Multi-Level Explainability in Radiomic-based Classification of Multiple Sclerosis and Ischemic Lesions[⋆]

Nighat Bibi[1,*,†], Kathleen M. Curran[2], Ronan P. Killeen[2,3] and Jane Courtney[1]

[1] *School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, Ireland*
[2] *School of Medicine, University College Dublin, UCD Belfield, Dublin, Ireland*
[3] *Department of Radiology, St Vincent's University Hospital, Dublin, ireland*

Abstract

Differentiating multiple sclerosis (MS) from ischemic stroke lesions on MRI remains a clinical challenge due to their similar appearances as white matter hyperintensities. We propose a radiomics-based machine learning framework that integrates multi-level explainable AI (XAI) techniques to support transparent and clinically meaningful lesion classification. Radiomic features are extracted from standardized MRI scans and used to train multiple classifiers, with Random Forest achieving the best performance (accuracy: 91.24%, F1: 86.54%). The framework incorporates four complementary explanation layers: global insights using SHAP, local interpretability via LIME, counterfactual reasoning with DiCE, and clinical narrative generation using GPT-based language models. This layered approach enhances interpretability at both dataset and lesion levels, enabling clinicians to understand, trust, and act upon model outputs. A radiologist who reviewed the results found the explanations helpful and confirmed that the overall analysis was clinically meaningful. Our results demonstrate the value of combining radiomics and advanced XAI techniques for differential diagnosis of brain lesions.

Keywords

Explainable AI, Radiomics, MRI, Brain Lesions, Multiple Sclerosis, Ischemic Stroke, SHAP, LIME, Counterfactual Explanations, GPT Narratives, Clinical Decision Support.

1. Introduction and Related Work

Differentiating multiple sclerosis (MS) from ischemic stroke lesions on magnetic resonance imaging (MRI) is a complex diagnostic task due to their overlapping appearance as white matter hyperintensities (WMHs). MS is a chronic inflammatory disease marked by demyelination, while ischemic lesions arise from vascular occlusion and subsequent tissue damage. Despite their distinct pathologies, both appear similar on common MRI sequences like FLAIR, complicating manual diagnosis and often requiring expert interpretation [1].

Radiomics offers a quantitative approach to analyze lesion characteristics by extracting texture, shape, and intensity-based features from medical images [2]. These features, when used with machine learning models, have shown promise in identifying subtle differences between MS and ischemic lesions. However, such models often behave like "black boxes," with limited transparency in how they make decisions, which restricts their adoption in clinical workflows.

Explainable AI (XAI) techniques have developed to tackle this issue by making model decisions interpretable. SHapley Additive exPlanations (SHAP) [3] and Local Interpretable Model-Agnostic Explanations (LIME) [4] are widely used to understand feature contributions at both global and local levels. In neuroimaging, these methods have been applied to improve the transparency of disease classifiers. For instance, Eitel et al. [5] used relevance propagation to explain CNN-based MS classification. Basu et al. [6] and Lopatina et al. [7] have explored similar approaches using clinical and imaging data. Leite et al. [8] demonstrated the use of texture features and SVM to distinguish between MS and ischemic lesions,

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

Nighat.Bibi@tudublin.ie (N. Bibi); kathleen.curran@ucd.ie (K. M. Curran); ronan.killeen@st-vincents.ie (R. P. Killeen); Jane.courtney@tudublin.ie (J. Courtney)
0000-0002-3586-7363 (N. Bibi); 0000-0003-0095-9337 (K. M. Curran); 0000-0002-9175-7855 (J. Courtney)

achieving notable accuracy on a small private dataset. Castillo et al. [9] used wavelet-transformed radiomics and machine learning to differentiate lesion types, but lacked interpretability mechanisms. Vuong et al. [10] proposed Radiomics Feature Activation Maps to enhance the interpretability of radiomic signatures by spatially localizing the regions contributing most to model predictions. Their method enables visual attribution of radiomic features at the voxel level, improving transparency and clinical trust in radiomics-based models.

