Radiomics Feature Profiling of Brodmann Regions in Structural MRI: A Machine Learning Study of Intensive Verbal Memorisation (Huffaz vs Controls)

Quantifying Memorisation-Related Neuroplasticity Using PyRadiomics and Nested Cross-Validation

Mohd Zulfaezal Che Azemin[1], Iqbal Jamaludin[1], Abdul Halim Sapuan[1], Mohd Izzuddin Mohd Tamrin[2]
Integrated Omics Research Group-Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, Kuantan, Malaysia[1]
Kulliyyah of ICT, International Islamic University Malaysia, Gombak, Kuala Lumpur, Malaysia[2]

***Abstract*—Memorisation-based cognitive training has been hypothesized to relate to experience-dependent brain plasticity; however, quantitative evidence at the regional level remains limited. We hypothesized that radiomics descriptors extracted from Brodmann-area volume-of-interest (VOI) regions in pre-processed structural MRI would contain sufficient information to discriminate Quran memorizers (Huffaz) from non-memorizers (controls), and we evaluated this hypothesis using a fully nested validation framework. T1-weighted MRI volumes were pre-processed using a voxel-based morphometry pipeline, and VOIs were defined using Brodmann-area masks. Using PyRadiomics, first-order and texture features were extracted per VOI and combined into a feature matrix for classification. Models were evaluated using repeated nested cross-validation (outer 5-fold × 10 repeats; inner 5-fold for tuning), with ROC-AUC as the primary metric. Random Forest achieved the strongest discrimination (AUC = 0.6704 ± 0.1792), followed by Logistic Regression (AUC = 0.5948 ± 0.2153), while SVM with an RBF kernel underperformed (AUC = 0.4356 ± 0.1927). One-sided testing against chance (AUC = 0.5) indicated above-chance performance for Random Forest and Logistic Regression, but not for SVM-RBF. These results should be interpreted as exploratory because the cohort is small (n = 47) and no independent external validation cohort was available. Practically, the observed effect sizes suggest that VOI-based radiomics may capture detectable group-associated imaging signatures under the current preprocessing and VOI assumptions, motivating validation on larger cohorts, sensitivity analysis (e.g., discretization/normalization settings), and assessment of probability calibration.**

Keywords*—Radiomics; PyRadiomics; Magnetic Resonance Imaging (MRI); voxel-based morphometry (VBM); neuroplasticity; Huffaz; Brodmann-areas; Volume of Interest (VOI); texture analysis; machine learning; nested cross-validation; ROC-AUC; Random Forest

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) provides a non-invasive window into brain structure, making it central to investigating neuroplasticity associated with intensive learning and memory training. Recent evidence from verbal memory experts indicates that long-term memorisation practice can be accompanied by measurable structural brain differences, reinforcing the plausibility of detectable morphology changes in high-repetition textual memorisation cohorts [1]. In our earlier work on Quran memorisers (Huffaz), a volume-of-interest (VOI) fractal framework was able to localise group differences that were not apparent in global measurements, suggesting that region-specific quantitative descriptors may be necessary to capture subtle experience-related neuroplasticity [2].

While voxel-based morphometry (VBM) remains a common approach for detecting regional grey matter differences, it typically summarises effects using voxel-wise statistics and may miss complex textural or distributional patterns within anatomically defined regions. Radiomics addresses this gap by extracting a large number of engineered quantitative features (e.g., first-order statistics and texture descriptors) from medical images, thereby encoding patterns that may not be obvious visually or via simple volumetry [3]. However, reproducibility has been a major barrier for radiomics, motivating standardised feature definitions and reporting practices. The Image Biomarker Standardisation Initiative (IBSI) provides consensus definitions for many radiomic features to improve comparability across studies and software implementations [4]. In addition, biological interpretation remains important; radiomic features are increasingly studied as proxies of underlying tissue heterogeneity, organisation, and microstructural variation [5].

Practical radiomics pipelines typically include: 1) image preparation, 2) region-of-interest definition, 3) feature extraction, and 4) predictive modelling with rigorous validation [6]. PyRadiomics is a widely used open-source implementation for IBSI-aligned radiomic feature extraction and supports multiple feature classes and filtered image types (e.g., wavelet decompositions) [3]. For clinical translation, recent guidance has emphasised methodological rigour, transparency, and robust evaluation, particularly in small-to-moderate datasets where model performance can be unstable and overly optimistic if validation is not handled carefully [7]. Reporting recommendations such as TRIPOD+AI further encourage clear documentation of data handling, modelling steps, and performance metrics for machine learning prediction studies [8].

