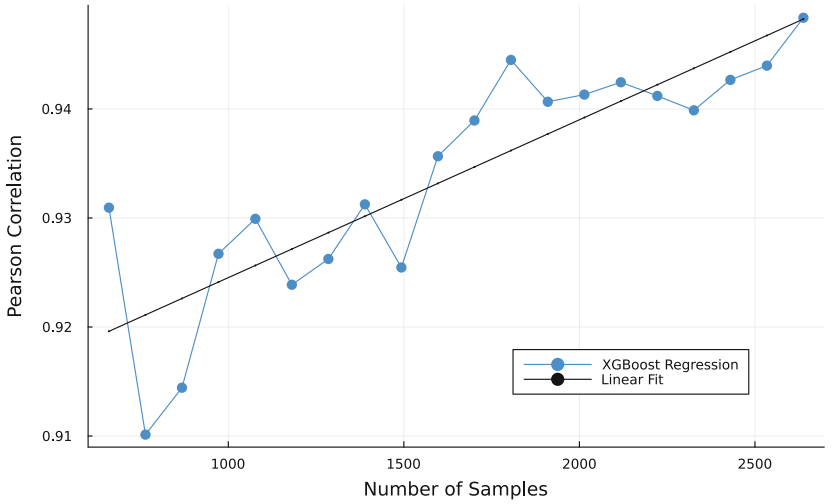

Fig. 5. Pearson correlation along the number of samples for the XGBoost model.

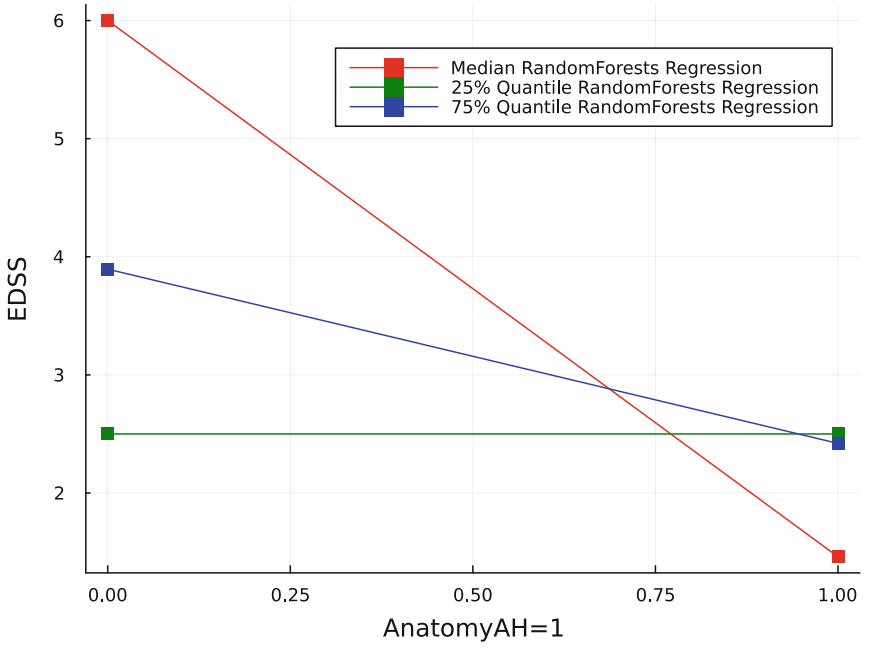

Fig. 6. Sensitivity analysis for AnatomyAH=1 based on the RandomForests model.

