













Explainable Few-Shot Learning for Multiple Sclerosis Detection in Low-Data Regime

Montassar Ben Dhifallah^{1,6}^{*}, Dalel Kanzari^{1,2}, Selma Naija^{3,4}, Sana Ben Amor^{3,4}, Ahmed Zrig^{5,7}, Mezri Maatouk^{5,7}, Mabrouk Abdelaali^{5,7}, Jamel Saad^{5,7}, Asma Achour^{5,7}, Sofiane Gaied Chortane^{5,7}, Maher Hadhri^{6,7}, Ahmed Dahmoul^{5,7}, Azza Ben Ali^{5,7}, Sahar Selim⁸, and Ahmed Nebli⁹

¹ Higher Institute of Applied Sciences and Technology, University of Sousse, Tunisia

² Operational Research, Decision and Process Control Laboratory (LARODEC), 41 Liberty Street, Bardo, 2000, Tunis, Tunisia

³ Sahloul Hospital, Department of Neurology, Sousse, Tunisia

⁴ University of Sousse, Faculty of Medicine of Sousse, Tunisia

⁵ Department of Radiology A, Fattouma Bourguiba Hospital, Monastir, Tunisia

⁶ Department of Neurosurgery, Fattouma Bourguiba Hospital, Monastir, Tunisia

⁷ Research Unity Interventional radiology LR18SP08, University of Monastir, Tunisia

⁸ School of Information Technology and Computer Science, Nile University, Giza, Egypt

⁹ Independent Researcher

Abstract. Diagnosing multiple sclerosis (MS) accurately is highly challenging due to symptom overlap with other demyelinating diseases. Here, we present DemyeliNeXt, an explainable few-shot learning framework designed to classify MS and other demyelinating diseases from MRI scans. This framework employs a prototypical network with a 3D DenseNet-121 backbone and uses Deep SHAP for feature importance visualization. We train our DemyeliNeXt on a dataset from African populations and we test it for different datasets including MICCAI MSSEG2 public dataset. Our findings demonstrate robust performance across diverse datasets highlighting the model’s potential to enhance diagnosis accuracy and generalizability in various clinical settings.

Keywords: Few-Shot Learning · Explainable AI · Multiple Sclerosis · 3D MRI · and Deep Learning

1 Introduction

Multiple sclerosis (MS) is a complex neurological condition that is often misdiagnosed due to its symptom overlap with other conditions such as vasculitis and vascular leukoencephalopathy. Studies indicate that over half of the patients were misdiagnosed for a period exceeding three years [2,12]. Moreover, 70% of these patients had been administered disease-modifying therapies (DMTs), and 31% suffered unnecessary morbidity due to the incorrect diagnosis and treatment

^{*} corresponding author: montassar.bendhifallah@issatso.u-sousse.tn

[2,12]. This diagnostic challenge results in a prolonged time to achieve a definitive diagnosis, often exceeding several months. Hence, accurate and timely diagnosis is crucial for effective management and treatment planning in MS patients. Advanced imaging techniques and biomarker analyses are increasingly important in differentiating MS from other similar presenting conditions, thereby reducing diagnostic errors and improving patient outcomes. Machine learning provides a robust approach for the analysis of medical images and the diagnosis of MS.

In this context, several studies have employed machine learning models for MS classification. For instance, Wang et al. [15] employed a multi-layer convolutional neural network (CNN) with data augmentation techniques to classify MS. However, the model’s explainability remains unexplored. To address this issue, Zhang et al. [17] proposed a classification model for MS subtypes based on VGG19 [10] with global average pooling and utilized Grad-CAM++ [1] for model explanation. While effective in performance and interpretability, this approach did not account for the diversity of MS data, particularly by not comparing it with other similar demyelinating diseases such as vasculitis. To rectify this concern, Huang et al. [3] leveraged a Transformer-based model with a Multiple Instance Learning (MIL) strategy to discriminate between MS and various demyelinating diseases. The authors used Grad-CAM to visualize feature extraction through activation heatmaps. Nevertheless, their study did not incorporate data from low-income countries, such as datasets from the African population. This omission underscores a critical gap, as regional genetic and environmental factors influence disease onset and progression [16]. These factors impact the timeliness and accuracy of MS diagnosis, thereby potentially threatening the patient’s life.