In contrast to previous work, our study proposes a multi-level XAI framework that integrates four layers of interpretability: (1) SHAP for global feature importance, (2) LIME for local explanations, (3) DiCE for counterfactual reasoning, and (4) GPT-generated clinical narratives for human-aligned interpretation. This layered approach supports both technical transparency and clinical relevance. We evaluate the method using lesion-wise radiomic features extracted from two public datasets—MSSEG (for MS) and ISLES (for stroke)—and compare multiple classifiers, identifying Random Forest as the best-performing model. To the best of our knowledge, this is the first radiomic framework to combine SHAP, LIME, counterfactual explanations, and language-based narratives for transparent brain lesion classification.

2. Materials and Methods

2.1. Dataset Description

This study utilizes two publicly available MRI datasets for lesion segmentation and classification. The first is the ISLES 2022 dataset [11], which includes diffusion-weighted and FLAIR images from 250 ischemic stroke cases collected across multiple centers in Europe. The second is the MSSEG 2016 dataset [12], containing T2-FLAIR MRI scans from 53 patients diagnosed with multiple sclerosis. Lesions in both datasets were manually segmented by clinical experts, providing high-quality annotations for radiomic analysis. In total, 13489 2D images were extracted: 6281 from multiple sclerosis patients and 7208 from ischemic stroke patients. We randomly split the cases into 70% training, 15% validation, and 15% testing sets, ensuring that images from the same patient appear in only one set.

2.2. Preprocessing Pipeline

All MRI volumes were converted into 2D slices along axial, coronal, and sagittal planes. The following preprocessing steps were applied to each slice:

- **Bias Field Correction:** N4 bias correction was applied to reduce intensity non-uniformities.
- **Intensity Normalization:** Pixel intensities were scaled to the [0, 255] range using min–max normalization.
- **Denoising:** Non-local means filtering [13] was used to suppress noise while preserving texture.
- **Mask Alignment:** Lesion masks were resampled to match MRI dimensions where needed.
- **Brain Region Cropping:** Slices with low brain content were excluded; valid slices were cropped to brain region with margin.
- **Resizing and Padding:** All slices and masks were resized to 224×224 pixels with padding where necessary.

2.3. Radiomic Feature Extraction and Selection

Radiomic features were extracted on a per-lesion basis using the PyRadiomics library [14] with default parameters. Each lesion was treated as a separate connected component, resulting in the extraction of 90 quantitative features covering intensity, texture, and structural characteristics. Diagnostic and non-informative metadata were excluded, and missing values were imputed using a mean-based strategy. Feature selection was performed exclusively on the training set to prevent information leakage. A univariate ANOVA F-test was applied using `SelectKBest` (method from scikit-learn python library), selecting the top 20 features based on their statistical significance with respect to the lesion class (MS vs. Ischemic). The selected features were then applied to the validation and test sets. These features span multiple radiomic families, including first-order intensity features, gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size zone matrix (GLSZM), gray-level dependence matrix (GLDM), and neighborhood gray-tone difference matrix (NGTDM).

2.4. Classification Models

We evaluated three classifiers: Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Each model was optimized using `RandomizedSearchCV` (method from scikit-learn python library) on a combined training and validation set with a predefined split. Hyperparameters were selected based on F1-score. The final models were retrained on the full training+validation set and evaluated on the held-out test set.