A key methodological risk in small neuroimaging datasets is biased performance estimation when model selection (e.g., hyperparameter tuning and feature selection) is not isolated from the outer evaluation loop. Nested cross-validation is recommended to reduce this bias, because it repeats all modelling steps within each training fold before testing on held-out data [9], and practical guidance for medical imaging AI specifically highlights cross-validation pitfalls and best practices [10]. Data leakage can also occur when resampling or preprocessing steps are applied outside the correct training-only scope, which has been shown to inflate radiomics performance estimates [11]. Finally, performance is commonly summarised with receiver operating characteristic (ROC) analysis and the area under the curve (AUC), and statistical tests or confidence estimation should be used to interpret whether observed AUCs exceed chance levels and to support model comparison [12], [13].

Motivated by these considerations, this study extends our earlier VOI-based neuroplasticity investigation in Huffaz [2] by replacing fractal descriptors with PyRadiomics-engineered features extracted from preprocessed grey matter MRI volumes within Brodmann-area-based VOIs. We then evaluate multiple classical machine learning models under a nested cross-validation framework and statistically test whether the observed AUC distributions are significantly above chance, aiming to provide an IBSI-aware, publication-ready workflow for quantitative neuroimaging analysis.

Research gap and objectives. Despite growing interest in radiomics for characterizing subtle tissue patterns, two gaps remain in the context of memorisation-related neuroimaging studies: 1) prior work on expert memorisers has more often focused on voxel-wise or volumetric summaries, which may not capture within-region distributional and texture patterns; and 2) radiomics findings can be sensitive to design choices (e.g., discretization/bin width, normalization, and leakage-free validation), yet these sensitivities are not consistently discussed or stress-tested in small neuroimaging cohorts.

While radiomics has shown promise in characterizing subtle imaging patterns across oncology and, increasingly, neuroimaging applications, several limitations are recurrent in the literature. First, many studies are conducted in small cohorts with high-dimensional feature spaces, which can inflate apparent performance and lead to unstable results when validation is not strictly nested or externally replicated. Second, reported performance is often difficult to compare across studies due to heterogeneity in preprocessing pipelines, ROI/VOI definitions, feature calculation settings, and model selection procedures. Third, radiomics features are known to be sensitive to methodological choices, particularly intensity discretization (e.g., fixed bin width vs. fixed bin count), intensity normalization, resampling resolution, and filtering options, all of which can shift feature distributions and downstream model behaviour. Consequently, reproducible radiomics reporting benefits from explicitly documenting feature-extraction settings and from conducting sensitivity analyses to quantify how binning and normalization decisions affect both features and model performance. These considerations motivated our emphasis on transparent reporting of PyRadiomics configuration and leakage-controlled model selection via repeated nested cross-validation; nevertheless, we treat our results as exploratory and highlight robustness testing and external validation as priorities for future work.

Accordingly, this study addresses the following research questions:

RQ1: Do VOI-based PyRadiomics first-order and texture features extracted from pre-processed T1-weighted MRI provide above-chance discrimination between Huffaz and controls under repeated nested cross-validation?

RQ2: Which classical classifier family (Logistic Regression, SVM-RBF, Random Forest) yields the most stable discrimination under leakage-controlled model selection?

Hypothesis (H1): Radiomics features extracted from the specified Brodmann-area VOIs will yield ROC-AUC statistically greater than 0.5 under a fully nested evaluation protocol.

Hypothesis (H2): Non-linear ensemble learning (Random Forest) will outperform linear (Logistic Regression) and kernel-based (SVM-RBF) models in this high-dimensional, low-sample setting.

II. Methodology

A. Overview

This study proposes a reproducible radiomics-based pipeline to quantify structural differences between Huffaz (H) and non-Huffaz (NH) participants using pre-processed structural MRI and atlas-derived Brodmann-area (BA) masks (see Fig. 1). The workflow consists of: 1) automated cohort/label parsing from filenames, 2) region-of-interest (VOI) preparation and spatial alignment of masks to subject space, 3) extraction of hand-crafted radiomics features using PyRadiomics, and 4) model development and evaluation using nested cross-validation with inner-loop hyperparameter tuning.