Additionally, the collection of MS and other demyelinating diseases data is challenging due to the variability in disease presentation, limited patient availability, and the high cost of medical imaging. Therefore, the application of few-shot learning is essential to leverage limited data effectively. Furthermore, a key finding in MS identification is the presence of white matter lesions in the brain, detectable via Fluid Attenuated Inversion Recovery (FLAIR) sequence of MRI.

This study focuses on distinguishing MS from other demyelinating diseases. We introduce DemyeliNeXt, an explainable few-shot learning framework for the classification of MS and other demyelinating diseases. Our approach employs a prototypical network with a 3D DenseNet-121 backbone, which integrates spatial information from FLAIR MR (Magnetic Resonance) images to classify them as MS vs other demyelinating diseases (NON-MS). Additionally, the framework provides model interpretability through the Deep SHAP model for visualizing the most important features leading to the classification of the input MRI. The primary contributions of our work are as follows:

1. Application of Few-Shot Learning: We apply few-shot learning for the detection of multiple sclerosis (MS).
2. Emphasis on Explainability: Our method integrates explainability mechanisms to enhance interpretability, making it more suitable for clinical settings.

3. Utilization of African 3D MRI Data: We trained our model using 3D MRI data from African populations, which are often underrepresented in medical datasets. By benchmarking our model against MICCAI MS public dataset, we demonstrated its robust performance, thereby validating its generalizability across diverse populations.

2 Proposed Method

In this section, we explain the key building blocks of our proposed DemyeliNeXt architecture for explainable MS identification from other demyelinating diseases.

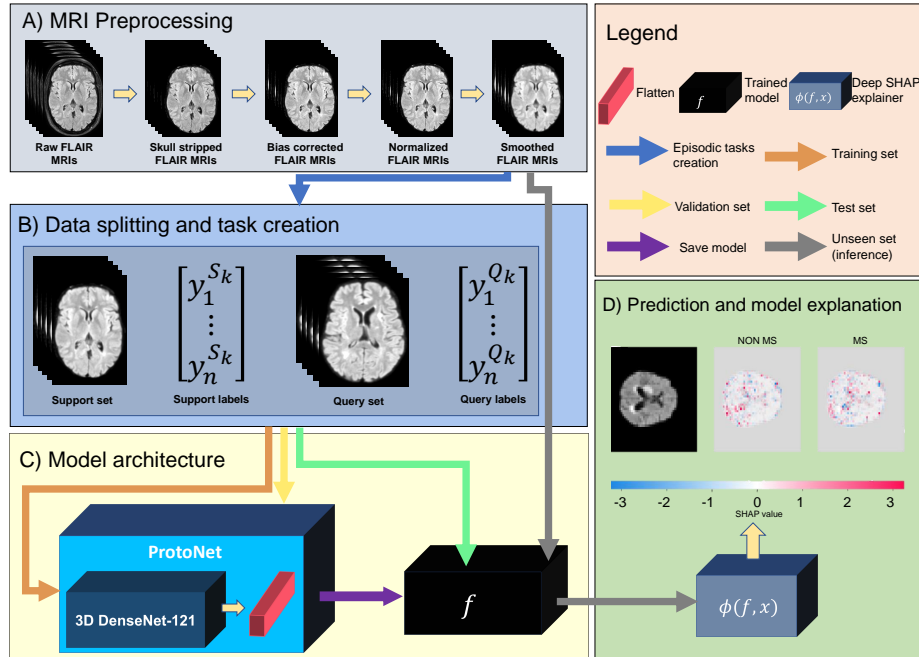


Fig. 1. *DemyeliNeXt Pipeline.* (A) Preprocessing MRI scans: includes skull stripping, bias correction normalization, and FLAIR MRI smoothing. (B) Data splitting into support and query sets. (C) Training a prototypical network with 3D DenseNet-121 backbone. (D) Model testing on unseen MRIs with explanations provided using Deep SHAP.

