A Comparative Study of Explainable AI models in the Assessment of Multiple Sclerosis

Andria Nicolaou[1](✉), Nicoletta Prentzas[1], Christos P. Loizou[2], Marios Pantzaris[3], Antonis Kakas[1], and Constantinos S. Pattichis[1]

[1] Department of Computer Science, University of Cyprus, Nicosia, Cyprus
{nicolaou.andria,prentzas.nicoletta,antonis,pattichi}@ucy.ac.cy
[2] Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, Limassol, Cyprus
christos.loizou@cut.ac.cy
[3] Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus
pantzari@cing.ac.cy

Abstract. Multiple Sclerosis (MS) is characterized by complex and heterogeneous nature and as a result, there's currently no cure. Medications can help control the progression and ease the symptoms of MS. The scientific interest in the field of explainable artificial intelligence (AI) comes to the surface and aims to assist computer-aided diagnostic systems to be established in medical use by providing understandable and transparent information to the experts. The objective of this study was to present different learning methods of explainable AI models in the assessment of MS disease based on clinical data and brain magnetic resonance imaging (MRI) lesion texture features and compare them by focusing on the main findings. The learning methods used machine learning and argumentation theory to differentiate subjects with relapsing-remitting MS (RRMS) from progressive MS (PMS) subjects and provide explanations. The results showed that the different learning methods achieved a high accuracy of 99% and gave similar explanations as they extracted the same set of rules. It is hoped that the proposed methodology could lead to personalized treatment in the management of MS disease.

Keywords: Multiple Sclerosis · Brain MRI · Lesions · Texture Features · Clinical Data · Machine Learning · Rule Extraction · Argumentation · Explainable AI

1 Introduction

Multiple Sclerosis (MS) is a complex autoimmune disease affecting the central nervous system and is the leading cause of non-traumatic neurological disability in young adults [1]. Both environmental and genetic factors are believed to contribute to MS susceptibility. Environmental influences such as smoking, childhood obesity, infectious mononucleosis, and low serum vitamin D are consistently associated with increased MS risk [1, 2]. While it may be possible to assess an individual's MS susceptibility based on genetic data and risk factor exposure, the practicality of routinely predicting MS development faces theoretical and practical challenges [1].

N. Tsapatsoulis et al. (Eds.): CAIP 2023, LNCS 14185, pp. 140–148, 2023.
https://doi.org/10.1007/978-3-031-44240-7_14

MS is traditionally seen as having two distinct stages. In the initial stage, there is inflammation that leads to relapsing-remitting disease (RRMS). In the later stage, there is neurodegeneration that results in a progressive form of the disease (PMS). This progression can be defined as either secondary progressive MS or primary progressive MS, depending on the symptoms and the disability [2]. The hallmark of MS is the appearance of white matter (WM) lesions that can be seen using Magnetic Resonance Imaging (MRI) to diagnose the progression of the disease [2].

Recent advancements in artificial intelligence (AI) have led to its widespread use, demonstrating exceptional performance in numerous tasks through complex machine learning (ML) systems. However, the increased complexity has made these systems function like "black boxes," raising concerns about their operation and decision-making processes [3]. This lack of transparency has hindered their adoption in healthcare. Consequently, explainable AI has gained significant attention, focusing on developing methods that can explain and interpret ML models [3].

The objective of this study was to present the learning method of two different explainable AI models focused on the assessment of MS disease progression and compare them by discussing their findings. Both learning methods are based on ML and argumentation theory.

2 Materials

A dataset of 87 MS subjects (34 males, and 53 females) was examined at different time points. MRI images of 66 RRMS and 21 PMS were obtained using different MRI scanners and different sequences (T1w, T2w, and FLAIR). The expert neurologist (co-author, M. Pantzaris) manually segmented the brain MS lesions in a blinded manner where the segmented areas were intensity normalized between the grayscale values of 0 and 255. Clinical data were also investigated including demographic, and neurological measurements, such as functional system (FS) scores defining 0: 'Normal', 1: 'Signs Only', 2: 'Mild', 3: 'Moderate', 4: 'Severe', and 5: 'Loss' [4].

