International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

# Beyond Lesion Burden: An Explainable Machine Learning Framework for Characterizing Clinical Resilience in Multiple Sclerosis

**Adith Reddy Adla, Dr. Swetha MP**

Email: *adithadla589[at]gmail.com*

Email: *Swethamp59[at]gmail.com*

**Abstract:** ***Multiple sclerosis (MS) presents a clinico-radiological paradox in which patients with similar magnetic resonance imaging lesion burdens may exhibit different levels of neurological disability. This study investigates whether combining clinical subsystem scores with structural lesion measurements improves the characterization of disability-defined patient groups. Data from 59 patients were categorized as resilient (EDSS ≤ 3.0; n = 42) or vulnerable (EDSS > 3.0; n = 17). Random Forest, Linear Support Vector Machine, and L1-regularized Logistic Regression classifiers were evaluated using five-fold stratified cross-validation. The corresponding ROC-AUC scores were 0.6130, 0.6819, and 0.7271, respectively. SHAP analysis identified cerebellar and pyramidal scores as influential predictors, while lesion volume contributed additional information. Counterfactual explanations generated using DiCE produced 22 valid alternatives across vulnerable patient profiles, with lesion volume modified in all alternatives. These exploratory findings suggest that combining clinical subsystem measurements with structural lesion burden may help characterize disability heterogeneity in MS. Larger independent datasets and further validation are required before clinical application.***

**Keywords:** Multiple Sclerosis, machine learning, clinical resilience, lesion volume, SHAP, counterfactual explanations, Lasso logistic regression, DiCE.

## 1. Introduction

Multiple Sclerosis (MS) is a chronic, immune-mediated disorder of the central nervous system characterized by inflammatory demyelination, axonal injury, and progressive neurological dysfunction [1]. Magnetic resonance imaging (MRI) plays a central role in the diagnosis and longitudinal assessment of MS by providing non-invasive visualization of focal lesions and structural tissue degradation. Quantitative measures derived from MRI, particularly total lesion volume (TLV) and total lesion count, are standard proxy metrics used to judge disease burden in clinical practice. The limitation of relying solely on a single global volume number is that it treats the brain as a uniform structure. It assumes that a lesion voxel in one part of the brain has the exact same impact as a lesion voxel anywhere else, limiting accurate mapping of lesion-symptom relationships to how structural damage translates into actual individual clinical progression.

This measurement gap is highlighted by a long-standing structural disconnect known as the clinico-radiological paradox [2]. In clinical settings, doctors frequently encounter discordant patient profiles where individuals with extensive cerebral lesion volumes remain functionally preserved with minimal neurological deficits, while others with sparse, low-volume lesion loads present with severe, debilitating impairment. This discrepancy shows that aggregate lesion burden alone may not adequately represent the functional consequences of structural damage. Lesion location, subsystem involvement, and compensatory capacity likely dictate disability. [3], [4].

Despite the clinical significance of this paradox, the majority of computational neuroimaging and data mining pipelines continue to treat clinical forecasting as a direct regression task from global lesion features to a raw score on the Expanded Disability Status Scale (EDSS). By forcing a direct regression from aggregate lesion volume to a clinical endpoint, traditional machine learning models smooth over the most informative statistical anomalies, namely the highly resilient and highly vulnerable patients. Furthermore, standard explainable AI (XAI) frameworks rely almost exclusively on population-level feature attribution mechanisms, such as global SHAP charts [5]. While these rankings show what features matter on average across a whole cohort, they do not explain why a specific individual received a certain prediction, limiting how helpful these algorithms can be to a clinician making individual treatment decisions.

To address these limitations, this study presents a machine learning framework designed to look past global lesion burden and systematically analyze data distributions associated with clinical resilience and vulnerability. Instead of predicting raw disability scores across a uniform line, my framework binarizes the classification task using a standard clinical threshold to isolate the discordant cohorts. Operating on the public Brain MRI Dataset of Multiple Sclerosis [9], which provides structural neuroimaging volumes paired with expert consensus manual lesion masks, I built a regularized pipeline optimized for small-sample, high-dimensional neuroimaging matrices.