2.1 Architecture overview

In this study, we introduce DemyeliNeXt, a four-stage pipeline designed for the classification of multiple sclerosis (MS) and other demyelinating diseases from

MRI scans, while also providing model interpretability. Figure 1 illustrates the first stage (Section 2.2), which involves a preprocessing pipeline for FLAIR MRI scans. Here, raw FLAIR images are normalized, while noise and artifacts are reduced. In the second stage, the MRI scans are divided into training, validation, and testing sets. Each set contains a support set (S) with labeled examples to update model parameters and a query set (Q) with unlabeled examples for performance evaluation.

The third stage (Section 2.3) involves training a 3D DenseNet-based (DenseNet-121) [4] prototypical network to classify the preprocessed MRIs. The training process utilizes N^{tr} training tasks, each comprising N_{shots} support examples for model weight updates and N_{query} query examples for performance assessment. In the final stage, we employ Deep SHAP [7] to approximate the model for interpretability. Deep SHAP, inspired by DeepLIFT [9], assigns importance scores to each input feature by propagating neuron contributions backward through the network. These scores are based on the difference from a reference input, known as the "baseline" or "background" input, representing a typical or neutral state for the input features. The importance scores are computed via the combination of the model's weights, the actual input and the baseline input. After training the explainer, we use the model and explainer to predict and interpret new examples of MS and other demyelinating diseases during inference.

2.2 Preprocessing Pipeline

We begin our preprocessing pipeline by anonymizing DICOM MRI scans, converting them to NIfTI format. This process removes patient metadata and consolidates each volume into a single file. Next, we perform skull stripping using the ROBEX algorithm [5] to eliminate non-brain tissues. We then apply bias field correction using the N4ITK algorithm [14] to remove low-frequency intensity non-uniformities. Following this, we normalize MRI intensities to a range of 0 to 1. We reduce the noise using a Gaussian filter. Finally, we reorient the images to the "IPL" (Inferior, Posterior, Left) orientation, resample them to isotropic voxels, and resize them to a standard format.

2.3 Few shot learning

Prototypical network. Prototypical Networks (ProtoNet) [11] seek to find a metric space in which samples from the same class are close to one another. This approach makes the model particularly useful in settings with limited labeled data. Based on the prototype concept [11], the model depicts each class using the mean of its embedded support set S . Prototypical Networks then determine query samples Q based on their proximity to these prototypes. To generate the image embeddings, we use a 3D DenseNet-121 [4] as a backbone. We employed Euclidian distance for our ProtoNet to calculate the distance between the support samples and query samples. We create dataset episodes using a sampler that follows uniform distribution to load data from the dataset for each label.

Loss function We use binary cross-entropy loss:

$$\mathcal{L} = -[y \log(p) + (1 - y) \log(1 - p)] \quad (1)$$

where y and p are the MS label and the predicted probability of MS from the model respectively. We use ADAM [6] as an optimizer with step LR scheduler to decay the learning rate.

2.4 Explainability with Deep SHAP

Deep SHAP [7] approximates explanations for deep neural network models using SHAP (SHapley Additive exPlanations) values to quantify feature importance. This method integrates concepts from a deep learning explanation technique called DeepLIFT [9] that uses Shapley values [8]. We apply Deep SHAP to interpret our trained 3D DenseNet-based ProtoNet model using preprocessed MRI scans from the testing dataset. This approach creates a simplified explanation model, assessing the importance of each voxel in our testing MRIs, visualized through feature importance plots.

2.5 Model inference and explanation

After training and evaluating the model, we perform inference on unseen examples where we pass them to the explainer to check the used feature importance of the model on the classification of the new examples.

3 Results and discussion

In this section, we provide a quantitative evaluation of our model on three distinct datasets and we display the findings of the used Deep SHAP.