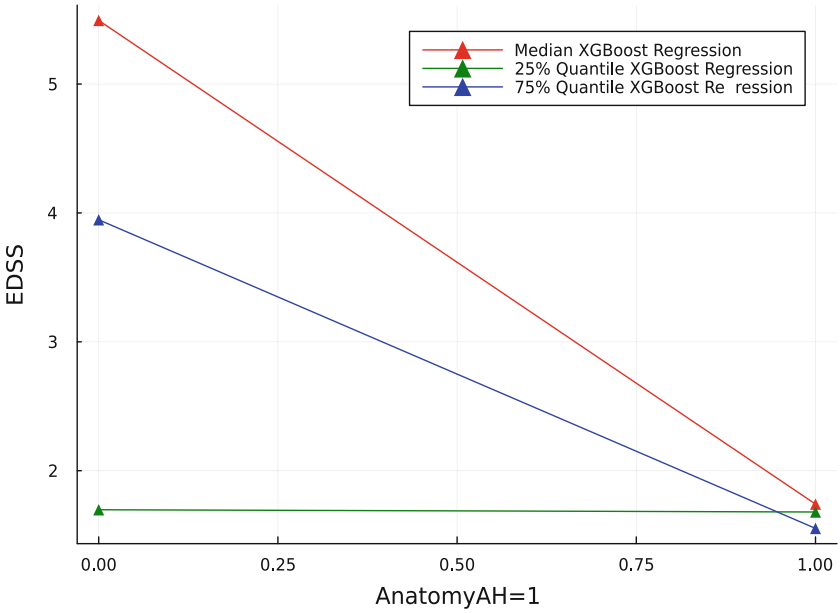

Fig. 7. Sensitivity analysis for AnatomyAH=1 based on the XGBoost model.

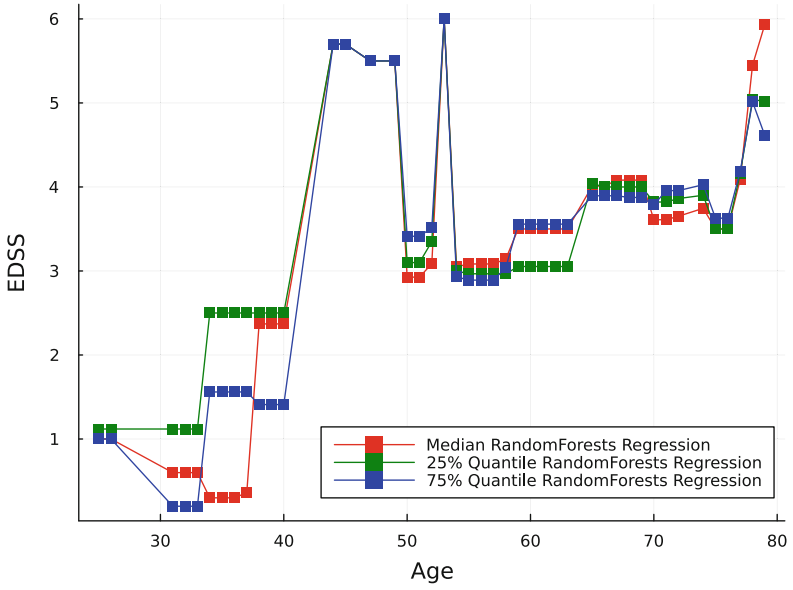

Fig. 8. Presentation of increasing pattern between the input parameter of age and EDSS emerged from the RandomForests model.

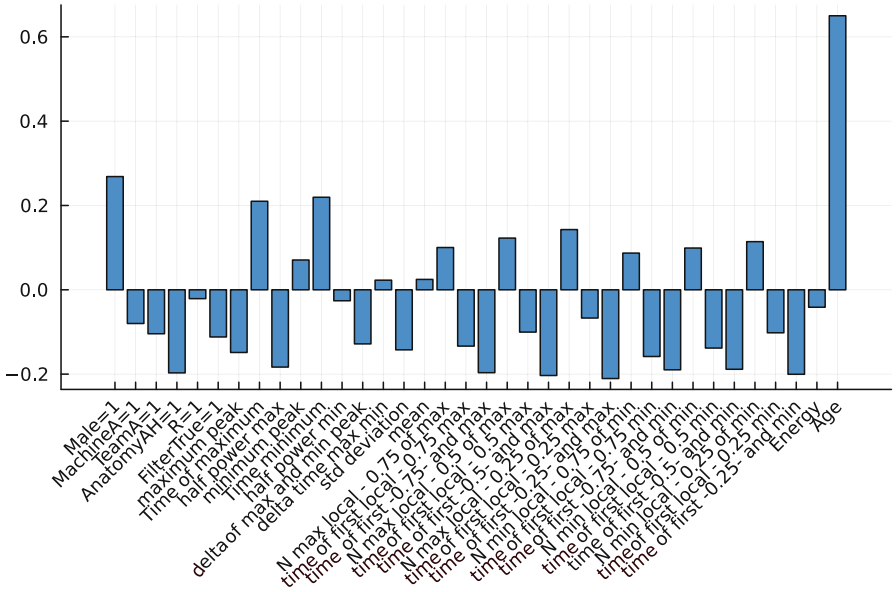

Fig. 9. Features importance derived by XGBoost model, which was proved to provide the most accurate prediction results.