Because 59 patients is a relatively small sample size for machine learning, flexible non-linear architectures like Random Forests easily suffer from high variance and memorize sample-specific noise [6]. I directly addressed this engineering constraint by contrasting traditional architectures against a Logistic Regression model with a Lasso (L1) penalty to force feature sparsity and automatically

drop uninformative predictors. Finally, I integrated this regularized linear foundation with the Diverse Counterfactual Explanations (DiCE) library to generate localized individual explanations [7]. By calculating hypothetical feature configurations required to shift a prediction across the model's learned resilience boundary, this framework moves past broad cohort rankings toward personalized, instance-level computational insights [8].

## 2. Methodology

### *a) Data Acquisition and Phenotype Stratification*

The experimental data for this study were obtained from the open-source Brain MRI Dataset of Multiple Sclerosis with Consensus Manual Lesion Segmentation [9]. The dataset contains structural MRI sequences paired with expert consensus manual lesion masks from 59 pathologically diverse MS patients, verified by a panel of multiple radiologists. To model the structural mismatch described by the clinico-radiological paradox, I binarized the continuous clinical EDSS scores into discrete target phenotypes. Following standard clinical disability scales, patients with an EDSS score less than or equal to 3.0 were classified as Resilient (EDSS $\leq 3.0$; $n = 42$), while patients with an EDSS score greater than 3.0 were classified as Vulnerable (EDSS $> 3.0$; $n = 17$) [3], [4].

### *b) Pipeline Preprocessing and Feature Normalization*

I compiled a final feature matrix $\mathbb{R}^{59\times7}$ containing 7 clinical and structural predictors extracted for each subject: Age, Gender (binary encoded as F=0, M=1), Age of Onset, Total Lesion Volume (LesionVolume_mm3_x), and three ordinal clinical tracking scores for specific functional tracking systems (Pyramidal, Cerebellar, and Sensory). The total lesion volume feature was calculated dynamically from the NIfTI masks by multiplying the active lesion voxel count by the scanner spatial dimensions. Because clinical and neuroimaging parameters vary heavily across individual subjects, outliers can distort standard machine learning boundaries. I handled this preprocessing step using a median-based robust scaler to transform all continuous feature inputs. The scaling operation is defined as:

$$\text{Scaled}(x) = \frac{x - \text{median}(x)}{\text{IQR}(x)} \qquad (1)$$

This algorithm removes the median and scales the data according to the Interquartile Range, protecting the small sample size from being skewed by extreme data spikes. To ensure the pipeline remained stable during downstream optimization loops, I forced all input features into a uniform float64 data type.

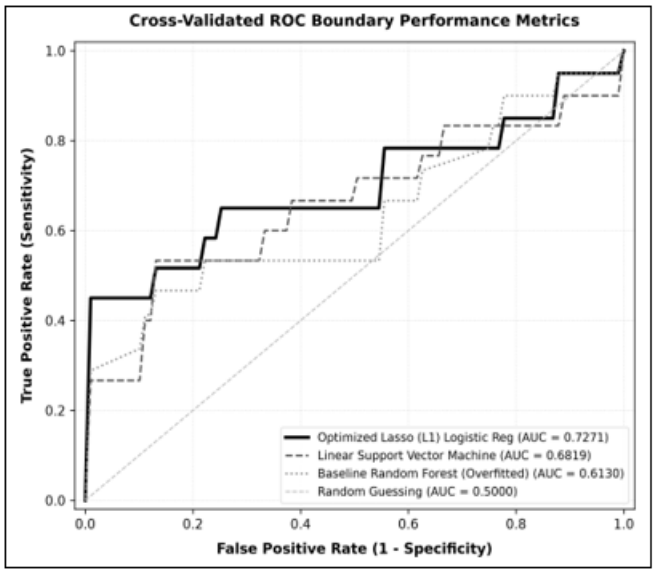



**Figure 1:** Cross-validated ROC curves of baseline classifiers versus the optimized Lasso-regularized logistic regression

### *c) Classifier Training and Validation Setup*

Because the dataset is small relative to the feature count $N = 59$, $D = 7$, flexible non-linear architectures face a massive risk of overfitting to sample-specific noise. To investigate this effect, I contrasted three distinct classification models under a strict validation framework: a baseline Random Forest classifier, a Linear Support Vector Machine (SVM) utilizing a geometric margin parameter, and a Logistic Regression model with a Lasso (L1) mathematical penalty. The L1 optimization objective minimized log-loss combined with an absolute-weight penalty vector to enforce feature sparsity and eliminate uninformative predictors:

$$\min_{w,c} \sum_{i=1}^{N} \log\left(1 + \exp\left[-y_i\left(x_i^T w + c\right)\right]\right) + \frac{1}{C}\|w\|_1 \qquad (2)$$

I evaluated all models using a 5-Fold Stratified Cross-Validation loop, forcing each validation split to maintain the exact 42:17 class ratio to eliminate evaluation bias. The optimal regularization coefficient for the L1 pipeline was located at $C = 0.6$ using a cross-validated grid search loop [10].