3.1 Employed datasets

In this work, we utilized three labeled datasets, summarized in Table 1. We trained, validated, and tested using a set that comprises 182 FLAIR MRI scans from 121 patients with multiple sclerosis (MS) and other demyelinating diseases (NON-MS). The dataset was split randomly and patient-wise into three different sets as follows: 70% for training, 15% for validation and 15% for testing. This dataset is sourced from the radiology department at CHU Fattouma Bourguiba Monastir (FBM), Tunisia. It includes 3D and axial scans: 91 scans from 52 MS patients and 91 scans from 69 patients with other demyelinating diseases such as vasculitis and vascular leukopathy.

We tested our model on a set containing 91 FLAIR MRI scans from 36 MS patients, obtained from the MRI center of CHU Sahloul Sousse (SS), Tunisia. Additionally, we used 80 3D FLAIR MRI scans from 40 patients in the MICCAI 2021 MS Segmentation Challenge (MSSEG-2) as a benchmark dataset. We randomly sampled data from each set to create episodes consisting of a support set and a query set. Prior to training, gamma correction was applied to all scans using $\gamma = 2.5$. No further data augmentation was performed.

Table 1. Datasets statistics

Source	Number of patients	Number of scans	Age	Gender
CHU FBM, Tunisia	MS: 52	MS: 91	21-63	MS: 22M/30F
	NON-MS: 69	NON-MS: 91		NON-MS: 19M/50F
CHU SS, Tunisia	36 MS	91	NA	4M/32F
MSSEG-2	40 MS	80	NA	NA

3.2 Experimental settings.

Parameter settings For model training, we used an ADAM optimizer [6] with a learning rate of 0.001. We applied learning rate decay for every single step by 0.1 using a step scheduler. We employed dropout with 20% rate. As for Deep SHAP explainer training, we adopted 90 background examples. We trained our model and our explainer on the Nvidia RTX 3090 GPU.

Hyperparameter Settings We conducted three distinct training experiments using 2-way ($K = 2$) classification. Validation was performed with 100 episodes ($N^{val} = 100$) every 500 training episodes. Testing was also conducted with 100 episodes. Each training lasted for 1000 episodes. Detailed hyperparameters for each experiment are listed below:

- **Experiment A:** Trained with 5 examples in both support and query sets ($N_{shots} = 5$, $N_{query} = 5$).
- **Experiment B:** Trained with 3 examples in both support and query sets ($N_{shots} = 3$, $N_{query} = 3$).
- **Experiment C:** Trained with 1 example in both support and query sets ($N_{shots} = 1$, $N_{query} = 1$).
- **Test 1:** We used the saved model from Experiment A to test on 91 scans from CHU SS MS dataset and on 13 scans from CHU FBM NON-MS test set.
- **Test 2:** We used the saved model from Experiment A to test on 80 scans from MSSEG-2 and on 13 scans CHU FBM NON-MS test set.

3.3 DemyeliNeXt evaluation

Table 2 shows the classification accuracy, precision, recall, specificity, and F1 scores for the different experiments detailed in Section 3.2. Across all experiments, Test 2, which involved training on an African dataset and testing on a combination of African and European datasets, achieved the highest classification accuracy. This result may indicate that our model has the ability to generalize well across different populations despite the differences in socio-economic conditions between the subjects in each of the datasets.

In contrast, Experiment C and B, which utilized one, and three shots and queries, respectively, demonstrated the lowest performance. This indicates that