Texture features were extracted from all the segmented MS lesions and were estimated by averaging the corresponding values for all lesions of each patient. The following selected group features were extracted [5]: first-order statistics (FOS), spatial grey level dependence matrix (SGLDM), neighborhood grey tone difference matrix (NGTDM), and Fourier power spectrum (FPS). Min-max normalization was performed between the values 0.0 and 1.0, where a fixed number of 3 bins that has the same number of observations to each bin (quantile strategy) was defined. The bins were encoded using the ordinal method, where 0 refers to 'Low', 1 refers to 'Medium' and 2 refers to 'High'. In addition, feature selection was applied by computing the analysis of variance (ANOVA) test. The 5 features with the highest F-value, from both clinical data and texture features, were selected (see Table 1).

Data were collected from 87 subjects coming from two groups: 66 RRMS (G1) and 21 PMS (G2) (see Table 2). As shown in Table 2, data were oversampled using the synthetic minority over-sampling technique (SMOTE) which creates new samples for the minority group of the model (G2) with the same statistical properties [6]. Splitting using 80% for the training and 20% for the evaluation set and the target class as a stratified parameter was applied.

Table 1. Selected clinical data and brain MRI lesion texture features.

Clinical data
cerebellarFS, slowtongueFS, facialFS, sensoryFS, dysarthriaFS
Brain MRI lesion texture features
contrastNGTDM, varianceFOS, variancesumsquaresSGLDM, sumvarianceSGLDM, angularsumFPS

FS: Functional Systems, NGTDM: Neighbourhood Grey Tone Difference Matrix, FOS: First-Order Statistics, SGLDM: Spatial Grey Level Dependence Matrix, FPS: Fourier Power Spectrum.

Table 2. Data distribution of the models.

Data sets	Subjects	RRMS (G1)	PMS (G2)
Initial	87	66	21
Over-sample minority	132	66	66
Training	106	53	53
Evaluation	26	13	13

RRMS: Relapsing-Remitting MS, PMS: Progressive MS.

3 Methods

3.1 Learning Method A

The first learning method utilized ML and argumentation theory. The ML algorithms random forest (RF), and gradient boosting (GB) were used. During model training, the grid search method was applied to find the optimal combination of hyper-parameters of each model [7], based on a stratified 10-fold cross-validation. Rules were extracted on training using the TE2rules algorithm [8] that converts a tree ensemble (TE) to a rule list (RL). Then, rule selection was performed selecting the models with high training accuracy and a minimum sample of rules. Argumentation-based reasoning was applied using Gorgias' theory [9], which involves constructing arguments using a basic argument scheme, connecting a set of premises to the claim of the argument. The extracted rules were modified as object-level arguments that can support contradictory claims, leading to arguments attacking one another. Moreover, the use of priority on object-level arguments can express a local preference between arguments and establish relative strength, tightening the attack relation between them. The performance of the learning method was based on the average evaluation set performance for 10 runs.

3.2 Learning Method B: ArgEML

The second learning method called ArgEML [10] is an argumentation framework for explainable machine learning, based on a novel approach that integrates sub-symbolic methods with logical methods of argumentation to provide explainable solutions to learning problems [11]. In the framework of ArgEML argumentation is used both as a target language for ML and the explanations of the ML predictions. The learning algorithm generates argumentation theories in the context of Gorgias argumentation framework [9], by processing a set of data and optionally a list of decision rules that represent some knowledge of the data (hybrid mode of operation). The ArgEML approach views the notion of prediction from a different perspective than that of a traditional ML model, by means of relaxing the requirement of accuracy by distinguishing two notions of *definite prediction* (single conclusion that can be either correct or wrong) and *ambiguity* (multiple conflicting conclusions) recognition. In this perspective, if we cannot uniquely or definitely predict, but can focus the prediction on a set of alternatives and can give justifications for the alternatives, then we consider that we still have a valuable output of learning. For these difficult cases, an argumentation theory will generate a *dilemma*. Dilemmas include multiple conflicting conclusions with explanations for each particular conclusion. A dilemma can be considered neither a correct nor a wrong prediction. For that reason, dilemmas are included in a new *learning assessment* (5) metric for evaluating the performance of a theory. More information on the framework and methodology can be found in [12]. The ArgEML system: α-version[1] is a Java implementation of the methodology that we can use to learn and evaluate Gorgias' argumentation theories.

