BMC Medical Imaging

RESEARCH

Open Access

Assessing AI-augmented training for multiple sclerosis classification in a basal ganglia radiomics model

Erick Eduardo López-Ríos[1] and Francisco J. Alvarez-Padilla[1*]

Abstract

Multiple sclerosis (MS) radiomics is hindered by multicenter variability and limited sample sizes. We evaluated whether GAN-based augmentation (GBA) improves MS classification versus traditional data augmentation (TDA) under center-wise external testing. A conditional GAN generated T1-weighted brain MRIs conditioned on class labels. Ten subcortical regions (including thalamus, putamen, caudate) were segmented with a 3D U-Net; radiomic features (shape, first-order, and texture families) were extracted and selected with LASSO. We used a leave-one-center-out (LOCO) design. All model development, segmentation, cGAN training, feature engineering, and tuning, were performed within the training centers only using inner 5-fold (subject-level 80/20) splits; the entire held-out center was reserved for a single external test. Across centers, GBA yielded small but consistent gains over TDA and real-only training, most evident for a tabular ResNet (average F1 up to 0.957), while confidence intervals overlapped for some metrics. SHAP analyses preserved the salience of basal-ganglia features, supporting biological plausibility. Limitations include a single-country cohort and no public external validation, which constrains generalizability. AI-augmented training provides incremental improvements for MS radiomics under site-held-out testing and motivates broader, international validation and clinically oriented utility analyses.

Keywords Generative AI, Feature engineering, Computer-aided diagnosis, Multiple sclerosis

Introduction

Multiple sclerosis (MS) is a chronic, debilitating neurological disease of the central nervous system [1–3] for which magnetic resonance imaging (MRI) is indispensable for diagnosis, monitoring disease progression, and evaluating therapeutic response [4, 5]. While white matter lesions have historically been the primary focus, subcortical gray matter structures, such as the basal ganglia, are increasingly recognized for their role in MS pathogenesis and their association with clinical manifestations ranging from cognitive impairment to fatigue [6–11].

Quantitative imaging techniques like radiomics, which extract high-dimensional handcrafted features describing tissue shape, intensity, and texture, offer a powerful framework for uncovering subtle pathological changes beyond conventional volumetry. In MS, radiomics has shown considerable promise [12–14]. Studies have successfully employed radiomic classifiers to distinguish active from inactive lesions, predict near-term disease progression, and aid in the differential diagnosis between MS and its mimics, such as neuromyelitis optica spectrum disorder (NMOSD) and myelin oligodendrocyte glycoprotein antibody-associated disease (MOGAD)

*Correspondence:
Francisco J. Alvarez-Padilla
francisco.alvarez@academicos.udg.mx
[1]Translational Bioengineering, University of Guadalajara, 1421 Blvd. Marcelino García Barragán, Guadalajara 44430, México

[15]. These efforts highlight the potential of radiomics to support clinical decision-making, potentially even reducing reliance on contrast-based imaging.

Despite these advances, the clinical translation of radiomic models is hampered by two significant and related challenges. First, multicenter heterogeneity, arising from variations in scanners, acquisition protocols, and post-processing steps, can degrade feature stability and limit the generalizability of learned models [16, 17]. Second, limited cohort sizes at individual centers increase the risk of overfitting, where a model learns site-specific noise rather than a true, transportable biological signal [17]. These realities underscore the critical need for robust validation strategies and data-centric approaches that can improve model generalization.

We hypothesize that synthetic data augmentation can help regularize radiomic pipelines by enriching the training distribution with plausible, yet diverse, examples. Generative adversarial networks (GANs) can produce high-fidelity MR images [18–20] that capture complex anatomical and textural patterns, offering a more sophisticated alternative to traditional data augmentation (TDA) techniques like rotation or scaling, which may alter high-order textures in non-biological ways [21]. However, any benefit derived from GAN-based augmentation (GBA) must be rigorously validated under a strict site-held-out evaluation to ensure that performance gains are genuine and not merely an artifact of data leakage or overfitting to a specific training set [22].

In this study, we investigate whether a radiomics pipeline trained with GBA improves classification performance for MS in the basal ganglia compared to TDA and real-data-only approaches. We employ a stringent leave-one-center-out (LOCO) cross-validation framework to provide an unbiased estimate of real-world generalization. Our primary contributions are fourfold: (1) we establish a rigorous external validation design where development, including GAN training, feature engineering, and hyperparameter tuning, is performed independently of the held-out test center; (2) we conduct fidelity and non-copying checks to ensure the quality and novelty of the synthetic images; (3) we report center-wise and overall performance metrics (AUC, accuracy, sensitivity, and specificity) to quantify the incremental benefit of GBA; and (4) we use interpretability analyses to confirm that GBA preserves the neuroanatomical importance of the caudate, putamen, and thalamus.

By adopting a transparent, leakage-controlled LOCO design, this work aims to provide a robust estimate of the value of synthetic data in multicenter MS radiomics and offer a reproducible foundation for future large-scale validation studies.

Materials and methods

Data description

This study employed a retrospective dataset comprising 368 T1-weighted MRI scans acquired at 3T from three radiology centers in Mexico, reflecting substantial clinical heterogeneity. Scans were obtained using Siemens ($n = 159$), Philips ($n = 117$), and GE ($n = 92$) systems, and included 193 images from 146 patients diagnosed with multiple sclerosis (MS), alongside 175 scans from healthy controls (HC). MS images were distributed as 116, 21, and 56 across Centers A, B, and C, respectively; HC images included 102, 34, and 39 from the same sites.

To mitigate technical variability, all scans underwent standardized preprocessing: skull stripping, reorientation to the RPS coordinate system, and resampling to isotropic $1 \times 1 \times 1$ mm voxel spacing. Volumes were then co-registered to a common template and cropped to a uniform dimension of $144 \times 160 \times 144$ voxels. Scanner-specific batch effects were addressed using the NeuroComBat harmonization algorithm [17] within a cross-validation framework, wherein transformations were trained exclusively on each fold's training set and applied to the corresponding test set.

We adopted a LOCO strategy across three sites. Given the strong class and sample imbalance, we performed B-out and C-out experiments (Center B held-out; Center C held-out). In each LOCO iteration, only the two remaining centers were used for model development. Within these training centers, we ran 5-fold cross-validation with distinct random seeds using subject-level splits (80/20) to tune hyperparameters and apply early stopping. All data from the held-out center remained unseen until the final evaluation of that experiment. For subjects with longitudinal scans, all visits were kept in the same split; if a subject had scans from multiple centers and one belonged to the held-out site for that iteration, that scan was excluded to avoid leakage. This strategy was employed for the training and evaluation of all the fitted models.

Segmentation atlas for radiomic extraction

We constructed a segmentation atlas to delineate the regions of interest (ROIs) for radiomic analysis, including bilateral thalamus, putamen, globus pallidus, caudate nucleus, as well as whole white and gray matter. All ROIs were initially segmented using a hybrid pipeline that combines a region-growing algorithm with a geodesic distance-based support vector machine (SVM) classifier [23], yielding candidate masks. Neuroanatomists subsequently performed manual refinements in 3D-Slicer [24]. Inter-rater reliability, assessed on a subset of 20 volumes, yielded a Dice similarity coefficient of 0.92 ± 0.03 for subcortical nuclei and 0.95 ± 0.02 for white/gray matter. An example of the resulting segmentations is shown in Fig. 1.