2.5. Multi-Level Explainability Framework

To ensure transparency and clinical relevance, the proposed framework integrates four complementary XAI strategies:

- **Global Explanations (SHAP):** SHAP values [3] were computed for the Random Forest model to rank feature importance across the test set.
- **Local Explanations (LIME):** For individual lesion predictions, LIME [4] provided local feature attribution.
- **Counterfactuals (DiCE):** We employed the DiCE framework [15] to generate counterfactual examples that would alter the model's prediction, identifying minimal changes required to flip class.
- **Clinical Narratives (GPT):** LIME outputs (feature names, values, and contribution direction) were passed into a structured GPT-4o prompt to generate clinician-friendly narratives. The prompt instructed GPT to summarize the predicted lesion type (MS or ischemic stroke), explain the most influential features supporting the prediction, discuss features contradicting the alternative diagnosis, and conclude with the primary reason for the prediction. The explanations avoided technical jargon, used real-world MRI interpretations, and followed a concise, bolded structure for readability.

This layered approach supports both technical and human-aligned interpretability, enhancing transparency and trust in the AI-assisted diagnosis process.

3. Results and Discussion

We evaluated three classifiers on the test set and analyzed the interpretability of the best-performing model using SHAP, LIME, DiCE, and GPT-based narratives.

Classification Performance

Table 1 shows that Random Forest achieved the highest F1-score and accuracy, making it the final model for interpretation.

Table 1
Performance comparison on the test set. Best values in **bold**.

Model	Accuracy	Precision	Recall	F1-Score
SVM	90.83%	95.3%	77.79%	85.66%
Logistic Regression	90.23%	**95.5%**	75.81%	84.52%
Random Forest	**91.24%**	94.21%	**80.02%**	**86.54%**

Global Interpretability (SHAP)

SHAP values were computed to explain the contribution of each radiomic feature across the dataset. Figure 1 shows that texture features like `glcm_Idmn` and `firstorder_Skewness` were consistently important.