Table 2. Experiments results

Experiments/Tests	Accuracy	MS specific Accuracy	NON-MS specific Accuracy	Precision	Recall	Specificity	F1 score
A: 5 shots 5 queries (Dataset: CHU FBM)	78.8%	-	-	0.75	0.87	0.71	0.8
B: 3 shots 3 queries (Dataset: CHU FBM)	63.83%	-	-	0.62	0.72	0.56	0.67
C: 1 shot 1 query (Dataset: CHU FBM)	65.0%	-	-	0.64	0.68	0.62	0.66
Test 1: 5 shots 5 queries (Dataset: CHU SS MS + CHU FBM NON-MS)	75.5%	68.6%	82.4%	0.8	0.69	0.82	0.74
Test 2: 5 shots 5 queries (Dataset: MSSEG-2 + CHU FBM NON MS)	87.8%	85%	90.6%	0.9	0.85	0.91	0.87

reducing the number of shots below a certain threshold adversely affects model accuracy. These findings suggest that while reducing shots can decrease computational demands, maintaining an adequate number of shots is critical for reliable performance (see experiment A). In particular, one could generally recommend using the model trained in Experiment A as a guide for practitioners in balancing computational efficiency with diagnosis accuracy for MS.

Figure 2 illustrates the explanation of our model backbone on unseen MS and NON-MS examples with lesion annotation. The plot highlights the features utilized by our trained ProtoNet model for classification that are explained by the Deep SHAP method. We evaluated the explainer results using the key diagnostic features outlined in the McDonald criteria [13], which include lesion size, number of lesions, lesion location, lesion contrast, and lesion shape. The Deep SHAP explainer seems to identify some of the key features for classification, specially the lesions in MS example (Fig.2 B). However, one should note that there is a risk that the included features in the explanation could be deemed irrelevant to clinicians.

Limitations and future studies. Despite the promising results, DemyeliNeXt has a few limitations that warrant further investigation. For instance, our approach currently utilizes only FLAIR MRI scans; incorporating other imaging modalities like T1-weighted and T2-weighted MRIs could potentially enhance diagnostic accuracy. While Deep SHAP provides some level of explainability, the clinical relevance of the highlighted features remains uncertain, indicating a need for further refinement. In future studies, we aim to benchmark against state-of-the-art methods. We will also focus on expanding the dataset to in-

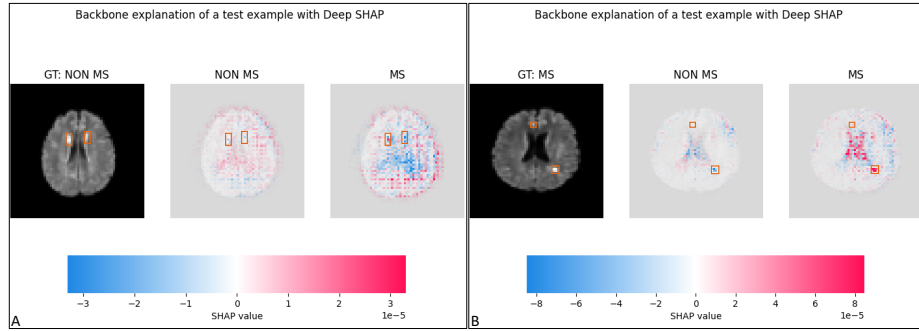


Fig. 2. *Deep SHAP Explanation of MS and NON-MS Examples.* **A:** Explanation of NON-MS example. **B:** Explanation of MS example. For each of the subfigures (A and B), the left panel displays an annotated MRI section of a patient with a NON-MS demyelinating disease (A) and a patient with MS disease (B). The center panel highlights the features identified by our model for classifying the case as NON-MS using Deep SHAP. The right panel shows the features identified for classification as MS using Deep SHAP. Lesions' locations are highlighted with orange rectangles across all panels. For the two right hand side panels, blue indicates the features excluded by the model, while red shows the important features for each class

clude diverse minority populations, integrating multimodal imaging techniques, as well as developing more clinically relevant explainability methods with their evaluation.

4 Conclusion

In this study, we introduced DemyeliNeXt, an explainable few-shot learning framework designed for the classification of multiple sclerosis (MS) and other demyelinating diseases in an African population. By incorporating the Deep SHAP model, we provided visual explanations for the model's decisions, enhancing its interpretability. Our findings, derived from MRI data of underrepresented African populations, demonstrate that this approach can generalize effectively to non-African datasets. Although the classification accuracy decreases with fewer shots, the method remains computationally efficient and can aid practitioners in improving diagnostic accuracy. In future work, we aim to extend our framework by including more minority populations and integrating additional neuroimaging modalities, thereby enhancing the generalizability and robustness of our model.