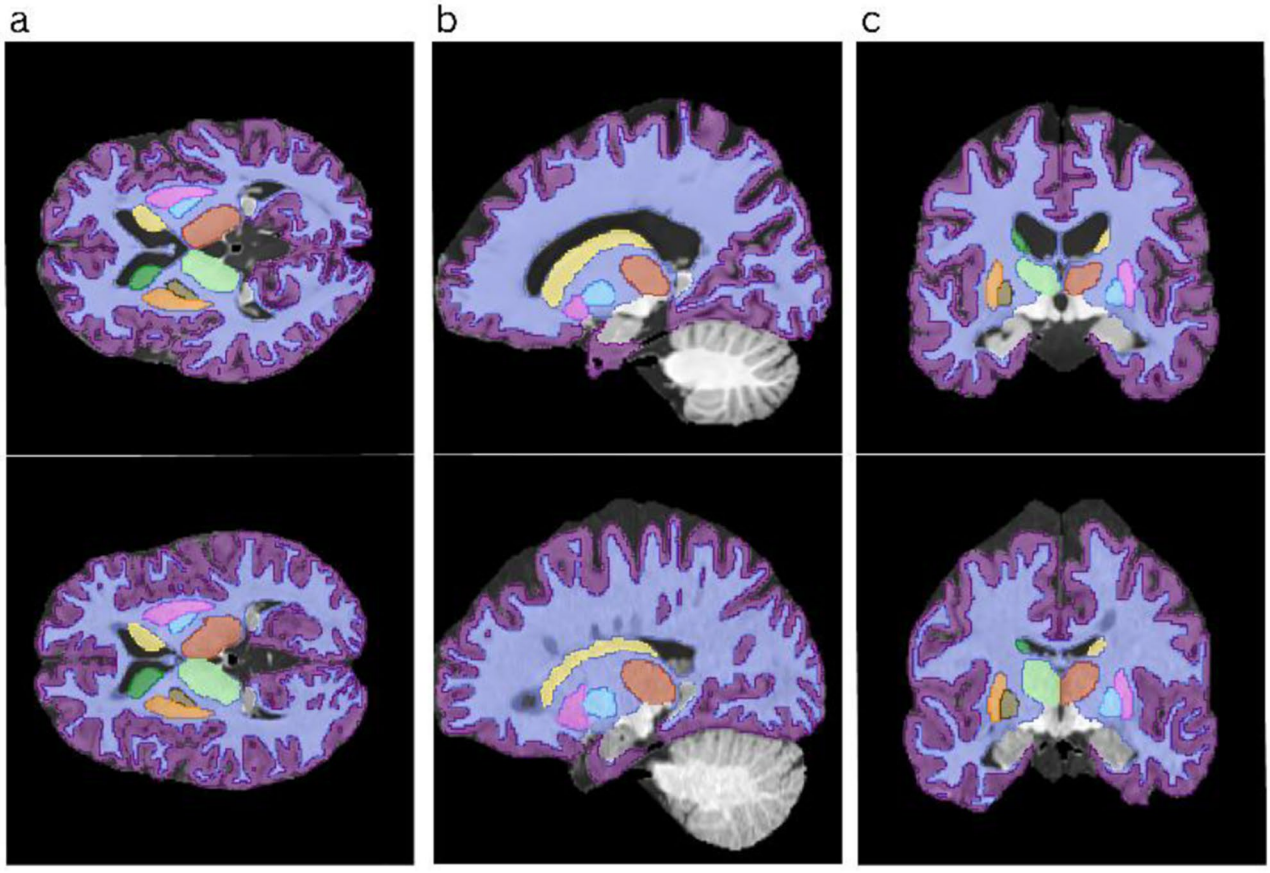

Fig. 1 Representation of segmentation in MRI slices. Axial (**a**), sagittal (**b**), and coronal (**c**) views. The top row exhibits generated images, while the bottom row presents real images. Segmented structures are color-coded, including white matter (blue), gray matter (purple), left and right caudate nuclei (green and yellow), left and right putamen (orange and pink), left and right globus pallidus (brown and cyan), and left and right thalamus (lime green and red)

To enable a scalable and automated pipeline for both real and synthetic data, a 3D U-Net model [25] was trained for the segmentation task. Crucially, to ensure its robustness, the U-Net was trained and validated within our training strategy. For each LOCO iteration, a 3D U-Net was trained only in the training centers and monitored on the inner validation partitions (80/20 within the training centers). The resulting model was then applied to the held-out center for segmentation prior to feature extraction, ensuring no information from the held-out site influenced training. The model's performance was then evaluated on the hold-out test set of real images for that fold, achieving a mean Dice similarity coefficient of 0.91 ± 0.04 across all ROIs.

The U-Net was then used to segment the GBA images. While direct validation on synthetic data is inherently challenging due to the absence of ground truth, this approach assumes that a segmentation model proven to be robust on real data can reliably delineate the anatomically plausible structures produced by the cGAN. The automated step was essential for extracting radiomic features from the augmented datasets in a consistent manner. To assess segmentation reliability, we computed the Dice coefficient (0.83 ± 0.03) and Jaccard Index (0.79 ± 0.04) on a representative subset of synthetic scans. Subsequently, all segmentations were reviewed and manually corrected by clinical experts.

Traditional data augmentation

Data augmentation was applied at multiple stages to enhance model robustness. To benchmark the performance of our GBA, we first implemented a conventional TDA pipeline for comparative evaluation. This allowed us to assess whether synthetic data generated by the cGAN offers measurable advantages over TDA in the final classification task.

The TDA pipeline operated on real training samples within each cross-validation fold, preserving the independence between training and testing sets throughout the evaluation. Transformations included spatial perturbations such as random horizontal flipping (mirroring along the sagittal plane), affine adjustments, controlled rotations (± 10°), pixel-level translations (up to 10 pixels), and isotropic scaling (90–110%), designed to simulate

plausible anatomical variations while maintaining structural fidelity.

Image generation

To explore synthetic data augmentation, we implemented a cGAN to generate realistic MRI scans for both MS patients and HC. The generator adopted a 3D U-Net architecture [25], while a PatchGAN discriminator [26] was used, as depicted in Fig. 2. Training employed a composite loss function combining adversarial loss with pixel-wise L1 reconstruction. The latter encouraged structural fidelity between synthetic and real images. To enhance stability and prevent discriminator overfitting due to limited data, we applied differentiable augmentation, introducing minor random geometric perturbations to both real and generated scans immediately prior to discriminator input, thereby fostering a balanced learning dynamic.

In each LOCO experiment, evaluation used a fixed hold-out test set comprising all cases from the held-out center, which remained entirely unseen throughout model development. Within the two training centers, we conducted 5-fold cross-validation with subject-level 80/20 splits to guide early stopping and stabilize hyperparameters. For each fold, a distinct cGAN was trained from scratch for up to 300 epochs, with early stopping after 10 consecutive epochs without validation improvement; training used only the fold-specific training data, fully isolating both the fold's validation split and the held-out center. After convergence, the cGAN outputs were assessed with SSIM, PSNR, FID, and KID. The validated generator then produced 100 synthetic images (50 per class), which were added solely to the training folds within that LOCO; the test set consisted exclusively of the held-out center.

Radiomic features

A comprehensive set of 1070 radiomic features (107 per structure) per subject was used for the analysis, comprising shape-based features, first-order statistics, and higher-order texture metrics (GLCM, GLDM, GLRLM, GLSZM, and NGTDM) extracted using the PyRadiomics library [27] in adherence with IBSI guidelines [28]. To assess the robustness of these features to scaling procedures, we conducted a comparative analysis across three normalization methods, RobustScaler, MinMax, and Z-Score, applied independently prior to dimensionality reduction. Feature distributions and clustering patterns were evaluated post-normalization to identify potential distortions or shifts in intergroup separability (see Figure S1).