3.3 Evaluation Metrics

The performance of the two learning methods was based on the evaluation set (see Table 2). The following evaluation metrics were used:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}\ (1),\ Precision = \frac{TP}{TP + FP}\ (2),$$

$$Recall = \frac{TP}{TP + FN}\ (3),\ F1\ score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}\ (4)$$

where TP and TN denote the number of true positive and true negative instances that are correctly identified, and FP and FN indicate the number of false positive and false negative instances that are incorrectly classified, respectively.

The Learning Assessment (LA) metric, introduced in [12] for the evaluation of the argumentation theories generated by ArgEML, is a generalization of the standard classification accuracy metric, that gives a holistic evaluation of an argumentation theory that balances definite errors and dilemmas:

$$\begin{aligned} &Learning\ assessment(LA) \\ &= \frac{definite\ correct\ predictions(TP + TN) + dilemmas * wa}{total\ number\ of\ predictions(TP + FP + TN + FN)} \end{aligned} \tag{5}$$

where *wa* corresponds to a weight factor for ambiguity, defined as 1/(number of labels in target class).

[1] https://github.com/nicolepr/argeml

4 Results

4.1 Learning Method A

Tables 3 and 4 illustrate the RL generated from the selected RF and GB models, respectively. It is shown that two rules consisting of only one feature can describe the target group G1 (see Table 3). In addition, some features are strong enough to differentiate the subjects into two different groups (G1 vs G2) as both RF and GB models extracted the same feature rules (e.g., *cerebellarFS*, *sensoryFS*). It's worth mentioning that the contrast from the NGTDM group is the only texture feature observed in the RL.

4.2 Learning Method B: ArgEML

In this work, we used the ArgEML system in hybrid mode to learn an argumentation theory from the dataset described in Sect. 2. Following the process of "Learning method A", we utilized 10 subsets of train/test sets, to train a RF model and extract decision rules using the inTrees algorithm [13]. We used the rules extracted from the best-performing models on the train and test sets with an Accuracy of 100% to run the ArgEML system, one time for each set of train/test/rules, and decide/learn the best-performing argumentation theory. ArgEML in hybrid mode processes the decision-rules given as input and generates an initial theory that contains a compact set of arguments (~rules) that cover the training data. Table 5 illustrates the compact set of rules extracted from a selected best-performing RF model and chosen by ArgEML for initialing the argumentation theory.

4.3 Evaluation of the Learning Methods

According to Tables 3, 4, and 5, it is observed that the different learning methods gave similar rules. These rules consisted of the clinical data and more specifically, the cerebellar and the sensory function systems' measures. Furthermore, it is highlighted that the brain MRI lesion texture features can be found in the rules with a length greater than 2 (see Tables 3 and 4). It's worth mentioning that the ArgEML learning method gave rules with a length equal to 1, meaning that the theory can differentiate the subjects of MS (G1 vs G2) with only one feature (see Table 5).

Table 6 summarizes the performance of the learning methods using the evaluation set based on a stratified 10-fold cross-validation (see Table 2). It is obvious that the use of the argumentation theory in both learning methods reached a high accuracy of 99% which makes the explainable AI models predict and provide explanations with high fidelity.

Table 3. Rules extracted from a selected RF model in learning method A.

Rules	Group
IF *cerebellarFS* = Normal **OR** SignsOnly **OR** Mild	**G1**
IF *sensoryFS* = Normal **OR** SignsOnly **OR** Mild	**G1**
IF (*cerebellarFS* = Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Moderate **OR** Severe **OR** Loss)	**G2**
IF (*contrastNGTDM* = Medium **OR** High) **AND** (*dysarthriaFS* = SignsOnly **OR** Mild **OR** Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Normal **OR** SignsOnly **OR** Mild)	**G2**

FS: Functional Systems, NGTDM: Neighborhood Grey Tone Difference Matrix, G1, G2: Subjects with RRMS and PMS, respectively.