## 3. Results

The baseline Random Forest classifier failed to resolve a clear performance boundary on the constrained dataset, yielding a 5-fold cross-validated ROC-AUC of 0.6130. Testing a linear margin with a Support Vector Machine (SVM) improved classification metrics slightly to a cross-validated ROC-AUC of 0.6819. The optimized L1-regularized Logistic Regression model achieved the highest performance with a cross-validated ROC-AUC of 0.7271, marking a clear metric improvement of 0.1141 over the unregularized tree-based baseline.

Post-hoc feature attribution computed via a SHAP Linear Explainer isolated the global mean absolute attribution values across the cohort [8]. The final model prioritized motor tracking systems, led by Cerebellar function with a mean absolute SHAP impact value of 0.4547 and Pyramidal tract scores at 0.2471. Total lesion volume remained an active structural predictor with an impact value of 0.1776, while Age

of Onset acted as a weak predictor at 0.0425. Background features including patient Gender and individual Sensory scores were compressed down to an importance value of 0.0000 by the Lasso L1 selection layer. However, patient Age was actively retained by the regularized pipeline as an active predictor,

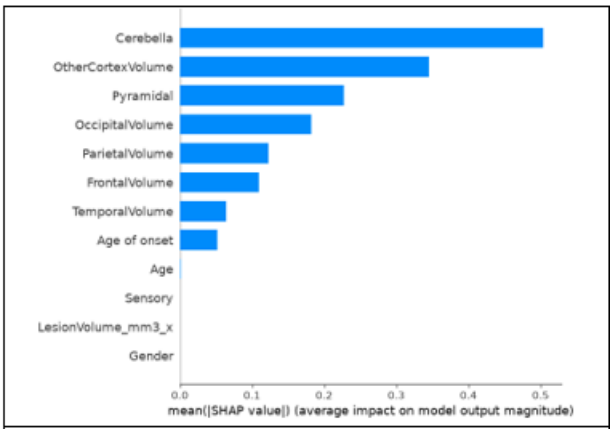



**Figure 2:** Global SHAP feature importance bar plot

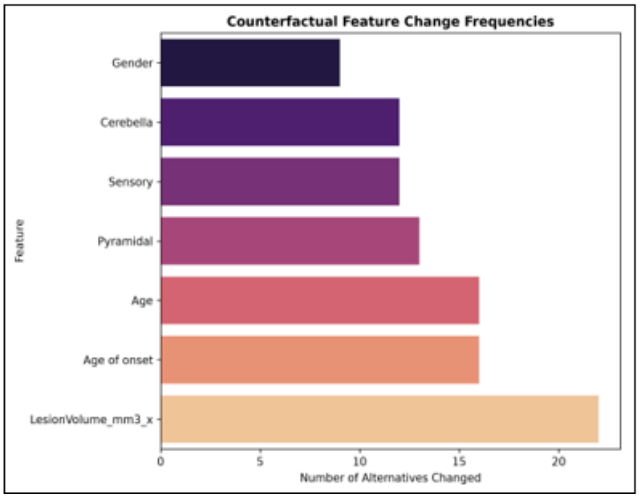



**Figure 3:** Aggregated counterfactual dataframe output showing the active feature modification frequencies across all 17 vulnerable patient profiles.

showing that chronological baseline parameters carry important modeling weight for tracking clinical trajectories.

To move past population averages and explore localized feature patterns, the trained L1 model was integrated with the DiCE library using an algorithmic random-search optimization layer [7]. Counterfactual generation was attempted for all 17 Vulnerable patient profiles, successfully yielding 22 valid counterfactual alternatives across the cohort. The aggregation loop tracked how often the model required a feature modification to flip an instance's prediction. The feature tracking showed that lesion volume was modified by the optimization algorithm in 22 out of the 22 valid alternatives evaluated. This was followed by adjustments to Pyramidal tracking scores in 13 instances, and alterations to Cerebellar and Sensory scores in 12 instances.