To identify the most predictive features while mitigating multicollinearity and overfitting, we implemented LASSO (Least Absolute Shrinkage and Selection Operator) regression with an L1-penalized logistic model. In each LOCO experiment, the fixed hold-out test set consisted exclusively of all cases from the held-out center, which remained entirely unseen during feature extraction, normalization, feature selection, hyperparameter tuning, and model fitting. Radiomic features were first extracted per ROI only from the training centers of that LOCO fold. LASSO feature selection was then performed within the inner 5-fold cross-validation on the training centers. Feature selection was performed exclusively within the inner CV of the training centers. The resulting feature list was then applied to train and evaluate models for that LOCO iteration; the held-out center was never used for selection or tuning. For synthetic data experiments, the base training set (real data from the training centers) was augmented with two scenarios: TDA and GBA (using the corresponding cGAN). For each scenario, five distinct folds of synthetic data were generated. This design enabled assessment of model performance and stability across varying synthetic distributions while

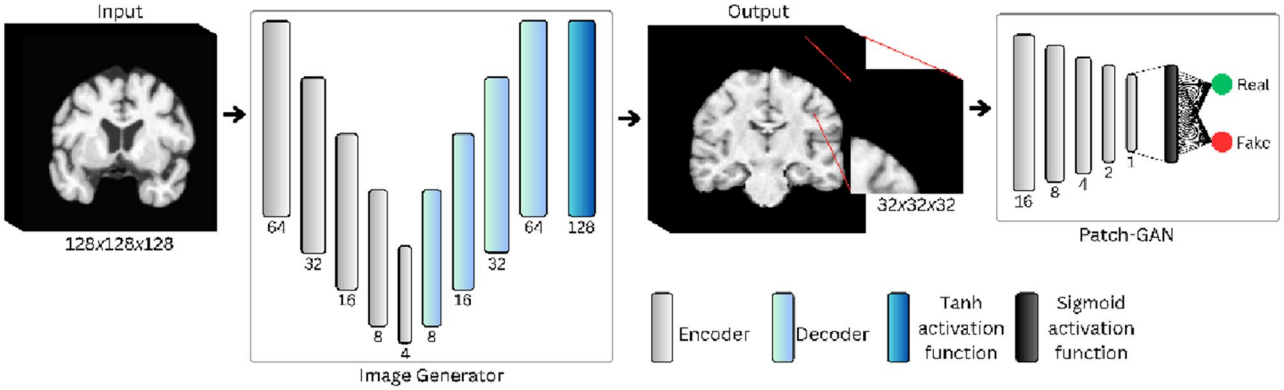

Fig. 2 Network architecture. The generator encoding path used a stride of 3, while the decoding path employed a stride of 2. All LeakyReLU activation layers maintained a negative slope of 0.01

preserving a strictly independent held-out center for final evaluation.

Classification algorithms

To evaluate the classification performance of the selected radiomic features, we trained and compared three models: a Support Vector Machine (SVM) [29], a Random Forest (RF) [30], and a custom Residual Neural Network adapted for tabular data (ResNet) [31]. For each LOCO experiment, the fixed test set consisted exclusively of all cases from the held-out center, kept completely isolated from feature extraction, hyperparameter tuning, and model training. The training set comprised only the two remaining centers and was augmented under two scenarios: (1) TDA and (2) GBA. For each scenario, five distinct folds of synthetic data were generated, and model training was repeated five times by combining the real training set with each synthetic fold. Hyperparameters for all classifiers were optimized within the inner 5-fold cross-validation on the training centers. This procedure identifies the optimal hyperparameters reported in Table S1. Final performance was evaluated once on the held-out center for each repetition, and we report per-center metrics as well as the mean ± SD across the five repetitions within each LOCO.

Model performance was quantified using accuracy, F1-score, precision, recall, and the area under the receiver operating characteristic curve (AUC). To compare classifiers, a non-parametric Friedman test was applied to the F1-scores across repetitions within each augmentation scenario. When significant differences were detected ($p < 0.05$), post-hoc pairwise Wilcoxon signed-rank tests were conducted, with p-values adjusted for multiple comparisons using the Holm-Bonferroni correction.

Model interpretability

To elucidate the decision-making process and identify the most influential radiomic biomarkers, we conducted an interpretability analysis using SHAP (SHapley Additive exPlanations) [32]. This analysis was specifically applied to the pipeline that demonstrated superior performance, trained under the Real + TDA and Real + GBA.

The SHAP analysis was configured to interpret the model's outputs in terms of probabilities by passing its prediction probability function to the explainer. The transformed training data was provided as the background dataset to establish a baseline for calculating the SHAP values, which represents the expected model output. Our interpretability focus centered exclusively on the SHAP values associated with the positive class (MS), enabling a targeted evaluation of how individual radiomic features influenced predictions related to disease presence. This approach ensured that the attribution scores reflected relevant diagnostic signals rather than global model behavior across classes.

Table 1 Evaluation of the quality of images generated by cGAN (mean ± SD across the 5 inner folds within the training centers)

Metric	Average Value	Interpretation
FID	18 ± 3	Lower values demonstrate better results
KID	0.042 ± 0.011	Lower values demonstrate better results
SSIM	0.904 ± 0.028	Higher values demonstrate better results
PSNR	38.9 ± 0.43db	Higher values demonstrate better results

Clinical utility assessment

To translate model performance into clinical relevance, we conducted a formal utility analysis on the top-performing pipeline. First, to avoid test-set optimization, the decision threshold was fixed a priori by selecting the operating point that maximized Youden's J index (Sensitivity + Specificity – 1) on the inner validation folds of the training centers. This single, pre-specified threshold was then applied to the held-out center data for final evaluation.

At this fixed operating point, we computed key diagnostic metrics, including sensitivity, specificity, positive and negative predictive values (PPV/NPV), with exact 95% confidence intervals (Clopper-Pearson). To contextualize the impact on clinical decision-making, we projected post-test probabilities across plausible MS prevalences (10%, 30%, 50%) using the derived likelihood ratios.

Results

Image generation results

Within each LOCO iteration, the quality and fidelity of GBA images were evaluated within the inner 5-fold cross-validation on the training centers. Results are reported as mean ± SD across the five inner folds (training centers only), as summarized in Table 1. Perception-based metrics, such as Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), indicate a high similarity between the feature distributions of real and synthetic images. Image similarity-based metrics, such as SSIM and PSNR, confirm high structural and intensity correspondence at the pixel level. Figure 3 presents examples of the generated images compared to real ones and transformed with data augmentation samples.

To formally evaluate whether the cGAN was generating replicas of original data, we conducted a nearest-neighbor distance analysis. We computed the cosine similarity between each synthetic sample and all real training scans within the same cross-validation fold. Across all inner validation folds within the training centers, synthetic–real distances were significantly greater ($p < 0.001$) than typical intra-cohort distances, indicating that the generative model does not simply reproduce training examples (see Fig. 4).