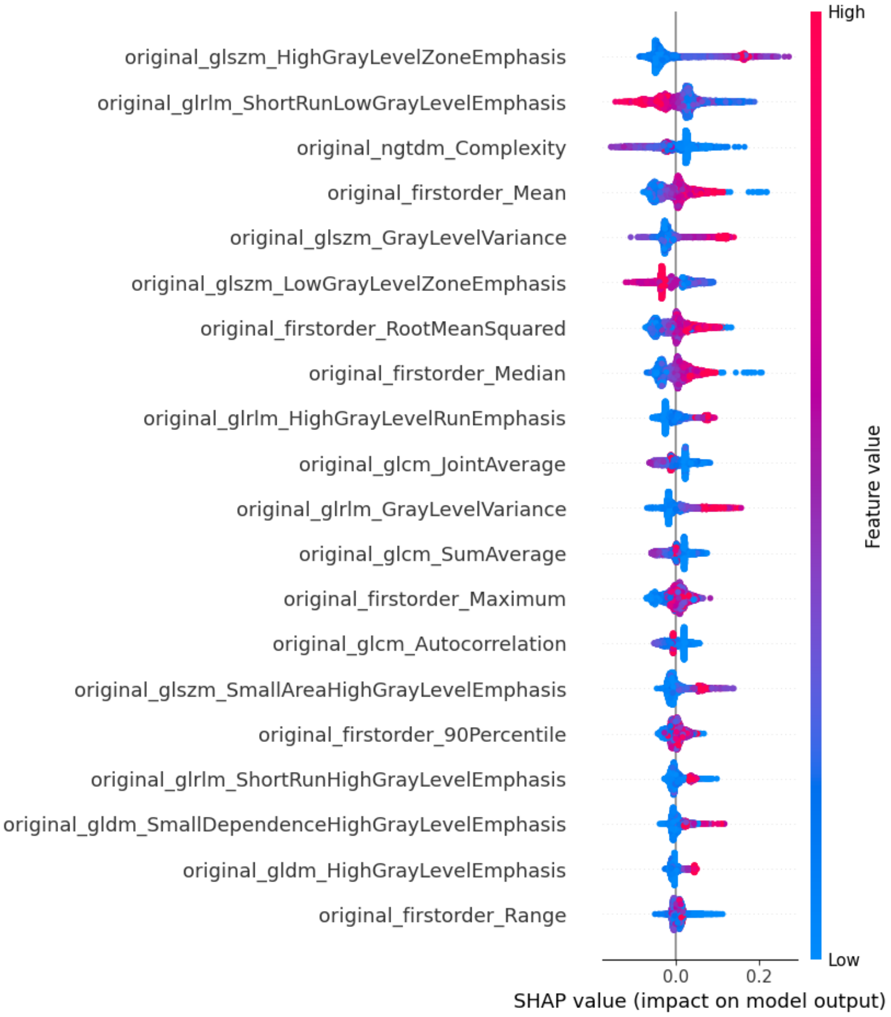


Figure 1: SHAP summary plot for Random Forest classifier showing globally important features.

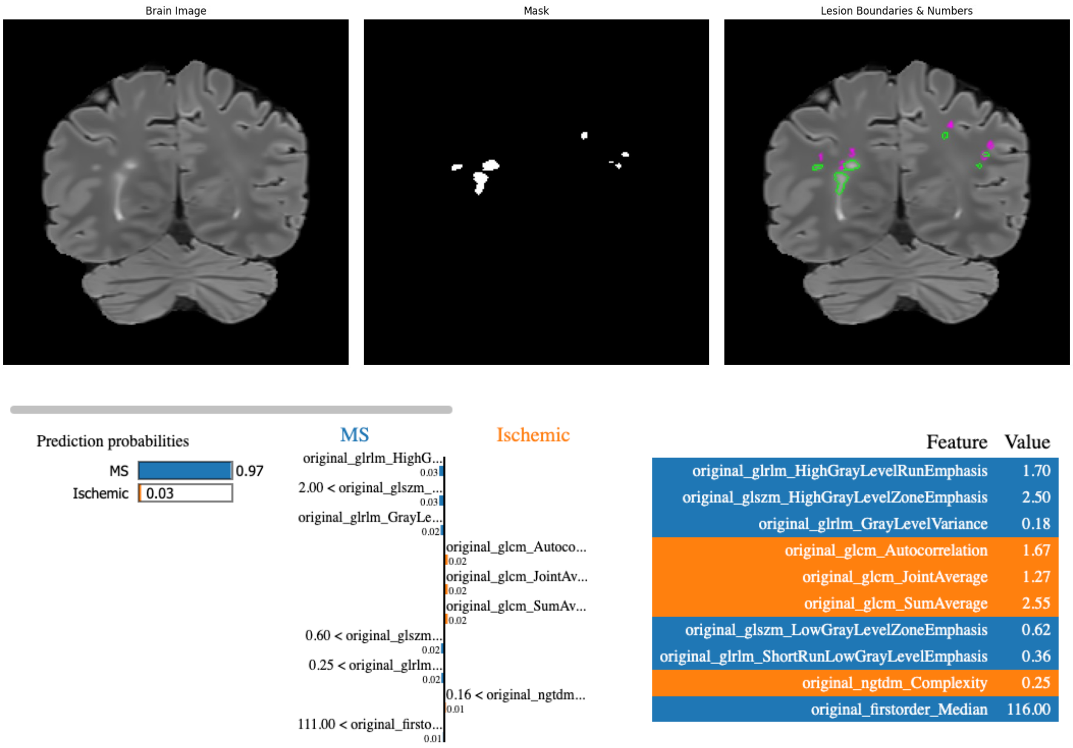
Local Interpretability (LIME + GPT Narratives)

LIME provided per-lesion explanations, and GPT transformed these into clinical narratives. Figure 2 illustrates a case predicted as MS. High uniformity and low contrast were influential, supporting the MS diagnosis. In contrast, Figure 3 shows a predicted ischemic lesion. High skewness and entropy supported the prediction, indicating structural heterogeneity.