B. Experimental Setup

1) Hardware: Experiments were executed on a Runpod GPU instance with 6 vCPU, 15 GB RAM, and 16 GB VRAM. The GPU primarily supported the CUDA-enabled runtime environment; however, the classical machine learning models were trained using CPU-based scikit-learn routines.

2) Software: The pipeline was implemented in Python 3.11. Key libraries and versions were: PyRadiomics 3.0.1 (feature extraction), SimpleITK 2.5.3 (PyRadiomics backend), NiBabel 5.3.3 (NIfTI handling and resampling), NumPy 1.26.4, SciPy 1.17.0, Pandas 2.3.3, and scikit-learn 1.8.0. PyRadiomics logging was set to error level to reduce non-essential output. All outputs (features, metadata, fold-wise and summary results) were saved to disk to ensure auditability and repeatability.

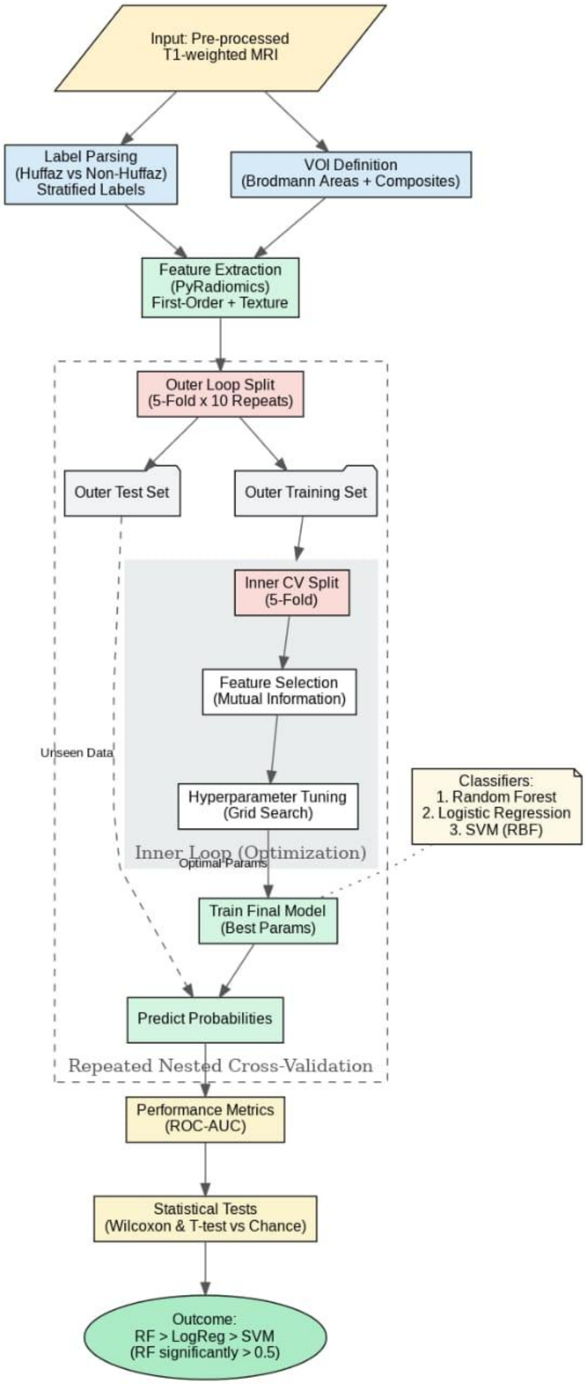
Input: Pre-processed T1-weighted MRI
Label Parsing (Huffaz vs Non-Huffaz) Stratified Labels
VOI Definition (Brodmann Areas + Composites)
Feature Extraction (PyRadiomics) First-Order + Texture
Outer Loop Split (5-Fold x 10 Repeats)
Outer Test Set
Outer Training Set
Inner CV Split (5-Fold)
Feature Selection (Mutual Information)
Unseen Data
Hyperparameter Tuning (Grid Search)
Classifiers: 1. Random Forest 2. Logistic Regression 3. SVM (RBF)
Inner Loop (Optimization)
Optimal Params
Train Final Model (Best Params)
Predict Probabilities
Repeated Nested Cross-Validation
Performance Metrics (ROC-AUC)
Statistical Tests (Wilcoxon & T-test vs Chance)
Outcome: RF > LogReg > SVM (RF significantly > 0.5)