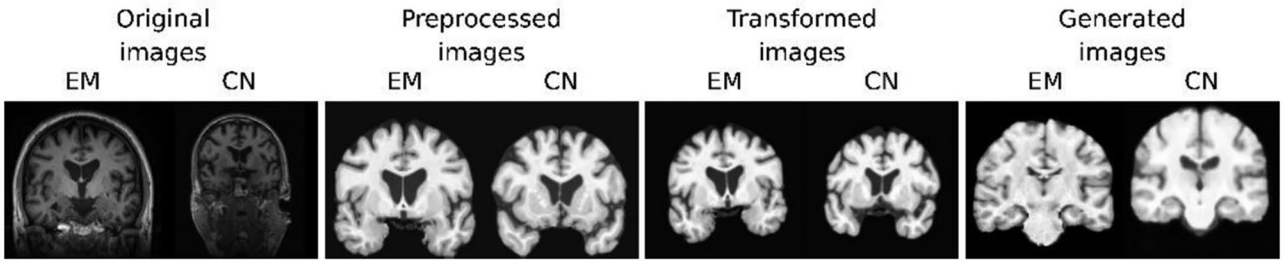

Fig. 3 Visual comparison of T1-weighted MRIs from MS and HC subjects across four stages: original, preprocessed, transformed, and synthetic. The first three columns correspond to the same subject, while the final synthetic samples do not represent the same individuals

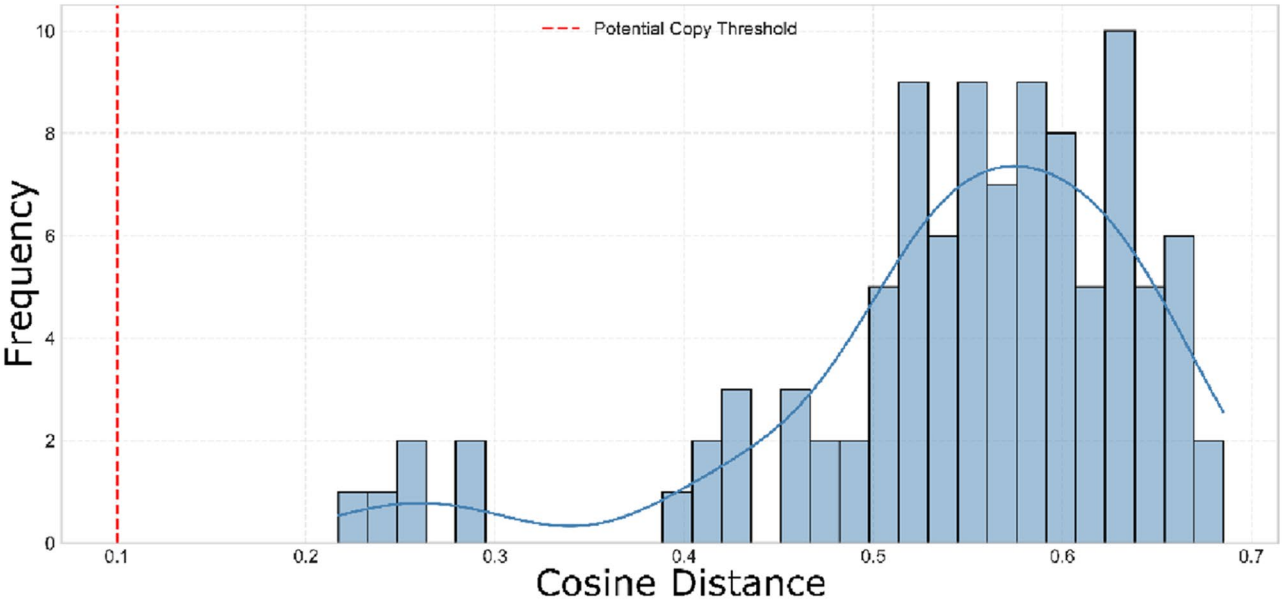

Fig. 4 Bar plots show the average cosine distance between real and GBA images across five cross-validation inner folds. Lower bars indicate image similarity, reflecting the risk of the GAN in replicating the original cohort

Evaluation of synthetic data fidelity

The fidelity of the synthetic data was examined via UMAP projection (Fig. 5), revealing a central cluster in the real data with notable overlap between the TDA and GBA samples. Notably, TDA tended to concentrate in the dense core of this manifold, while GBA showed slightly broader dispersion. The average energy distance to the real data was 0.1974 for TDA images and 0.2348 for GBA ones. Clustering analysis further assessed structure preservation: K-Means on real data yielded an ARI of 0.1552, slightly reduced to 0.1481 with TDA, and slightly decreased to 0.1370 when including GBA data, indicating that class distinction was retained, albeit with minor degradation (see Table 2).

To complement this assessment, a Kolmogorov–Smirnov (KS) test was conducted on each radiomic feature to quantify distributional shifts between augmented datasets. Most features exhibited non-significant deviations ($p > 0.05$), suggesting that both TDA and GBA preserved the statistical integrity of individual radiomic descriptors. Aggregated KS statistics revealed measurable differences in feature distributions: the average KS divergence was 0.1442 ± 0.0019 for TDA and 0.2244 ± 0.0021 for GBA when compared to the original cohort. While both strategies maintain radiomic coherence, TDA introduces less perturbation at the feature level.

Classification performance

Three experiments were conducted to evaluate classification performance. Experiment 1 used radiomic features from the real set; Experiment 2 combined real with TDA-derived features; and Experiment 3 merged real with GBA-derived features. In every case, only the features LASSO selected within each inner fold on the training centers were employed.

Table 3 presents the overall external performance under the LOCO scheme, aggregating B-out and C-out as Overall. Model development used inner 5-fold splits within the training centers only, and external testing was performed once per LOCO on the held-out center.