Counterfactual Reasoning with DiCE

To simulate alternative diagnostic scenarios and examine model robustness, we used DiCE to generate counterfactual examples. These identify minimal, actionable changes to radiomic features that would result in a different prediction outcome. This approach enhances transparency by answering the question: "What would need to change for the lesion to be classified differently?"

Figure 4 shows a counterfactual explanation for a lesion originally classified as MS. DiCE suggests that reducing the mean intensity (`original_firstorder_Mean`) from 97.45 to 31.60, and increasing the



Feature	Value
original_glrlm_HighGrayLevelRunEmphasis	1.70
original_glszm_HighGrayLevelZoneEmphasis	2.50
original_glrlm_GrayLevelVariance	0.18
original_glcm_Autocorrelation	1.67
original_glcm_JointAverage	1.27
original_glcm_SumAverage	2.55
original_glszm_LowGrayLevelZoneEmphasis	0.62
original_glrlm_ShortRunLowGrayLevelEmphasis	0.36
original_ngtdm_Complexity	0.25
original_firstorder_Median	116.00

Predicted Diagnosis: MS (Multiple Sclerosis) – *Lesion 3*

Prediction Probability: 97%

Key Features Supporting MS:

•**High Gray Level Run Emphasis (1.70):** Indicates concentrated areas with high signal intensity runs, commonly seen in MS plaques due to demyelination and chronic inflammation.

•**High Gray Level Zone Emphasis (2.50):** Reflects large, homogeneously intense regions, consistent with typical MS lesion appearance in FLAIR MRI.

•**Gray Level Variance (0.18):** Low heterogeneity in gray levels suggests uniform lesion intensity, a hallmark of MS plaques.

•**Low Gray Level Zone Emphasis (0.62):** Reflects reduced presence of darker, ischemia-associated zones, favoring MS.

•**Short Run Low Gray Level Emphasis (0.36):** Indicates less fragmented low-signal areas, again pointing toward uniform MS lesions.

•**First-order Median Intensity (116.00):** Falls within the expected range for MS lesions, indicating moderately hyperintense signal.

Why Ischemic Stroke is Unlikely:

•**Autocorrelation (1.67):** Below typical ischemic levels; ischemic lesions often show higher internal texture correlation.

•**Joint Average (1.27) & Sum Average (2.55):** These GLCM features are lower than typical ischemic values, indicating less internal gray-level co-occurrence and spread.

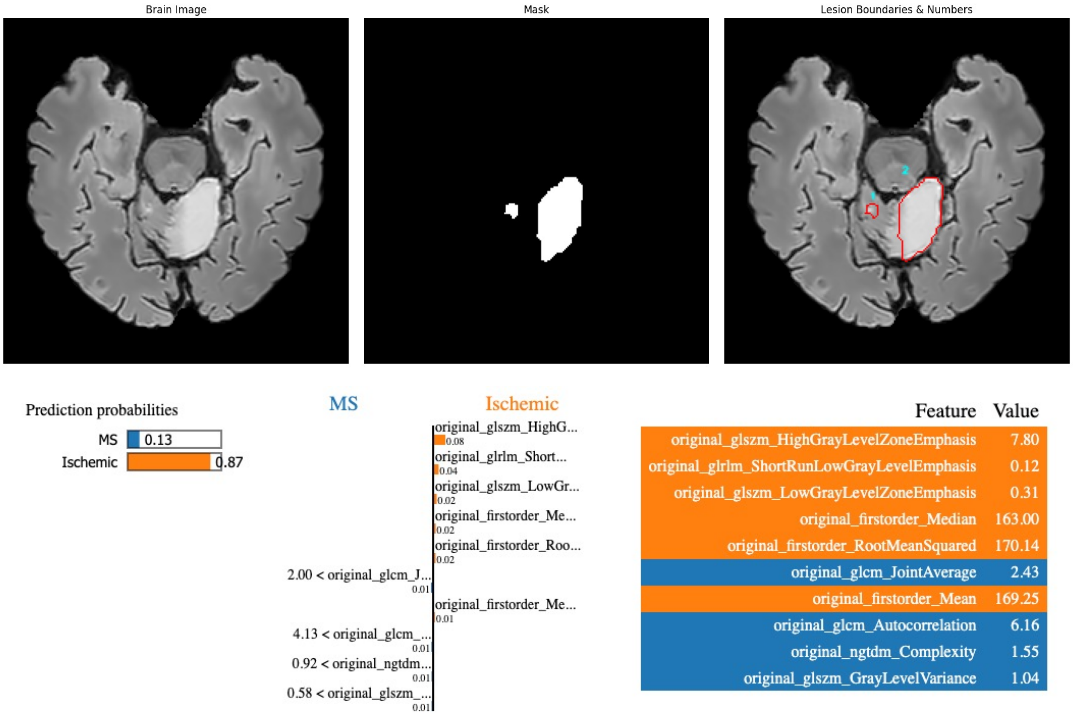
•**NGTDM Complexity (0.25):** Much lower than expected in ischemic strokes, which usually present with more complex and heterogeneous textures.

