An Explainable Artificial Intelligence model in the assessment of Brain MRI Lesions in Multiple Sclerosis using Amplitude Modulation - Frequency Modulation multi-scale feature sets

Andria Nicolaou, Antonis Kakas and Constantinos S. Pattichis
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
email: {nicolaou.andria, antonis, pattichi}@ucy.ac.cy

Marios S. Pattichis and Kevin Fotso
Department of Electrical Engineering, and Computer Engineering
University of New Mexico
USA
email: {pattichi, kfotsotagne}@unm.edu

Christos P. Loizou
Department of Electrical Engineering, and Computer Engineering and Informatics
Cyprus University of Technology
Limassol, Cyprus
email: christos.loizou@cut.ac.cy

Marios Pantzaris
Cyprus Institute of Neurology and Genetics
Nicosia, Cyprus
email: pantzari@cing.ac.cy

***Abstract*—The objective of this study was to implement an explainable artificial intelligence (AI) model with embedded rules to assess Multiple Sclerosis (MS) disease evolution based on brain Magnetic Resonance Imaging (MRI) multi-scale lesion evaluation. Amplitude Modulation-Frequency Modulation (AM-FM) features were extracted from manually segmented brain MS lesions obtained using MRI and were labeled with the Expanded Disability Status Scale (EDSS). Machine learning models were used to classify the MS subjects with a benign course of the disease and subjects with advanced accumulating disability. Rules were extracted from the selected model with high accuracy and then were modified to perform argumentation-based reasoning. It is demonstrated that the proposed explainable AI modeling can distinguish MS subjects and give meaningful information to track the progression of the disease. Future research will examine more subjects and add new feature sets and models.**

Keywords—Multiple Sclerosis; Brain MRI; Lesions; AM-FM features; Classification analysis; Rule extraction; Explainable AI.

I. Introduction

One of the greatest challenges for effective personalized treatments in Multiple Sclerosis (MS) is the difficulty to predict the disease evolution due to its heterogeneous nature. Explainable artificial intelligence (AI) can help in detecting and monitoring disease progression and providing transparent and understandable explanations both to physicians and patients. As MS is a complex autoimmune disease of the central nervous system, several lines of evidence indicate that both genetic and environmental variables may have a major impact in determining the vulnerability to the disease, even if the actual origin of MS is not entirely understood [1], [2]. MS is mainly characterized by the appearance of lesions in the white matter (WM) which show inflammatory activity, myelin damage, and axonal loss [1]. The MS lesions are visualized by magnetic resonance imaging (MRI) [3], and evaluated by expert neurologists following the McDonald criteria [4]. The clinical disability is assessed at each MRI scan using the expanded disability status scale (EDSS) [5].

Previous studies [6]-[8] showed that feature analysis of MS lesions and more specifically AM-FM features can be used to assess the evolution of MS disease. Furthermore, another study [9] took one step forward in the assessment of the disease with the extraction of quantitative data from the lesion features, in the form of rules, using machine learning models. The objective of this study was to investigate the usefulness of explainable AI in the assessment of MS disease based on brain MRI multi-scale lesion evaluation.

II. Methodology

The proposed implemented explainable AI model consisted of six main processing steps, including MRI acquisition, preprocessing, segmentation, feature extraction and selection, classification analysis, and argumentation-based reasoning. Below is a detailed analysis of each step.

A. MRI acquisition

A total of 38 subjects (17 males, and 21 females) with a clinically isolated syndrome (CIS) of MS were investigated. MRI scans were carried out at the initial stage of the disease ($Time_0$) and after 6-12 months ($Time_{6\text{-}12}$). The transverse MRI images used for analysis were obtained using a T_{2w} turbo spin-echo pulse sequence (repetition time=4408 ms, echo time=100 ms, echo spacing=10.8 ms). The reconstructed image had a slice thickness of 5 mm and a field of view of 230 mm with a pixel resolution of 2.226 pixels per mm. Standardized planning procedures were followed during each MRI examination. The MRI images were acquired using a 1.5 T whole-body Philips ACS NT MR imager. The EDSS score of each subject was

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

estimated by the neurologist (co-author, M. Pantzaris), at two, five, and ten years after the initial diagnosis to quantify future disability progression. In this paper, we use MRI images at $Time_0$ and their interrelation with the EDSS score at year ten.