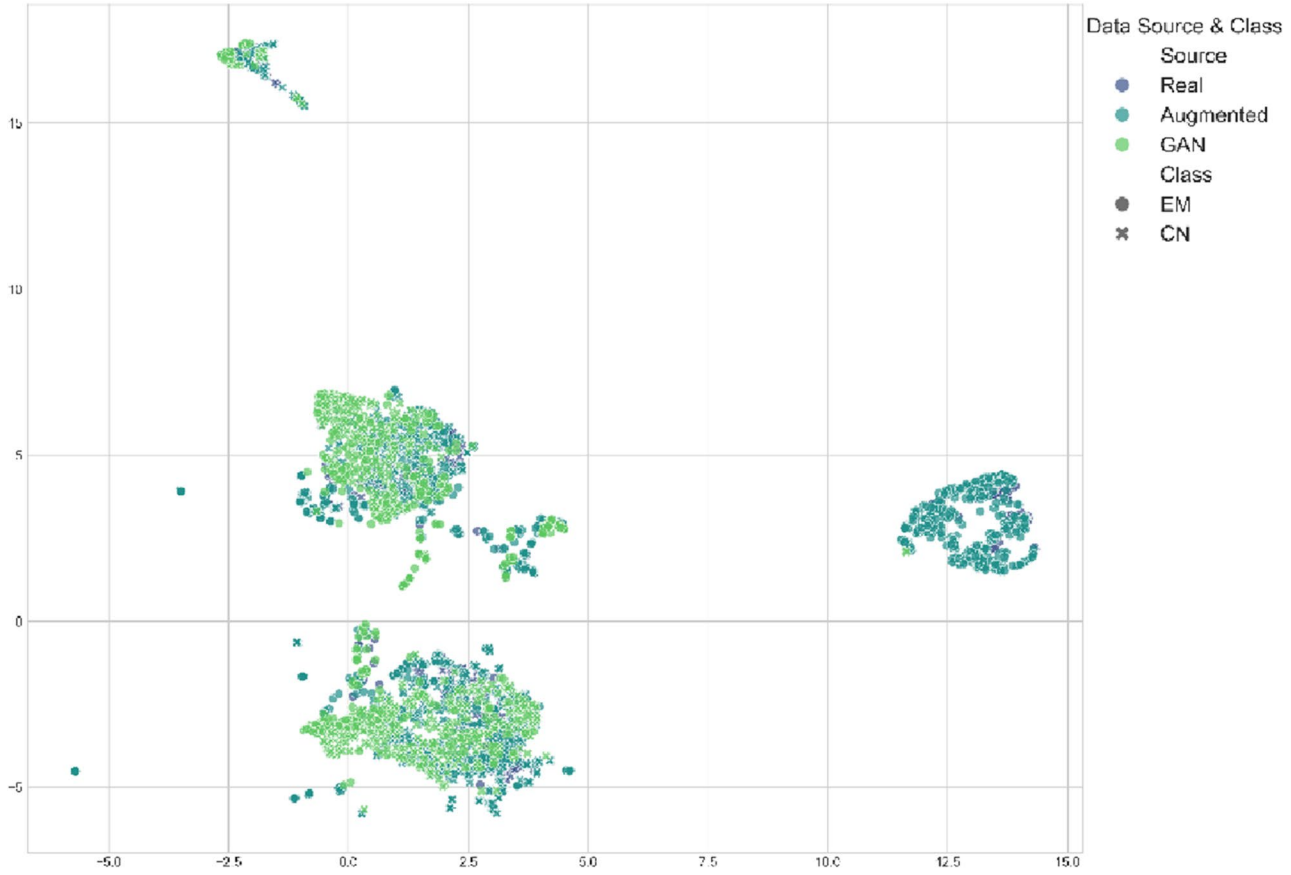

Fig. 5 UMAP projection of radiomic feature distributions. A two-dimensional representation of the high-dimensional feature space for real, TDA, and GBA data. Each point corresponds to a single subject scan. Colors denote the data source, while marker styles distinguish between the two clinical conditions

Table 2 Quantitative metrics for synthetic data fidelity

Dataset	Average energy distance (compared to real)	Average ARI
Real	—	0.1552
TDA	0.1974 ± 0.0031	0.1481 ± 0.0116
GBA	0.2248 ± 0.0024	0.1370 ± 0.0007

Overall, both augmentation methods (TDA and GBA) improved performance compared to training with real data alone. ResNet emerged as the top-performing model, achieving the highest average F1-score of 0.957 ± 0.018 in this scenario.

A formal statistical analysis was conducted on the external test predictions from the held-out center. The Friedman test revealed no statistically significant difference in the F1-scores among the five models ($p = 0.112$),

Table 3 Overall performance under the LOCO strategy

Classifier	Dataset	F1-Score	Accuracy	Recall	AUC
SVM	Real	0.899 ± 0.02	0.899 ± 0.02	0.855 ± 0.02	0.978 ± 0.01
SVM	+ GBA	0.943 ± 0.0	0.939 ± 0.01	0.975 ± 0.02	0.991 ± 0.00
SVM	+ TDA	0.938 ± 0.02	0.932 ± 0.02	**0.983 ± 0.01**	**0.994 ± 0.00**
RF	Real	0.919 ± 0.01	0.916 ± 0.02	0.907 ± 0.04	0.979 ± 0.01
RF	+ GBA	0.926 ± 0.01	0.923 ± 0.01	0.921 ± 0.01	0.983 ± 0.00
RF	+ TDA	0.936 ± 0.01	0.932 ± 0.01	0.954 ± 0.01	0.989 ± 0.00
ResNet	Real	0.920 ± 0.04	0.921 ± 0.04	0.881 ± 0.08	0.985 ± 0.01
ResNet	+ GBA	**0.957 ± 0.01**	**0.954 ± 0.01**	0.975 ± 0.02	0.991 ± 0.00
ResNet	+ TDA	0.947 ± 0.01	0.943 ± 0.01	0.975 ± 0.02	0.992 ± 0.00

Models were trained on the remaining centers using inner 5-fold cross-validation. Results are reported as the mean ± SD across two experiments: leaving out center B and leaving out center C

suggesting that while there are minor variations in their average performance, all tested pipelines perform comparably when trained with both augmentation techniques. For further interpretability, the confusion matrix for the top-performing model is presented in Supplementary Figure S2.

Model interpretability

To elucidate the decision logic of our ResNet classifier, we applied a SHAP-based interpretability framework to the LASSO-selected radiomic features under both the Real + GBA and Real + TDA training scenarios. Figure 6 provides a swarm-style overview of feature impacts, with each marker representing how one feature influenced an individual prediction; features are sorted by their mean absolute contribution to spotlight the most decisive biomarkers. In Fig. 7, we aggregate these contributions into a global ranking, making it easier to compare average effect sizes across all samples. A detailed description of each feature, including its mathematical and potential biological interpretation, is provided in Supplementary Table S2.

The model identified GLRLM run length non-uniformity in the left caudate as the top predictor, reflecting increased tissue texture heterogeneity in MS cases (Fig. 6). Morphological descriptors, such as shape flatness in the right thalamus and shape elongation in the right putamen, also ranked highly, with elevated values indicating less spherical structures and suggesting subtle anatomical distortions.