Table 4. Rules extracted from a selected GB model in learning method A.

Rules	Group
IF *sensoryFS* = Normal **OR** SignsOnly **OR** Mild	**G1**
IF *cerebellarFS* = Normal **OR** SignsOnly **OR** Mild	**G1**
IF (*cerebellarFS* = Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Moderate **OR** Severe **OR** Loss) **AND** *slowtongueFS* = SignsOnly	**G1**
IF (*cerebellarFS* = Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Moderate **OR** Severe **OR** Loss)	**G2**
IF (*contrastNGTDM* = Medium **OR** High) **AND** (*facialFS* = Mild **OR** Moderate **OR** Severe **OR** Loss) **AND** (*sensoryFS* = Normal **OR** SignsOnly **OR** Mild)	**G2**

FS: Functional Systems, NGTDM: Neighborhood Grey Tone Difference Matrix, G1, G2: Subjects with RRMS and PMS, respectively.

Table 5. Rules extracted from a selected RF model and ArgEML(theory initialization) in learning method B.

Rules	Group
IF *sensoryFS* = Normal **OR** SignsOnly **OR** Mild	**G1**
IF *cerebellarFS* = Normal **OR** SignsOnly **OR** Mild	**G1**
IF *slowtongueFS* = Normal	**G1**
IF *sensoryFS* = Moderate **OR** Severe **OR** Loss	**G2**
IF *slowtongueFS* = SignsOnly **OR** Mild **OR** Moderate **OR** Severe **OR** Loss	**G2**

FS: Functional Systems, G1, G2: Subjects with RRMS and PMS, respectively.

Table 6. Evaluation of the two learning methods.

Learning method	Accuracy	Precision	Recall	F1 score
A: RF + ARG	99%	99%	99%	99%
A: GB + ARG	99%	99%	100%	99%
B: ArgEML	99%	99%[a]	99%[a]	99%

RF: Random Forest, GB: Gradient Boosting, ARG: Argumentation theory.
[a]Dilemmas were considered both as FP and FN.

5 Discussion

The objective of this study was to compare two learning methods of explainable AI models in the assessment of MS disease based on clinical data and brain MRI lesion texture features. Both learning methods used ML and argumentation theory to differentiate subjects with RRMS from PMS subjects, providing explanations with a high accuracy of 99%. The main findings showed that:

1) Different learning methods can give the same explanation as long as they extracted the same rules.
2) Cerebellar and sensory function systems' rules were strong enough to identify and explain the type of MS disease. The contrast from the NGTDM group was the only brain MRI lesion texture feature found in the rules.

A previous study from our group [14] performed rule extraction from brain MRI lesion texture features using decision trees to assess MS disease progression. The main findings showed that simple rules including only one texture feature group (e.g. FPS) without the combination of other feature groups can achieve high accuracy greater than 70%. Another recent study from our group [15] implemented an explainable AI model with embedded rules in the assessment of brain MRI lesions in MS disease based on Amplitude Modulation – Frequency Modulation (AM-FM) multi-scale feature sets. Different ML models were used to classify the MS subjects with a low disability and subjects with a high disability. Argumentation-based reasoning was performed using the extracted rules from models with a high accuracy of 98%. It was demonstrated that the proposed model could differentiate the MS subjects by providing understandable information for the progression of the disease.

Other MS studies investigated explainability using local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP). More specifically, Basu *et al.* [16] developed multivariate ML models to predict MS disease activity using extreme GB and applied SHAP methods to identify the predictive covariates for early identification of MS. A large-scale study was used including demographic, neurological, and laboratory measures, as well as MRI assessment. The models achieved a balanced accuracy of 80%. The findings showed that the number of treatment weeks, the new combined unique active lesion count, the new T1 hypointense lesion count, and the age-related MS severity score were the top predictive covariates. In addition, Olatunji *et al.* [17] used different ML models and interpreted them utilizing SHAP and LIME methods for early screening of MS. The input data of the models included clinical features, such as demographic and other laboratory measures. The results indicated that Extra Trees outperformed the rest of the models with an accuracy of 95%. The greatest impact on the model's prediction was shown by age, systolic blood pressure, and alkaline phosphatase.