B. Preprocessing

The brain MRI images were intensity normalized between the grayscale values of 0 and 255 using histogram normalization as documented in [6]-[8], where all additional details about the algorithm may be found. All detectable brain lesions were identified and segmented by the experienced MS neurologist.

C. Segmentation

The segmentation was performed manually in a blinded manner without the possibility of identifying the subject, the time-point of the exam, or the clinical findings. The selected points and delineations were saved to be used for feature extraction and analysis. In addition, the MS subjects were separated into two different groups (i.e. G_1: EDSS≤3.5 and G_2: EDSS>3.5). The reason for selecting an EDSS cut-off point of 3.5 is that for EDSS>3.5, the physician can assess neurological signs, meaning that the patient starts accumulating disability. Thus, any patient having an EDSS≤3.5 can be regarded as having a rather benign course of the disease.

D. Feature extraction and selection

Over each segmented MS lesion, a multiscale AM-FM decomposition was computed using the:

$$I(x,y) = \sum_{n=1}^{M} a_n(x,y)\cos\phi_n(x,y) \quad (1)$$

where M denotes different scales, $n = 1, 2, 3$ correspond to the low, medium, and high scales, $a_n(x,y)$ denote the instantaneous amplitude (IA) components, and $\varphi_n(x,y)$ denote the instantaneous phase components. It is noted that FM components $\cos\varphi_n(x,y)$ describe fast changing texture components. For each AM-FM component, the associated instantaneous frequency (IF): $\nabla\varphi_n(x,y)$ was estimated as described in [10].

The bandpass filters were grouped into low (LF), medium (MF), and high (HF) components. For sampling at 2.226 pixels/mm, multiplying the discrete spatial frequencies by $2.226/(2\pi)$ converts them (component wise) into cycles per millimeter. To see this, note that the π-frequency produces samples of $1, -1, 1, -1, \ldots$ at 0.5 cycles/pixel. For the low frequencies, we have discrete frequencies from the minimum IF magnitude of $(0, \pi/8)$ corresponding to 0.1391 cycles/mm, and to a maximum IF magnitude of $(\pi/4, \pi/4)$ at 0.3935 cycles/mm. For the medium frequencies, we have the minimum at $(0, \pi/4)$ corresponding to 0.2782 cycles/mm, and a maximum at $(\pi/2, \pi/2)$ corresponding to 0.7870 cycles/mm. For the high frequencies, we have the minimum frequency at $(0, \pi/2)$ corresponding to 0.5565 cycles/mm, and a maximum of (π, π) at 1.5740 cycles/mm.

For each lesion, for each frequency-scale band, the median value from 32-bin histograms of the dominant IA, IF magnitude (|IF|), and IF angle components were computed. The IF magnitude estimates were then normalized to cycles per millimeter, providing a physically meaningful interpretation of the texture measurements.

Before performing the classification analysis, the features were preprocessed using the scikit-learn [11], a machine learning library in Python. More specifically, min-max scaler was used to normalize only the IA features between the values 0.0 and 1.0. K-bins discretizer was also applied to discretize the values of all the features in intervals, called bins. 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, the select k-best method was used to select 5 of the features according to the highest score which was defined by computing the analysis of variance (ANOVA), F-value.

E. Classification analysis

Classification modeling was developed to predict EDSS of MS subjects with EDSS≤3.5 (G_1) versus those with EDSS>3.5 (G_2) based on the extracted lesion AM-FM features. The classification models were implemented in Python using the scikit-learn library [11]. Different classifiers were used, such as decision tree (DT), random forest (RF), gradient boosting (GB), k-nearest neighbors (kNN), gaussian naïve Bayes (Gaussian NB), and support vector machine (SVM). As shown in TABLE I, data were split into a training and an evaluation group, using 80% for the training and 20% for the evaluation set, and were re-arranged to have an equal size of the two classes on the evaluation set. The synthetic minority over-sampling technique (SMOTE) [12] was applied during the model training to improve the performance of the model and avoid overfitting. SMOTE creates new samples for the minority group of the model (G_2) with the same statistical properties. Applying the over-sampling technique on the training set, G_1 and G_2 were generated having the same number of subjects.