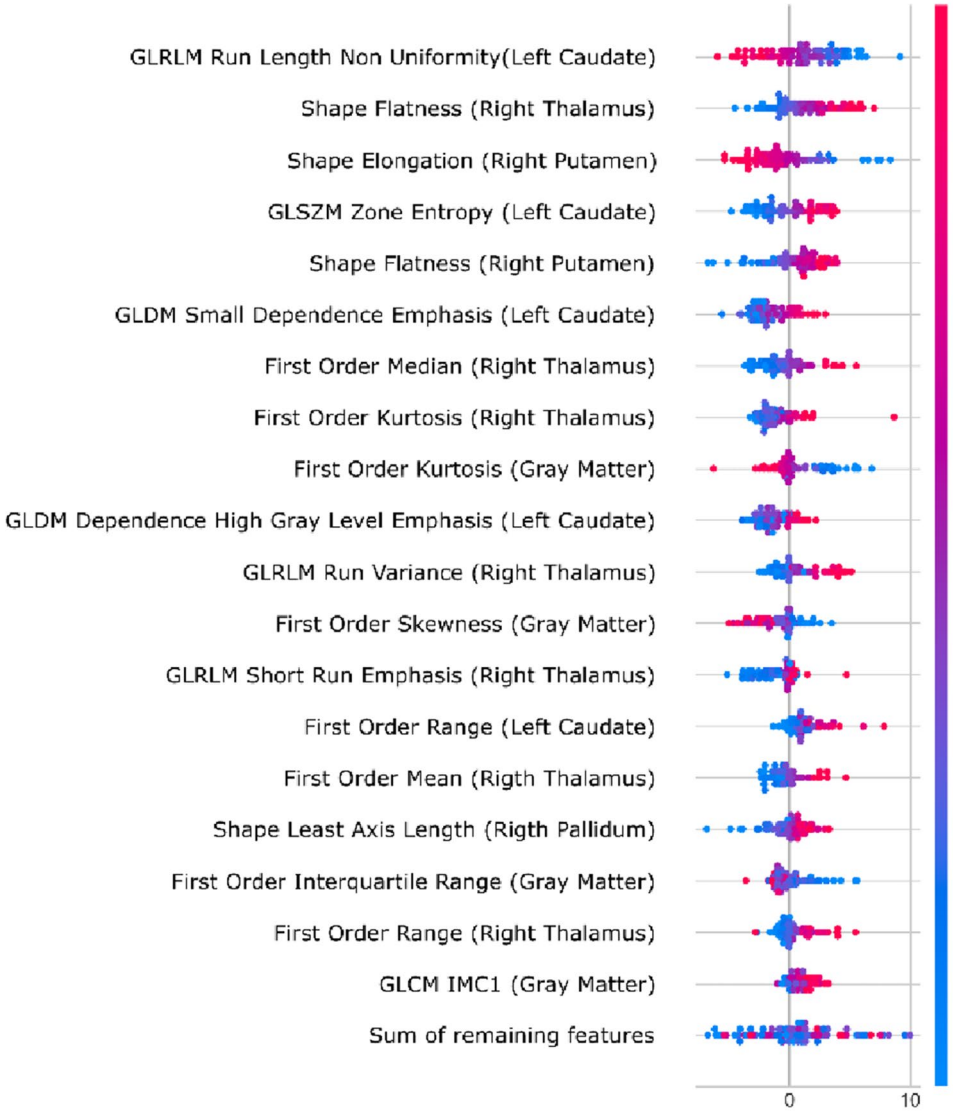

Fig. 6 Contribution of radiomic features to individual predictions. SHAP swarm plot showing the impact of the 20 most important features. Each point represents a subject in the test set. The X-axis indicates the SHAP value, where positive values push the prediction toward MS and negative values toward HC. The color represents the value of the feature for that subject, from low (blue) to high (red)

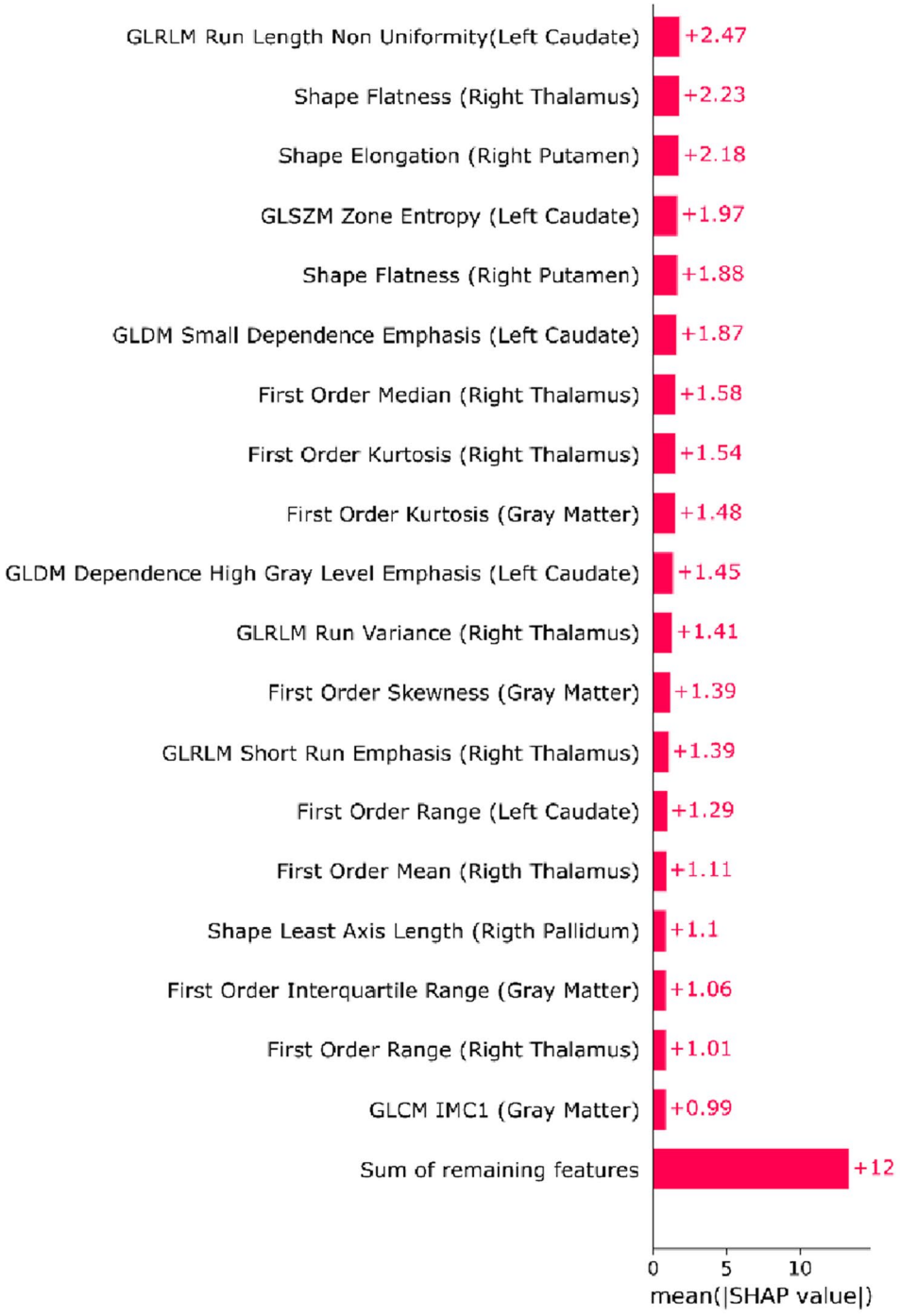

Fig. 7 Overall importance of radiomic features. Bar chart ranking the 20 most influential features according to their average impact on the model (mean absolute SHAP value). Run Length Non-Uniformity of GLRLM in the left caudate was the most important feature, followed by Flatness in the right thalamus. The results underscore the predominant role of texture and shape descriptors for classification

To probe relationships in more detail, Fig. 8 presents a dependence analysis for the top contributing features, demonstrating how combination of two features can increase the model's propensity to predict MS. Finally, Supplementary Figure S3 illustrates a case-level breakdown for a singular patient, showing exactly which features drove the model toward an MS diagnosis and which opposed it.

Across both experimental settings, features from Caudate, Thalamus and Putamen emerge consistently as top predictors. Their SHAP distributions remain stable when synthetic samples are introduced, indicating that the

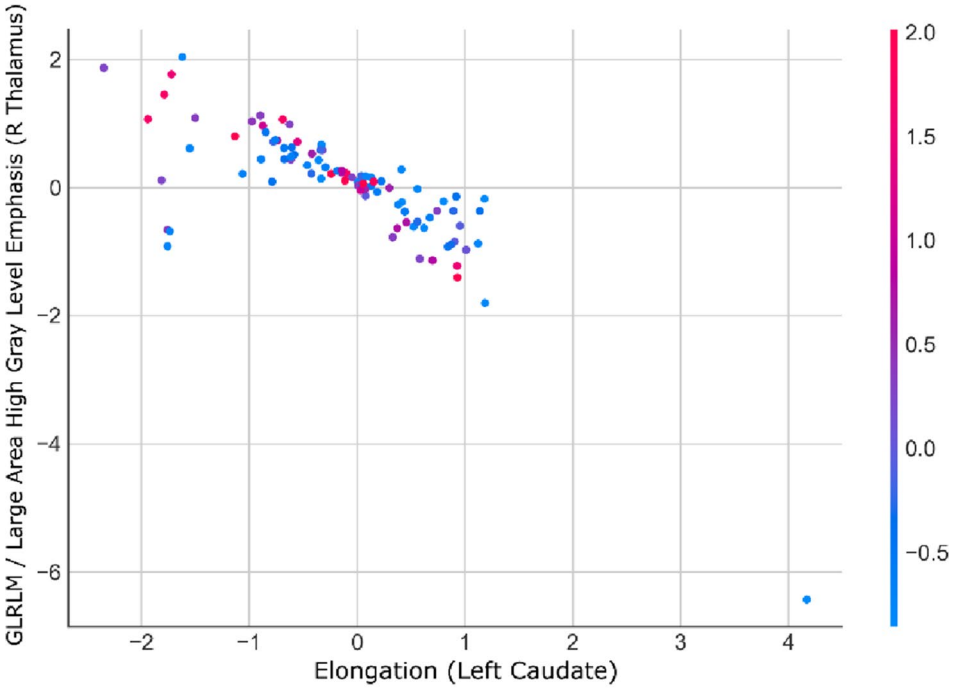

Fig. 8 SHAP dependence plot for left caudate nucleus Elongation: X shows elongation values, Y shows SHAP values for right thalamus Large Area High Gray Level Emphasis (GLRLM), revealing a negative trend; point color encodes the interacting feature's magnitude

inclusion of GBA data does not substantially distort the model's reliance on these key anatomical features.