6 Concluding Remarks

In a medical diagnosis system, clarity and transparency are crucial factors for gaining the trust of medical experts. Since the underlying causes of MS are still not that clear, it is essential to develop an explainable AI model in the assessment of MS disease. This study presented two different learning methods of explainable AI models which used ML and argumentation theory to identify the progression of the disease and explain its causes. By comparing the two learning methods, it's concluded that they can give the same explanation as selected features are strong enough to assess the disability and differentiate the MS subjects. Further work needs to be carried out using more subjects.

References

1. Hone, L., Giovannoni, G., Dobson, R., Jacobs, B.M.: Predicting multiple sclerosis: challenges and opportunities. Front. Neurol. **12**, 1–8 (2022)
2. Dobson, R., Giovannoni, G.: Multiple sclerosis-a review. Eur. J. Neurol. **26**, 27–40 (2019)
3. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**, 1–45 (2021)
4. Kurtzke, J.F.: Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology **33**(11), 1444–1452 (1983)
5. Loizou, C.P., Petroudi, S., Seimenis, I., Pantziaris, M., Pattichis, C.S.: Quantitative texture analysis of brain white matter lesions derived from T2-weighted MR images in MS patients with clinically isolated syndrome. J. Neuroradiol. **42**(2), 99–114 (2015)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 108–122 (2013)
8. Lal, G.R., Chen, X., Mithal, V.: TE2Rules: extracting rule lists from tree ensembles, pp. 1–17 (2022)
9. Kakas, A.C., Moraitis, P., Spanoudakis, N.I.: GORGIAS: applying argumentation. Argument Comput. **10**, 55–81 (2019)
10. Prentzas, N., Gavrielidou, A., Neophytou, M., Kakas, A.: Argumentation-based Explainable Machine Learning (ArgEML): a real-life use case on gynecological cancer. In: CEUR Workshop Proceedings, vol. 3208 (2022)
11. Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A., Pattichis, C.: Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. In: Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019, pp. 817–821. Institute of Electrical and Electronics Engineers Inc. (2019)
12. Prentzas, N., Pattichis, C., Kakas, A.: Explainable machine learning via argumentation. In: Communications in Computer and Information Science. Springer (2023)
13. Deng, H.: Interpreting tree ensembles with inTrees. Int. J. Data Sci. Anal. **7**(4), 277–287 (2018). https://doi.org/10.1007/s41060-018-0144-8

14. Nicolaou, A., Loizou, C.P., Pantzaris, M., Kakas, A., Pattichis, C.S.: Rule extraction in the assessment of brain mri lesions in multiple sclerosis: preliminary findings. In: Tsapatsoulis, N., Panayides, A., Theocharides, T., Lanitis, A., Pattichis, C., Vento, M. (eds.) CAIP 2021. LNCS, vol. 13052, pp. 277–286. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89128-2_27
15. Nicolaou, A., et al.: An explainable artificial intelligence model in the assessment of brain MRI lesions in multiple sclerosis using amplitude modulation – frequency modulation multi-scale feature sets. In: 24th International Conference on Digital Signal Processing (DSP), pp. 1–4. Rhodes, Greece (2023)
16. Basu, S., Munafo, A., Ben-Amor, A.F., Roy, S., Girard, P., Terranova, N.: Predicting disease activity in patients with multiple sclerosis: an explainable machine-learning approach in the Mavenclad trials. CPT Pharm. Syst. Pharmacol. **11**, 843–853 (2022)
17. Olatunji, S.O., Alsheikh, N., Alnajrani, L., Alanazy, A., Almusairii, M., et al.: Comprehensible machine-learning-based models for the pre-emptive diagnosis of multiple sclerosis using clinical data: a retrospective study in the Eastern province of Saudi Arabia. Int. J. Environ. Res. Public Health **20** (2023)