Conclusion:

Lesion 3 is classified as **MS** primarily due to its **homogeneous intensity, lack of low-gray regions, and texture simplicity**, which are characteristic of chronic demyelinating lesions rather than ischemic damage.

Figure 2: LIME explanation and GPT narrative for an MS lesion. Radiomic features suggest homogeneity typical of demyelinating plaques.

`original_glszm_SmallAreaHighGrayLevelEmphasis` from 0.14 to 33.46, would be sufficient to flip the prediction to ischemic. These feature adjustments reflect plausible variations in lesion brightness and structural homogeneity that align with known imaging patterns of ischemic pathology. This layer of "what-if" analysis enables clinicians to explore how small radiomic shifts could affect classification outcomes, making the model more interpretable and clinically interactive. Although DiCE successfully generated plausible counterfactual examples, interpreting the proposed feature modifications requires clinical expertise, as some radiomic changes may not correspond directly to observable anatomical changes.



Feature	Value
original_glszm_HighGrayLevelZoneEmphasis	7.80
original_glrlm_ShortRunLowGrayLevelEmphasis	0.12
original_glszm_LowGrayLevelZoneEmphasis	0.31
original_firstorder_Median	163.00
original_firstorder_RootMeanSquared	170.14
original_glcm_JointAverage	2.43
original_firstorder_Mean	169.25
original_glcm_Autocorrelation	6.16
original_ngtdm_Complexity	1.55
original_glszm_GrayLevelVariance	1.04

Predicted Diagnosis: Ischemic Stroke – *Lesion 1*

Prediction Probability: 87%

Key Features Supporting Ischemic Stroke:

•**High Gray Level Zone Emphasis (7.80):** Indicates large, bright uniform regions—typical of acute/subacute infarcts with well-demarcated necrotic tissue.

•**Short Run Low Gray Level Emphasis (0.12):** Suggests a lack of short, low-intensity textures—supporting ischemia, where lesions tend to be sharply defined and not mottled.

•**Low Gray Level Zone Emphasis (0.31):** Minimal presence of dark areas aligns with the uniform brightness of ischemic lesions.

•**First-order Median (163.00), Root Mean Squared (170.14), and Mean Intensity (169.25):** All indicate high signal intensity, characteristic of ischemic tissue with edema or gliosis in FLAIR MRI.

Why MS is Unlikely:

•**Joint Average (2.43):** While elevated, it's less distinctive for MS compared to more specific texture patterns.

•**Autocorrelation (6.15):** Although relatively high, it does not offset the dominant ischemic characteristics.

•**NGTDM Complexity (1.55):** Moderate complexity, but not high enough to suggest the varied texture seen in MS plaques.

•**Gray Level Variance (1.04):** Suggests some heterogeneity but insufficient to match the typical diverse texture of MS lesions.