Clinical utility assessment

When evaluating performance at the pre-specified operating point, the GBA strategy demonstrated a modest but superior diagnostic advantage compared to traditional data augmentation (TDA). This improvement is reflected in Youden's J index, which increased from 0.78 with TDA to 0.82 with GBA. Clinically, this translated into a reduction of the overall error rate from 11% to 9%, which is equivalent to avoiding two misclassifications for every 100 cases evaluated. Furthermore, the model's discriminative power improved, with an increase in the positive likelihood ratio (LR+) from 8.8 to 11.3 and a decrease in the negative likelihood ratio (LR−) from 0.133 to 0.109, indicating a greater capacity to both confirm and rule out the disease.

Discussion

This study demonstrated that augmenting radiomic feature sets with synthetically generated data enhances the performance of machine learning classifiers for distinguishing MS patients from healthy controls. Our findings indicate that while both TDA and GBA improve model performance over using real data alone, GBA, particularly when paired with a ResNet classifier, yields a marginal yet consistent advantage, achieving a top F1-score of 0.957 ± 0.018.

A primary concern with generative models is their potential to either replicate training data or produce low-fidelity samples that degrade model performance. Our results address both issues. The quantitative evaluation of the GBA images showed high quality and fidelity, with an SSIM of 0.904 and a PSNR of 38.9 dB, indicating strong structural and pixel-level correspondence to real MRI scans. Critically, our nearest-neighbor analysis confirmed that the synthetic images are novel and not replicas of the training set, mitigating concerns of data leakage or overfitting.

Although absolute gains with GBA are modest (≈ 1–2 pp), at clinically relevant thresholds they translate into fewer missed MS cases and fewer false alarms under external, site-held-out evaluation. At a 30% pre-test probability, a positive test increases post-test probability from 79.0% (TDA) to 82.8% (GBA), and a negative reduces it from 5.41% to 4.45%. Scaled to 1,000 MS-suspect referrals by year, this operating point would prevent ~ 10 additional false negatives and ~ 10 false positives, which can be consequential in settings prioritizing early treatment initiation and resource stewardship.

UMAP projections (Fig. 5) reveal that both TDA and GBA successfully reproduce the same high-dimensional manifold as real MRI samples, yet they differ subtly in distribution. TDA generates tight clusters at the core of the data (average energy distance 0.1974), while GBA yields a slightly broader dispersion (energy distance 0.2348). This increased spread marginally weakens class

coherence (ARI 0.1370 vs. 0.1481 for TDA) but appears to drive the small, consistent gains in F1-score and overall accuracy observed with GBA when training a ResNet classifier. A Friedman test ($p = 0.112$) indicates no statistically significant difference among leading models, raising questions about whether the added complexity and computational cost of GANs are warranted. However, in high-stakes clinical contexts such as multiple sclerosis diagnosis, where even a 1–2% improvement in recall can be critical, these modest enhancements suggest that GBA remains a compelling option when resources allow.

Our work is situated within a broader effort to overcome the primary obstacles in clinical radiomics: multicenter data heterogeneity and the limitations of small cohorts. As outlined in [33], these challenges often lead to overfitting and poor generalizability, hindering the translation of radiomic models into routine practice. Our study directly confronts this by proposing a data-centric solution. While some research focuses on improving robustness at the acquisition level as [34] rigorously evaluated the stability of QSM-based radiomic features against variations in echo times and grey levels, our approach provides a complementary, model-level strategy. By using GBA to enrich the training data, we aim to build models that are inherently more resilient to the site-specific variations that persist even after image harmonization.

Our application of radiomics for MS diagnosis complements other recent investigations targeting different, yet equally critical, clinical questions. For example [35], developed a radiomic warning sign by analyzing normal-appearing white matter to predict future lesion development, demonstrating the prognostic power of features invisible to the naked eye. Similarly [36], designed a contrast-free model to classify lesion activity by integrating features from five MRI sequences, including susceptibility-weighted imaging (SWI). Their multi-sequence approach and our data augmentation strategy represent two distinct paths toward the shared goal of reducing reliance on gadolinium. While their work leverages the unique information from different MRI physics, our study demonstrates how the robustness of a single-sequence (T1-weighted) model can be enhanced through synthetic data. Our approach aligns with the work of [15], who created a joint model of radiomics and lesion spatial distribution for the more complex task of discriminating between MS, NMOSD, and MOGAD. Similarly [37], showed that fusing novel radiomic features directly with imaging data enhanced MS lesion segmentation. Together, these studies underscore that advanced data-centric and feature-engineering strategies are key to unlocking higher performance in neuroimaging AI.

Furthermore, the interpretability analysis via SHAP provides confidence that our model is learning biologically relevant patterns. The consistent identification of the Caudate, Thalamus, and Putamen as key predictive regions aligns with known MS pathophysiology involving deep gray matter structures. This finding contrasts with the work of [38], who, in a pediatric cohort, identified the GLCM "contrast" feature from T2-FLAIR images as the most significant predictor for lesion classification. This divergence highlights that radiomic signatures may be population-specific, and the features driving classification in adult MS focused on subcortical atrophy may differ from those characterizing inflammatory lesions in pediatric MS.

Taken together, the center-wise external validation frames GAN-based augmentation as one of several viable levers for strengthening MS radiomics pipelines when training data are sparse and heterogeneous. Its gains over TDA are reproducible but measured, reinforcing the view that augmentation strategies alone are unlikely to close the generalization gap without parallel advances in feature engineering, harmonization, and clinical integration. In this context, the modest improvements observed here should be read less as a destination than as a proof-of-principle, demonstrating that carefully controlled synthetic data can enhance robustness without eroding biological plausibility. Situating these findings alongside complementary work on acquisition-level stability, multi-sequence fusion, and spatial priors points to a converging agenda: building neuroimaging AI systems that are resilient across sites, diagnostically transparent, and clinically actionable.

Conclusion

This study examined whether AI-augmented training improves a basal-ganglia radiomics classifier for multiple sclerosis under center-wise external evaluation. Using a LOCO design, the entire held-out center served as the external test set, while inner 5-fold cross-validation (80/20, subject-level) within the training centers segmentation, GAN stabilization, feature selection, and classifier tuning. Across centers, GAN-based augmentation produced small but consistent improvements over TDA and real-only pipelines, most evident in accuracy and precision for a tabular ResNet, with overlapping confidence intervals for some metrics. These findings indicate incremental, rather than transformative, benefits. It is important to note that higher-order texture radiomic features are highly sensitive to segmentation inaccuracies; even minor boundary perturbations can induce substantial shifts in feature values. Caution is therefore warranted when segmenting and extracting features from GBA images.