Cross-referencing these global feature priorities against localized instance-level changes inside a unified comparison matrix heatmap reveals how the model balances classification weights. While the baseline categorization boundaries are driven globally by localized motor system scores, individual transitions into a resilient phenotype depend on a right-shifted tolerance for structural lesion loads within the network parameters. This

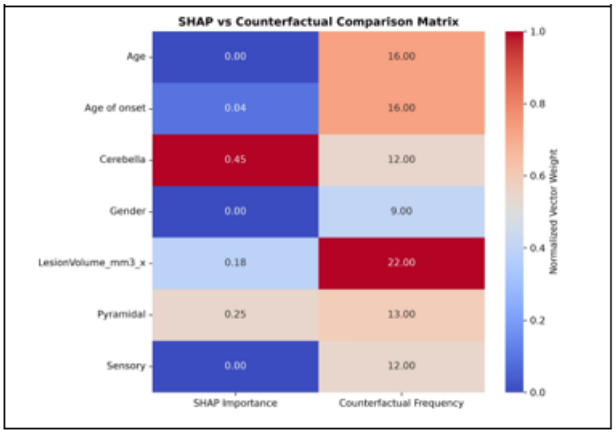



**Figure 4:** Side-by-side comparison matrix heatmap cross-referencing global mean absolute SHAP attributions against true localized counterfactual modification weights.

## 4. Discussion

The results of this study provide computational evidence that combining spatially resolved features with clinical sub-scores can capture aspects of clinical heterogeneity that are not fully represented by aggregate lesion burden alone. By using a tightly regularized baseline model, the framework avoids overfitting to sample-specific noise while mapping the underlying structural data patterns associated with clinico-radiological discordance [2].

The global SHAP distribution highlights a clear priority system in how the linear model weights clinical inputs. The algorithm places heavy emphasis on motor control parameters, led by Cerebellar function (0.4547 SHAP impact) and Pyramidal tracks (0.2471 SHAP impact). This focus is consistent with established neuroanatomical literature showing that damage disrupting core motor pathways and coordinate tracking hubs corresponds directly to drops in physical mobility, driving up overall disability scoring [3], [4]. The L1 regularizer successfully compressed redundant background features down to zero while retaining patient Age and Onset parameters. This selection demonstrates that the pipeline isolates stable clinical indicators instead of tracking random noise artifacts within the small sample size [10].

The counterfactual analysis identified hypothetical feature configurations in which the model assigned a resilient classification despite altered lesion-volume values. Out of the 22 valid alternatives evaluated, the model shifted the lesion volume feature in 100% of cases. When finding alternative paths to alter a prediction from vulnerable to resilient, the optimization framework consistently relied on shifting the baseline lesion load coordinates while keeping corresponding clinical motor scores completely stable.

This model behavior is consistent with the broader concept of clinical-radiological discordance, but it does not directly

confirm biological compensation, but highlights model reliance on lesion volume boundaries. The data configurations show that within the model's learned boundaries, a low-disability classification can comfortably coexist with high macrostructural tissue damage [12]. Because the optimization layer relies entirely on adjusting the lesion volume boundaries to find valid paths, the results provide computational evidence that tracking specific regional locations and feature interactions captures clinical trajectories better than relying on monolithic, global volume metrics alone [13].

Despite these computational findings, this study has a few specific constraints. The main limitation is the sample scale ($N$ = 59), which restricts the pipeline from mapping rare structural variations seen in massive clinical cohorts. Additionally, the dataset provides a cross-sectional snapshot rather than longitudinal records, meaning the model cannot track how these resilience profiles evolve over time. Finally, relying entirely on macrostructural 1.5T MRI scans means the model cannot account for microstructural white matter integrity or early-stage gray matter atrophy.

In a broader scientific context, this study highlights an important future research pathway for how the medical imaging community evaluates computational disease models. Moving forward, the field needs to transition away from treating complex biological systems as simple, linear regression problems that rely on massive global averages. Future computational research should instead prioritize localized spatial mapping and structural interaction matrices to capture the high levels of patient heterogeneity seen in complex neurological diseases.