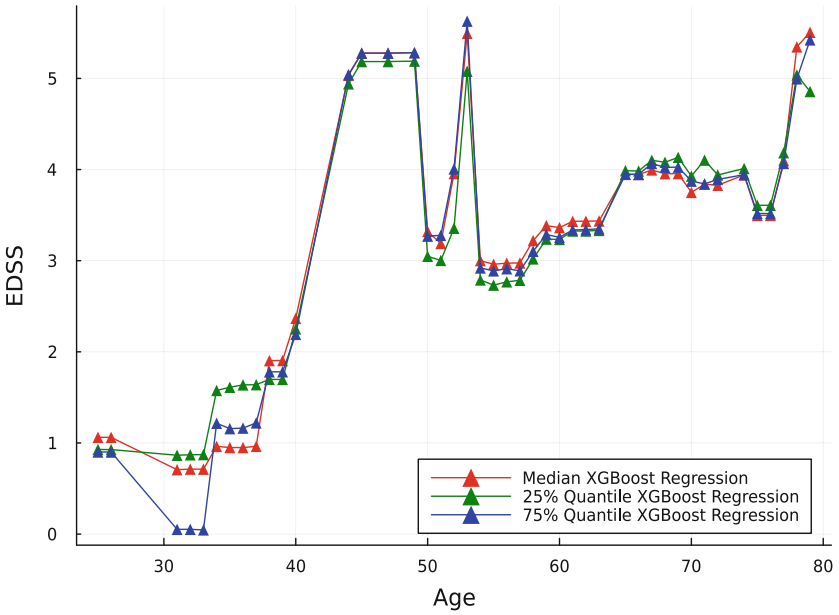

Fig. 10. Resentation of increasing pattern between the input parameter of age and EDSS emerged from the XGBoost model.

5 Conclusions

The present research analysis in concert with former descriptive Figs. 1-3 manifest that the highest Pearson correlation was achieved through the implementation of XGBoost, reaching approximately 96% accuracy. On the contrary, the lowest accuracy of 58% was obtained by the artificial neural network. Thus, the most appropriate ML model to predict the value of EDSS from a propagating signal on the nervous system is XGBoost.

One of the most critical conclusions drawn in the current MS analysis is depicted in Fig. 9 wherein the most effective input parameters of the prediction model are presented. In particular, the age, gender, anatomyAH and maximum peak are the prominent ones. Among the former significant input parameters are also the energy, time difference of local peaks as well as the time points at either local maximums or minimums occur.

Besides, the Pearson correlation follows an increasing linear trend along with the increase of input training data, Fig. 5. Hence, the suggested prediction model can be clarified as efficient and robust.

The early results of the current research work dictate its extension to the prediction of the forthcoming evolution of EDSS for a certain MS patient. This is of great importance as it will determine the most appropriate long-term medical treatment to enhance the quality of patient's life.

References

1. Ali, N.M., Shaheen, M., Mabrouk, M.S., Aborizka, M.A.: A novel approach of transcriptomic microrna analysis using text mining methods: an early detection of multiple sclerosis disease. IEEE Access **9**, 120024–120033 (2021). https://doi.org/10.1109/access.2021.3109069
2. Aslam, N., et al.: Multiple sclerosis diagnosis using machine learning and deep learning: Challenges and opportunities. Sensors **22**(20) (2022). https://doi.org/10.3390/s22207856, https://www.mdpi.com/1424-8220/22/20/7856
3. Bakas, N.P., Langousis, A., Nicolaou, M., Chatzichristofis, S.A.: A gradient free neural network framework based on universal approximation theorem. arXiv preprint arXiv:1909.13563 (2019)
4. Bakas, N.P., Plevris, V., Langousis, A., Chatzichristofis, S.A.: ITSO: a novel inverse transform sampling-based optimization algorithm for stochastic search. Stoch. Env. Res. Risk Assess. **36**(1), 67–76 (2021). https://doi.org/10.1007/s00477-021-02025-w
5. Bejarano, B., et al.: Computational classifiers for predicting the short-term course of multiple sclerosis. BMC Neurol. **11**(1) (2011). https://doi.org/10.1186/1471-2377-11-67
6. Breiman, L.: Random Forest. Machine Learning **45**(1), 5–32 (2001)
7. Chen, T., Guestrin, C.: Xgboost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). https://doi.org/10.1145/2939672.2939785
8. Ettema, A.R., Lenders, M.W., Vliegen, J., Slettenaar, A., Tjepkema-Cloostermans, M.C., de Vos, C.C.: Detecting multiple sclerosis via breath analysis using an eNose,