Conclusion:

Lesion 1 is classified as **Ischemic Stroke** due to its **high uniform brightness, low texture fragmentation, and elevated mean intensities**, all of which are classic features of infarcts rather than demyelinating MS lesions.

Figure 3: LIME explanation and GPT narrative for an ischemic lesion, emphasizing irregular textural features.

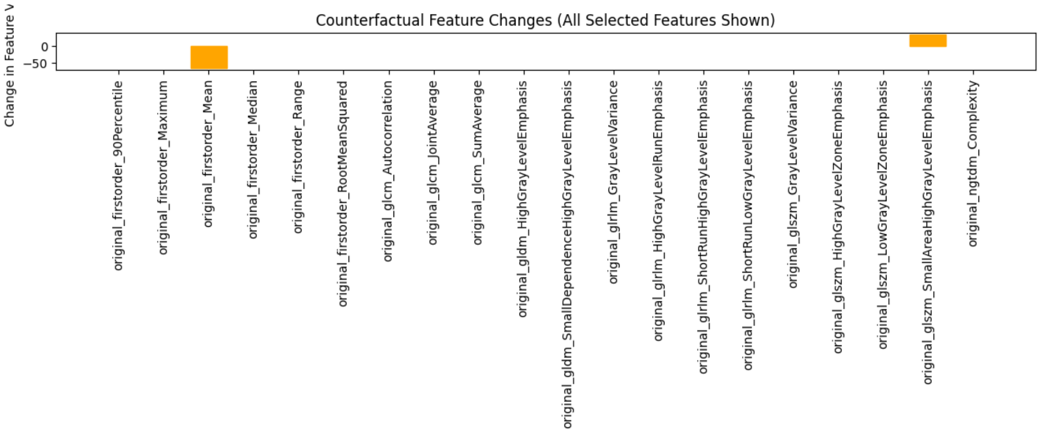


Figure 4: Counterfactual explanation generated by DiCE, showing feature changes required to flip prediction from MS to ischemic.

Discussion

The combination of radiomics and multi-level XAI revealed consistent and interpretable patterns in lesion classification. SHAP highlighted globally dominant features, while LIME showed per-lesion factors contributing to predictions. DiCE further provided hypothetical scenarios for counter-diagnosis, and GPT-based narratives completed the pipeline by offering concise, clinician-aligned explanations. Together, these methods support not only technical validation but also clinical usability, making the system suitable for diagnostic support in real-world settings. Each explanation modality brings distinct strengths and limitations. SHAP provides both global and local feature attributions but can be computationally intensive. LIME generates interpretable local explanations but may suffer from instability across small perturbations. DiCE enables actionable counterfactuals but sometimes proposes feature changes that may not correspond directly to plausible anatomical variations. GPT narratives offer intuitive, clinician-friendly summaries, but they are prone to occasional hallucinations, especially in ambiguous cases. Understanding these trade-offs is crucial for practical clinical adoption.

To assess the clinical relevance of the generated explanations, we shared the model outputs and narratives with a practicing radiologist. The radiologist noted that the explainability provided was very helpful and that the overall analysis made sense. However, broader evaluation is needed. Future work will involve multi-expert validation with inter-rater agreement scoring and quantitative assessment of explanation fidelity to strengthen the clinical robustness of the framework. Additionally, integrating counterfactual outputs from DiCE into GPT-based narrative generation presents an exciting opportunity to create "what-if" clinical explanations that can further enhance the usability of the system.