TABLE I. DATA DISTRIBUTION OF THE CLASSIFICATION MODELS

Data sets	Patients	EDSS≤3.5 (G_1)	EDSS>3.5 (G_2)
Initial	38	26	12
Training	30	21	9
Over-sample training	46	23	23
Evaluation	6	3	3

G_1, G_2: Subjects with 0.0≤EDSS≤3.5 and 3.5>EDSS≤10.0, respectively.

Furthermore, the grid search method was performed to find the optimal combination of hyper-parameters of each model [11], based on a stratified 10-fold cross-validation. Overfitting of data by cross-validation was avoided. The classification analysis performance in this study was based on the average evaluation set performance for 10 runs. The following evaluation metrics were used:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Sensitivity/Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

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

By selecting models with high accuracy, lesion features' rules were extracted during the model training using the TE_{2rules} algorithm [13], a novel approach to convert a tree ensemble (TE) for binary classification to a rule list (RL). This algorithm is characterized by high fidelity, as it generates rules from leaf nodes of individual trees and captures the interactions between trees of TE.

F. Argumentation-based reasoning

Gorgias is a structured argumentation framework and was used to couple learning with reasoning [14]. In detail, arguments are constructed using a basic argument scheme to link a set of premises with the claim of the argument. The premises are a set of conditions describing a scenario and the claims are options. Two types of arguments are constructed within a Gorgias argumentation theory: object-level arguments and priority arguments. Object-level arguments are literal claims and can support contradictory claims where arguments attack each other. Priority arguments express a local preference between arguments and their purpose is to give relative strength, tightening the attack relation between them.

Gorgias Cloud **Error! Reference source not found.** is an implemented platform that offers argumentation as a service and was used to visualize the rule tabulation as an internal explanation and the application-level explanation. Gorgias can represent the knowledge by describing the application in terms of object-level arguments. Explanation in a physical language is the final output of the model to provide understandable information both to the experts and to the patients.

III. RESULTS

A. Classification analysis

The proposed models were trained and evaluated for the lesion features of the data at the initial stage of the disease ($Time_0$). TABLE II tabulates the results of the evaluation metrics for model classification based on the evaluation set. It is shown that the AM-FM features can be used to differentiate subjects with a benign course of the disease (EDSS≤3.5) and subjects with advanced accumulating disability (EDSS>3.5), achieving an average accuracy (ACC) of 75%.

TABLE II. MS LESION MODEL EVALUATION RESULTS BETWEEN THE TWO DIFFERENT GROUPS (EDSS≤3.5 VS EDSS>3.5 AT YEAR 10) AT $TIME_0$ AVERAGED AT 10 RUNS

Classifiers	ACC	SEN	SPE	PR	REC
DT	0.72	0.82	0.71	0.60	0.82
RF	0.73	0.91	0.70	0.57	0.91
GB	0.75	0.83	0.77	0.63	0.83
kNN	0.73	0.93	0.69	0.53	0.93
Gaussian NB	0.73	0.81	0.75	0.67	0.81
SVM	0.75	0.90	0.73	0.60	0.90

ACC: Accuracy, SEN: Sensitivity, SPE: Specificity, PR: Precision, REC: Recall, DT: Decision Tree, RF: Random Forest, GB: Gradient Boosting, kNN: k-nearest neighbors, NB: Naïve Bayes, SVM: Support Vector Machine.

A RL was generated from a selected GB model, which achieved high accuracy at training (ACC=98%), using the TE_{2rules} algorithm. TABLE III shows an example of the rule extraction.