a pilot study. J. Breath Res. **15**(2), 027101 (2021). https://doi.org/10.1088/1752-7163/abd080
9. Fiorini, S., Verri, A., Tacchino, A., Ponzio, M., Brichetto, G., Barla, A.: 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), pp. 4443–4446 (2015). https://doi.org/10.1109/EMBC.2015.7319381
10. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
11. Gates, R.L.: Cuthillmckee.jl (2022). https://github.com/rleegates/CuthillMcKee.jl
12. Ion-Mărgineanu, A., et al.: Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. Front. Neurosci. **11**, 398 (2017). https://doi.org/10.3389/fnins.2017.00398
13. Kocevar, G., et al.: Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses. Front. Neurosci. **10**, 478 (2016). https://doi.org/10.3389/fnins.2016.00478
14. Kurtzke, J.F.: Rating neurologic impairment in multiple sclerosis. Neurology **33**(11), 1444–1444 (1983). https://doi.org/10.1212/WNL.33.11.1444, https://n.neurology.org/content/33/11/1444
15. Liu, W.H., Sherman, A.H.: Comparative analysis of the Cuthill-Mckee and the reverse Cuthill-Mckee ordering algorithms for sparse matrices. SIAM J. Numer. Anal. **13**(2), 198–213 (1976)
16. Martynova, E., et al.: Serum and cerebrospinal fluid cytokine biomarkers for diagnosis of multiple sclerosis. Mediat. Inflamm. **2020**, 1–10 (2020). https://doi.org/10.1155/2020/2727042
17. Mezzaroba, L., et al.: Antioxidant and anti-inflammatory diagnostic biomarkers in multiple sclerosis: a machine learning study. Mol. Neurobiol. **57**(5), 2167–2178 (2020). https://doi.org/10.1007/s12035-019-01856-7
18. Nabizadeh, F., et al.: Artificial intelligence in the diagnosis of multiple sclerosis: a systematic review. Multiple Sclerosis Related Disord. **59**, 103673 (2022). https://doi.org/10.1016/j.msard.2022.103673, https://www.sciencedirect.com/science/article/pii/S2211034822001882
19. Plevris, V., Bakas, N.P., Solorzano, G.: Pure random orthogonal search (pros): a plain and elegant parameterless algorithm for global optimization. Appl. Sci. **11**(11), 5053 (2021). https://doi.org/10.3390/app11115053
20. Sadeghi, B.: Decisiontree.jl (2013)
21. Stafford, I.S., Kellermann, M., Mossotto, E., Beattie, R.M., MacArthur, B.D., Ennis, S.: A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ Digit. Med. **3**(1) (2020). https://doi.org/10.1038/s41746-020-0229-3
22. Xu, B., Chen, T.: Xgboost.jl (2014)
23. Yperman, J., et al.: Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis. BMC Neurol. **20**(1), 1–15 (2020). https://doi.org/10.1186/s12883-020-01672-w
24. Yperman, J., Popescu, V., Wijmeersch, B.V., Becker, T., Peeters, L.: Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset. Sci. Data **9**(1), 207 (2022). https://doi.org/10.1038/s41597-022-01335-0

25. Zeng, C., Gu, L., Liu, Z., Zhao, S.: Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. Front. Neuroinformatics **14**, 610967 (2020). https://doi.org/10.3389/fninf.2020.610967
26. Zhang, Y., et al.: Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. Simulation **92**(9), 861–871 (2016). https://doi.org/10.1177/0037549716666962
27. Zhao, Y., et al.: Exploration of machine learning techniques in predicting multiple sclerosis disease course. PLOS ONE **12**(4), e0174866 (2017). https://doi.org/10.1371/journal.pone.0174866