Interpretability analyses supported anatomical plausibility: SHAP attributions preserved the salience of caudate, thalamus, and putamen, suggesting that synthetic

samples did not distort disease-relevant signal. Methodologically, the center-held-out reporting mitigates optimistic bias and clarifies generalization across sites; overall aggregates summarize cross-center tendencies, while center-wise metrics preserve site-level evidence.

Two limitations frame the scope of our claims. First, although multicenter, the cohort originates from a single country and lacks public external validation, this limits generalizability. Second, segmentation variability, even with internal validation and expert review, can propagate to higher-order textures, affecting feature stability. From a clinical perspective, the observed gains may still be operationally relevant in workflows where marginal error reductions matter, provided calibration and thresholding are tuned to the intended use.

In sum, AI-augmented training offers a scalable, data-centric lever for modestly improving MS radiomics under site-held-out testing without compromising interpretability. Future work should extend this method to additional centers and countries, incorporate public datasets for external validation, and quantify clinical utility with calibration analyses and decision-analytic assessments; comparative studies against synthetic-data generalization frameworks and spatially informed radiomics will further clarify where synthetic augmentation delivers the greatest return.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01985-7.

Supplementary Material 1

Supplementary Material 2

Author contributions

This research was conceptualized by EE and FJ. EE led the design and implementation of the methodology, carried out the investigation, including data curation, software development, and formal analysis. EE also prepared the original draft of the manuscript and created the visualizations. FJ supervised the entire research process, guided the theoretical and methodological framing, and contributed significantly to the critical revision and refinement of the manuscript. Both authors reviewed and approved the final version of the manuscript.

Funding

This research received no external funding.

Data availability

The data presented in this study are not publicly available due to internal privacy policies regarding their collection and use. However, the dataset can be made available at reasonable request to the corresponding author, provided that the requesting party complies with the ethical and privacy considerations established by the collaborating institutions.

Declarations

Ethics approval and consent to participate

This study was a retrospective analysis of fully anonymized and deidentified medical images. The Hospital Civil de Guadalajara Committee determined that this research did not constitute human subjects research and waived the requirement for formal ethics committee approval and individual informed consent. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable. This study does not include any identifiable personal data or images.

Competing interests

The authors declare no competing interests.

Received: 2 June 2025 / Accepted: 7 October 2025
Published online: 11 November 2025

References

1. Wang L, et al. Human autoimmune diseases: a comprehensive update. J Intern Med. 2015;278(4):369–95.
2. Kalinin I, et al. The impact of intracortical lesions on volumes of subcortical structures in multiple sclerosis. Am J Neuroradiol. 2020;41(5):804–8.
3. Dobson R, Giovannoni G. Multiple sclerosis–a review. Eur J Neurol. 2019;26(1):27–40.
4. McGinley MP, et al. Diagnosis and treatment of multiple sclerosis: a review. JAMA. 2021;325(8):765–79.
5. Murray T. Diagnosis and treatment of multiple sclerosis. BMJ. 2006;332(7540):525–7.
6. Jakimovski D, et al. Long-standing multiple sclerosis neurodegeneration: volumetric magnetic resonance imaging comparison to Parkinson's disease, mild cognitive impairment, Alzheimer's disease, and elderly healthy controls. Neurobiol Aging. 2020;90:84–92.
7. Schoonheim MM, et al. Subcortical atrophy and cognition: sex effects in multiple sclerosis. Neurology. 2012;79(17):1754–61.
8. Azevedo CJ, et al. Thalamic atrophy in multiple sclerosis: a magnetic resonance imaging marker of neurodegeneration throughout disease. Ann Neurol. 2018;83(2):223–34.
9. Zivadinov R, et al. Thalamic atrophy is associated with development of clinically definite multiple sclerosis. Radiology. 2013;268(3):831–41.
10. Schoonheim MM, et al. Disability in multiple sclerosis is related to thalamic connectivity and cortical network atrophy. Multiple Scler J. 2022;28(1):61–70.
11. Trufanov A, et al. Basal ganglia atrophy as a marker of multiple sclerosis progression. Biomarkers Neuropsychiatry. 2023;9:100073.
12. Mayerhoefer ME, et al. Introduction to radiomics. J Nucl Med. 2020;61(4):488–95.
13. Tavakoli H, et al. Investigating the ability of radiomics features for diagnosis of the active plaque of multiple sclerosis patients. J Biomedical Phys Eng. 2023;13(5):421.
14. Luo X, et al. Multi-lesion radiomics model for discrimination of relapsing-remitting multiple sclerosis and neuropsychiatric systemic lupus erythematosus. Eur Radiol. 2022;32(8):5700–10.
15. Luo X, et al. Joint radiomics and spatial distribution model for MRI-based discrimination of multiple sclerosis, neuromyelitis optica spectrum disorder, and myelin-oligodendrocyte-glycoprotein-IgG-associated disorder. Eur Radiol. 2024;3(7):4364–75.
16. Reiazi R et al. The impact of the variation of imaging factors on the robustness of computed tomography radiomic features: A review. medRxiv preprint medRxiv: 0 9.20137240. 2020.
17. Ligero M, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol. 2021;31(3):1460–70.
18. Barile B, et al. Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis. Comput Methods Programs Biomed. 2021;206:106113.
19. Brugnara G, et al. Addressing the generalizability of AI in radiology using a novel data augmentation framework with synthetic patient image data: proof-of-concept and external validation for classification tasks in multiple sclerosis. Radiology: Artif Intell. 2024;6(6):e230514.
20. Suganthi K et al. Review of medical image synthesis using GAN techniques. In: ITM Web of Conferences. vol. 37. EDP Sciences; 2021. p. 01005.
21. Whybra P, et al. Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. Sci Rep. 2019;9(1):1–10.

22. de Farias EC, et al. Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. Sci Rep. 2021;11(1):21361.
23. Park S, et al. Seed growing for interactive image segmentation using SVM classification with geodesic distance. Electron Lett. 2017;53(1):22–4.
24. Pieper S et al. 3D Slicer. In: 2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821). IEEE; 2004. pp. 632–635.
25. Çiçek Ö et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19. Springer; 2016. pp. 424–432.
26. Isola P et al. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. pp. 1125–1134.
27. Van Griethuysen JJ, et al. Computational radiomics system to Decode the radiographic phenotype. Cancer Res. 2017;77(21):e104–7.
28. Zwanenburg A et al. Image biomarker standardization initiative. arXiv preprint arXiv:161207003. 2016.
29. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 2020;408:189–215.
30. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
31. LeCun Y, et al. Deep Learn Nat. 2015;521(7553):436–44.
32. Lundberg SM et al. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
33. Kazemzadeh K. The role of radiomics in multiple sclerosis imaging: an updated review of literature. Neurol Lett. 2025;4(2):107–13.
34. Fiscone C, et al. Assessing robustness of quantitative susceptibility-based MRI radiomic features in patients with multiple sclerosis. Sci Rep. 2023;13(1):16239.
35. Kelly BS, et al. A radiomic warning sign of progression on brain MRI in individuals with MS. Am J Neuroradiol. 2024;45(2):236–43.
36. Shahbazi–Gahrouei D. Machine learning-based classification of multiple sclerosis lesion activity using multi-sequence MRI radiomics: A complete analysis of T1, T2, FLAIR, DWI, and SWI features. Pol J Radiol. 2025;90:394–403.
37. Alsahanova N et al. Integrating Radiomics with Deep Learning Enhances Multiple Sclerosis Lesion Delineation. arXiv preprint arXiv:2506.14524. 2025.
38. Faustino R, et al. Neuroimaging characterization of multiple sclerosis lesions in pediatric patients: an exploratory radiomics approach. Front NeuroSci. 2024;18:1–9.

Publisher's note