TABLE III. AN EXAMPLE OF GB RULE EXTRACTION USING TE2 RULES

Rules	Group
IF (*amplitudeHF* = Medium **OR** High) **AND** (*amplitudeLF* = Medium **OR** High) **AND** (*angleHF* = Low **OR** Medium)	G_1
IF (*amplitudeMF* = Low **OR** Medium) **AND** *angleHF* = High **AND** *magnitudeMF* = High	G_1
IF *amplitudeHF* = High **AND** (*angleHF* = Low **OR** Medium)	G_1
IF *amplitudeLF* = High **AND** *amplitudeMF* = High	G_1
IF (*amplitudeLF* = Medium **OR** High) **AND** *amplitudeMF* = Low **AND** (*angleHF* = Low **OR** Medium)	G_1
ELSE	G_2

G_1, G_2: Subjects with 0.0≤EDSS≤3.5 and 3.5>EDSS≤10.0, respectively, HF: High Frequency, LF: Low Frequency, MF: Medium Frequency.

B. Argumentation-based reasoning

Applying Gorgias' argumentation theory, object-level and priority arguments were constructed. Object-level arguments were the selected rules extracted from the classification models which were modified to have the syntax of the logic programming language, Prolog. Priority arguments were determined by prioritizing the object-level arguments. Gorgias Cloud was used to visualize the explanations of the predicted disability in MS. An example of a scenario is illustrated in TABLE IV, providing the input and output of Gorgias Cloud as well as the explanation in a physical language.

TABLE IV. AN EXAMPLE OF A SCENARIO USING GORGIAS CLOUD

input	*amplitudeHF*(p20, Medium). *amplitudeLF*(p20, High). *amplitudeMF*(p20, High). *angleHF*(p20, High). *magnitudeMF*(p20, Low).
output	prove([lowEDSS(p20)], InternalExplanation). Solution 1 Internal Explanation: [c4(p20),pr4(p20),r3(p20)], Application Level Explanation The statement "lowEDSS(p20)" is supported by: - "*amplitudeLF*(p20,High)" and "*amplitudeMF*(p20,High)" This reason is : - Stronger than the general reason of supporting "highEDSS(p20)"
	The patient p20 is predicted with low disability as the *instantaneous amplitude of low frequency* is High and the *instantaneous amplitude of medium frequency* is High, and this reason is stronger than the general reason supporting the opposite prediction of high disability.

HF: High Frequency, LF: Low Frequency, MF: Medium Frequency.

IV. DISCUSSION

The objective of this study was to investigate the usefulness of explainable AI in the assessment of MS disease based on brain MRI multi-scale lesion evaluation. The main findings showed that the implemented explainable AI model can differentiate MS subjects with an EDSS≤3.5 from those with an EDSS>3.5 at year ten of the disease (ACC=75%), and provide explanations with high fidelity based on the TE_{2rules} algorithm to follow up the disease evolution based on AM-FM features extracted from lesions at the baseline ($Time_0$).

Previous studies from our group investigated the analysis of AM-FM features focused on MS disease using brain MRI images. More specifically, Loizou *et al.* [7] suggested that AM–FM characteristics succeeded in differentiating between lesions and WM tissue (normal and normal-appearing). SVM classifier was used to differentiate subjects with an EDSS≤2 from those with an EDSS>2, combining different scales of frequency and achieving a correct classification score (CC) of 86%. In another recent study, Loizou *et al.* [8] proposed a methodology for the early detection of AM-FM features that can be used to predict the severity of MS disease. The classification was performed to differentiate subjects with an EDSS≤3.5 from those with an EDSS>3.5 at year ten of the disease, including both texture and AM-FM lesion features and achieving a CC=94%. In addition, a preliminary work from our group studied the rule extraction in the assessment of MS disease focusing on MS lesion texture features [9]. There were a few other studies reported in the literature focused on MS disease which explained the decision of the implemented convolutional neural network (CNN) using attribution methods in heatmaps [15], [17] and Shapley additive explanations (SHAP) plots to show the feature importance of a machine learning model [18]. However, there is no other study reported in the literature that has implemented an explainable AI model which provides explanations in the form of rules based on AM-FM feature analysis of the MS lesions.

V. CONCLUDING REMARKS

Personalized treatment of MS disease is a challenge in the medical domain as the cause of MS remains opaque. The proposed explainable AI model aims to assist physicians in the MS assessment and follow-up of disease evolution. Future work will include further feature sets and incorporate models based on sequential MRI scans at different time points of image acquisition. The proposed methodology should also be evaluated on more subjects in a future study.