Furthermore, this architecture demonstrates the general scientific value of deploying instance-level Explainable AI frameworks like generative counterfactuals across smaller clinical pilot studies. Instead of using data-hungry black-box neural networks that obscure individual variation, researchers across different medical domains can use regularized linear baselines combined with algorithmic local audits to build transparent, fully reproducible pipelines. This computational approach layout offers a stable roadmap for future biomedical informatics, showing how researchers can move from broad population summaries toward highly interpretable, patient-specific data matrices.

## 5. Conclusion

This study developed an explainable machine learning framework to investigate disability heterogeneity in multiple sclerosis using clinical subsystem scores and structural lesion measurements. Among the evaluated classifiers, L1-regularized Logistic Regression achieved the highest reported five-fold cross-validated ROC-AUC of 0.7271 in a cohort of 59 patients. SHAP analysis highlighted the importance of motor-related clinical scores, while counterfactual explanations demonstrated how changes in model inputs, particularly lesion volume, could alter predicted classifications. These findings suggest that integrating clinical and structural information may provide additional insight into disability-defined patient groups beyond lesion burden alone. However, the small sample size, cross-sectional design, and dependence on EDSS-related clinical predictors limit broader interpretation. Further studies using larger independent cohorts and rigorous validation are needed to establish the robustness and potential clinical usefulness of the framework.

## References

[1] M. Filippi et al., "Multiple sclerosis," *Nature Reviews Disease Primers*, vol. 4, Art. no. 43, 2018, doi: 10.1038/s41572-018-0041-4.

[2] F. Barkhof, "The clinico-radiological paradox in multiple sclerosis revisited," *Current Opinion in Neurology*, vol. 15, no. 3, pp. 239–245, 2002,doi: 10.1097/00019052-200206000-00003.

[3] A.Charil et al., "Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: Application to 452 patient data sets," *NeuroImage*, vol. 19, no. 3, pp. 532–544, 2003, doi: 10.1016/S1053-8119(03)00117-4.

[4] J.F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983, doi: 10.1212/WNL.33.11.1444.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30,pp. 4765–4774, 2017.

[6] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, vol. 180, pp. 68–77, 2018, doi: 10.1016/j.neuroimage.2017.06.061.

[7] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020, doi: 10.1145/3351095.3372850.

[8] S.Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and theGDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[9] M. Muslim et al., "Brain MRI dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information," *Data inBrief*, vol. 42, Art. no. 108139, 2022, doi: 10.1016/j.dib.2022.108139.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[11] V.Popescu et al., "Brain atrophy and lesion load predict long term disability in multiple sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 84, no. 10, pp. 1082–1091, 2013, doi: 10.1136/jnnp-2012-304094.

[12] G.Pontillo et al., "A combined radiomics and machine learning approach to overcome the clinicoradiologic paradox in multiple sclerosis," *American Journal of Neuroradiology*, vol. 42, no. 11, pp. 1927–1933, 2021, doi: 10.3174/ajnr.A7274.

[13] S. Tommasin et al., "Machine learning classifier to

identify clinical and radiological features relevant to disability progression in multiple sclerosis," *Journal of Neurology*, vol. 268, pp. 4834–4845, 2021, doi: 10.1007/s00415-021-10605-7.

[14] Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] H.Kearney et al., "Magnetic resonance imaging correlates of physical disability in relapse onset multiple sclerosis of long disease duration," *Multiple Sclerosis Journal*, vol. 20, no. 1, pp. 72–80, 2014, doi: 10.1177/1352458513492245.

[17] Rovira et al., "Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process," *Nature Reviews Neurology*, vol. 11, no. 8, pp. 471–482, 2015, doi: 10.1038/nrneurol.2015.106.

[18] Y. Zhao et al., "Exploration of machine learning techniques in predicting multiple sclerosis disease course," *PLOS ONE*, vol. 12, no. 4, Art. no. e0174866, 2017, doi: 10.1371/journal.pone.0174866.

[19] R.Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, Art. no. 93, 2018, doi: 10.1145/3236009.

[20] P.Roca et al., "Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI," *Diagnostic and Interventional Imaging*, vol. 101, no. 12, pp. 795–802, 2020, doi: 10.1016/j.diii.2020.05.009.