4. Conclusion

This study presents a radiomics-based classification framework for differentiating multiple sclerosis and ischemic stroke lesions on MRI, enhanced with a multi-level explainable AI pipeline. Among the evaluated classifiers, Random Forest achieved the best overall performance, with an accuracy of 91.2% and an F1-score of 86.5%. Beyond predictive performance, the framework integrates global (SHAP), local (LIME), counterfactual (DiCE), and natural language (GPT) explanations to provide transparent and clinically meaningful insights. SHAP identified texture-based radiomic features as globally influential, while LIME and GPT enabled per-lesion interpretability in clinician-friendly language. DiCE offered hypothetical reasoning to explore how small changes in feature values could lead to different diagnoses. Together, these methods create a robust and interpretable decision support system that bridges technical and clinical domains. Future work will focus on experimenting with different radiomic extraction parameters, such as varying angles, distances, and bin widths, to explore their impact on classification and explainability. We also plan to expand this framework to include multi-modal MRI inputs, increase sample diversity for better generalization, and integrate visual explanation methods, such as attention-based heatmaps or Grad-CAM visualizations, to enhance clinical trustworthiness. Finally, fine-tuning the narrative generation process using domain-specific language models will be explored. By combining radiomics with layered explainability, this approach supports the safer, more transparent deployment of AI tools in neuroimaging diagnostics.

Acknowledgments

This research is funded by the SFI Center for Research Training in Machine Learning (18/CRT/6183).

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o and Grammarly for grammar and spelling checks. The author(s) reviewed and edited the output as necessary and assume full responsibility for the content of the publication.

References

[1] P. Wildner, M. Stasiołek, M. Matysiak, Differential diagnosis of multiple sclerosis and other inflammatory cns diseases, Multiple sclerosis and related disorders 37 (2020) 101452.

[2] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, et al., Radiomics: extracting more information from medical images using advanced feature analysis, European journal of cancer 48 (2012) 441–446.

[3] S. Lundberg, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).

[4] M. T. Ribeiro, S. Singh, C. Guestrin, 'why should i trust you?' explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[5] F. Eitel, E. Soehler, J. Bellmann-Strobl, A. U. Brandt, K. Ruprecht, R. M. Giess, J. Kuchling, S. Asseyer, M. Weygandt, J.-D. Haynes, et al., Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation, NeuroImage: Clinical 24 (2019) 102003.

[6] S. Basu, A. Munafo, A.-F. Ben-Amor, S. Roy, P. Girard, N. Terranova, Predicting disease activity in patients with multiple sclerosis: An explainable machine-learning approach in the mavenclad trials, CPT: Pharmacometrics & Systems Pharmacology 11 (2022) 843–853.

[7] A. Lopatina, S. Ropele, R. Sibgatulin, J. R. Reichenbach, D. Güllmar, Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis, Frontiers in neuroscience 14 (2020) 609468.

[8] M. Leite, L. Rittner, S. Appenzeller, H. H. Ruocco, R. Lotufo, Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging, Journal of Medical Imaging 2 (2015) 014002–014002.

[9] D. P. Castillo, R. J. Samaniego, Y. Jiménez, L. A. Cuenca, O. A. Vivanco, J. M. Álvarez-Gómez, M. J. Rodriguez-Alvarez, Identifying demyelinating and ischemia brain diseases through magnetic resonance images processing, in: 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), IEEE, 2019, pp. 1–3.

[10] D. Vuong, S. Tanadini-Lang, Z. Wu, R. Marks, J. Unkelbach, S. Hillinger, E. I. Eboulet, S. Thierstein, S. Peters, M. Pless, et al., Radiomics feature activation maps as a new tool for signature interpretability, Frontiers in oncology 10 (2020) 578895.

[11] M. R. Hernandez Petzsche, E. de la Rosa, U. Hanning, R. Wiest, W. Valenzuela, M. Reyes, M. Meyer, S.-L. Liew, F. Kofler, I. Ezhov, et al., Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset, Scientific data 9 (2022) 762.

[12] O. Commowick, M. Kain, R. Casey, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, et al., Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset, Neuroimage 244 (2021) 118589.

[13] A. Buades, B. Coll, J.-M. Morel, Non-local means denoising, Image Processing On Line 1 (2011) 208–212.

[14] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer research 77 (2017) e104–e107.

[15] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 607–617.