REFERENCES

[1] R. Dobson and G. Giovannoni, "Multiple sclerosis-a review," *Eur. J. Neurol.*, vol. 26, pp. 27–40, 2019.

[2] L. Hone, G. Giovannoni, R. Dobson, and B. M. Jacobs, "Predicting Multiple sclerosis: Challenges and opportunities," *Front. Neurol.*, vol. 12, 2022.

[3] M. Filippi, P. Preziosa, B. L. Banwell, F. Barkhof, *et al.*, "Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines," *Brain*, vol. 142, no. 7, pp. 1858–1875, 2019.

[4] W. I. McDonald, A. Compston, G. Edan, D. Goodkin, *et al.*, "Recommended diagnostic criteria for multiple sclerosis: Guidelines from the International Panel on the Diagnosis of Multiple Sclerosis," *Ann. Neurol.*, vol. 50, no. 1, pp. 121–127, 2001.

[5] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983.

[6] C. P. Loizou, S. Petroudi, I. Seimenis, M. Pantziaris, and C. S. Pattichis, "Quantitative texture analysis of brain white matter lesions derived from T2-weighted MR images in MS patients with clinically isolated syndrome," *J. Neuroradiol.*, vol. 42, no. 2, pp. 99–114, 2015.

[7] C. P. Loizou, V. Murray, M. S. Pattichis, I. Seimenis, M. Pantziaris, and C. S. Pattichis, "Multiscale amplitude-modulation frequency-modulation (AM-FM) texture analysis of multiple sclerosis in brain MRI images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp. 119–129, 2011.

[8] C. P. Loizou, K. Fotso, A. Nicolaou, M. Pantzaris, M. S. Pattichis, and C. S. Pattichis, "Multiple sclerosis disease evolution assessment in brain MRI lesions based on texture and multi-scale amplitude modulation-frequency modulation (AM-FM) features," *IEEE Access*, vol. 11, pp. 29918–29933, 2023.

[9] A. Nicolaou, C. P. Loizou, M. Pantzaris, A. Kakas, and C. S. Pattichis, "Rule extraction in the assessment of brain MRI lesions in multiple sclerosis: Preliminary findings," in *Comput. Anal. Images Patterns CAIP 2021. Lecture Notes Comput. Sci.*, vol. 13052, N. Tsapatsoulis, A. Panayides, T. Theocharides, A. Lanitis, C. Pattichis, M. Vento, Eds. Springer Cham., 2021, pp. 277-286.

[10] K. P. Constantinou, I. Constantinou, C. S. Pattichis, and M. Pattichis, "Medical image analysis using AM-FM models and methods, *IEEE Reviews in Biomed. Engin.*, vol. 14, pp. 270-289, 2020

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 108-122, 2013.

[12] N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. *Artif. Intell. Res*., vol.16, pp. 321-357, 2002.

[13] G. R. Lal, X. Chen, and V. Mithal, "TE2Rules: Extracting rule lists from tree ensembles," pp. 1–17, 2022, [Online]. Available: http://arxiv.org/abs/2206.14359.

[14] A. C. Kakas, P. Moraitis, and N. I. Spanoudakis, "GORGIAS: Applying argumentation," *Argument Comput.*, vol. 10, no. 1, pp. 55–81, 2019.

[15] N. I. Spanoudakis, G. Gligoris, A. C. Kakas, and A. Koumi, "Gorgias cloud: On-line explainable argumentation," *Front. Artif. Intell. Appl.*, vol. 353, pp. 371–372, 2022.

[16] A. Lopatina, S. Ropele, R. Sibgatulin, J. R. Reichenbach, and D. Güllmar, "Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis," *Front. Neurosci.*, vol. 14, pp. 1–12, 2020.

[17] F. Eitel, E. Soehler, J. Bellmann-Strobl, A. U. Brandt, *et al.*, "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation," *NeuroImage Clin.*, vol. 24, 2019.

[18] A. Conti, C. A. Treaba, A. Mehndiratta, V. T. Barletta, C. Mainero, and N. Toschi, "An interpretable machine learning model to predict cortical atrophy in multiple sclerosis," *Brain Sci.*, vol. 13, no. 2, p. 198, 2023.