Acknowledgments. The data collection for this study was conducted under the agreement of the head of radiology department of CHU Fattouma Bourguiba Monastir, Tunisia, the head of neurology department of CHU Sahloul Sousse, Tunisia and the director of MRI center of Sahloul, Sousse, Tunisia.

Code availability. We provide the code repository of our method on GitHub at this link: <https://github.com/Montassar-bdh/DemyeliNeXt>

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks pp. 839–847 (2018). <https://doi.org/10.1109/WACV.2018.00097>
2. Gaitán, M.I., Correale, J.: Multiple sclerosis misdiagnosis: a persistent problem to solve. *Frontiers in Neurology* **10**, 466 (2019)
3. Huang, C., Chen, W., Liu, B., Yu, R., Chen, X., Tang, F., Liu, J., Lu, W.: Transformer-Based Deep-Learning Algorithm for Discriminating Demyelinating Diseases of the Central Nervous System With Neuroimaging. *Frontiers in Immunology* **13**, 897959 (Jun 2022). <https://doi.org/10.3389/fimmu.2022.897959>, <https://www.frontiersin.org/articles/10.3389/fimmu.2022.897959/full>
4. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (July 2017)
5. Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods **30**(9), 1617–1634. <https://doi.org/10.1109/TMI.2011.2138152>
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions pp. 4765–4774 (2017), <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
8. Shapley, L.S.: A value for n-person games pp. 307–317 (1953). <https://doi.org/doi:10.1515/9781400881970-018>
9. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences p. 3145–3153 (2017)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR* **abs/1409.1556** (2014), <https://api.semanticscholar.org/CorpusID:14124313>
11. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning **30** (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf
12. Solomon, A.J., Bourdette, D.N., Cross, A.H., Applebee, A., Skidd, P.M., Howard, D.B., Spain, R.I., Cameron, M.H., Kim, E., Mass, M.K., et al.: The contemporary spectrum of multiple sclerosis misdiagnosis: a multicenter study. *Neurology* **87**(13), 1393–1399 (2016)
13. Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S., Fujihara, K., Galetta, S.L., Hartung, H.P., Kappos, L., Lublin, F.D., Marrie, R.A., Miller, A.E., Miller, D.H., Montalban, X., Mowry, E.M., Sorensen, P.S., Tintoré, M., Traboulsee, A.L., Trojano, M., Uitdehaag, B.M.J., Vukusic, S., Waubant, E., Weinshenker, B.G., Reingold, S.C., Cohen, J.A.: Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria **17**(2), 162–173. [https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2), <https://linkinghub.elsevier.com/retrieve/pii/S1474442217304702>
14. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: Improved n3 bias correction. *IEEE Transactions on Medical*

- Imaging **29**(6), 1310–1320 (Jun 2010). <https://doi.org/10.1109/tmi.2010.2046908>, <http://dx.doi.org/10.1109/TMI.2010.2046908>
15. Wang, S.H., Tang, C., Sun, J., Yang, J., Huang, C., Phillips, P., Zhang, Y.D.: Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling. *Frontiers in Neuroscience* **12**, 818 (Nov 2018). <https://doi.org/10.3389/fnins.2018.00818>, <https://www.frontiersin.org/article/10.3389/fnins.2018.00818/full>
16. Waubant, E., Lucas, R., Mowry, E., Graves, J., Olsson, T., Alfredsson, L., Langer-Gould, A.: Environmental and genetic risk factors for ms: an integrated review. *Annals of clinical and translational neurology* **6**(9), 1905–1922 (2019)
17. Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., Slaney, G.: Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods* **353**, 109098 (Apr 2021). <https://doi.org/10.1016/j.jneumeth.2021.109098>, <https://linkinghub.elsevier.com/retrieve/pii/S0165027021000339>