Fig. 1. Radiomics methodology flowchart showing the end-to-end pipeline: input pre-processed T1-weighted MRI, automated label parsing (Huffaz vs. non-Huffaz), and VOI definition using Brodmann-area masks, PyRadiomics feature extraction, and repeated nested cross-validation with inner-loop feature selection and hyperparameter tuning, followed by ROC-AUC–based evaluation and statistical testing.

C. *Data Organisation and Label Encoding*

The Neuroimaging Informatics Technology Initiative (NIfTI) pre-processed Magnetic Resonance Imaging files followed the naming convention:

smwc1{Group}{Sex}_{ID}.nii, where {Group} ∈ {H, NH} and {Sex} ∈ {F, M}.

An automated parser extracted the study label and participant attributes:

- Class label $y = 1$ for Huffaz (H) and $y = 0$ for non-Huffaz (NH),
- Sex (F/M),
- A combined stratification label {Group}_{Sex} ∈ {H_F, H_M, NH_F, NH_M}.

This stratification label was used to preserve both class and sex proportions in cross-validation splits.

Class balance. The dataset comprised n = 47 participants split into Huffaz (H) and non-Huffaz (NH) groups. We report the group counts as H: n = 23, NH: n = 24 to make any imbalance explicit. To reduce bias from potential imbalance during training, all classifiers were trained with balanced class weights, and cross-validation splits were stratified using the combined {Group}_{Sex} label to preserve the joint distribution of group and sex across folds.

D. *Volume-of-Interest (VOI) Definition*

Based on prior VOI findings, the following regions were included:

- BA20, BA30, BA33,
- BA39,
- BA21,
- BA37,

Two composite VOIs were created to represent broader functional regions:

1) Anterior cingulate cortex (ACC) as the union of BA24 and BA32:

$$\text{ACC} = \text{BA24} \cup \text{BA32}$$

2) Frontal composite as the union of BA08, BA09, BA10, BA11, BA44, BA45, BA46, and BA47:

$$\text{Frontal} = \bigcup_{b\in\{8,9,10,11,44,45,46,47\}} \text{BA}b$$

Composite masks were computed programmatically using voxel-wise logical OR and saved for reproducibility.

E. *Radiomics Feature Extraction (PyRadiomics)*

Radiomics features were extracted per subject and per VOI using PyRadiomics 3.0.1. The extractor was configured to produce scalar features (not voxel-wise feature maps) by setting:

- voxelBased = False,
- force2D = False.

Intensity discretisation and preprocessing settings were:

- fixed bin width: binWidth = 0.01,
- no intensity normalisation: normalize = False,

- no resampling of the image intensity grid: resampledPixelSpacing = None.

Only PyRadiomics feature families were used, explicitly enabling:

- First-order,
- GLCM, GLRLM, GLSZM, GLDM, and NGTDM.

Shape features were excluded because atlas-defined VOIs are fixed anatomical regions after normalisation, and shape descriptors may be largely invariant across participants.

Each feature was stored with a unique VOI-qualified identifier:

$$\text{FeatureName} = \text{VOI} \ \| \ \text{FeatureKey}$$

E.g., BA20__original_firstorder_Mean. PyRadiomics sometimes returns scalar values as NumPy arrays of length 1; these were converted to float scalars to construct the feature matrix consistently.

The extracted feature matrix was saved as features_{mode}_bw0.01.csv and metadata (case ID, group, sex, label, stratification label) as metadata.csv.

Reproducibility considerations. Radiomics features can be sensitive to discretization and intensity preprocessing choices. In this study, we used a fixed bin width (binWidth = 0.01) and disabled intensity normalisation (normalise = False) to keep the pipeline explicit and reproducible. Because such settings may influence downstream performance, future work will include sensitivity analyses across plausible bin widths and normalization strategies, reported alongside the primary results.

F. Classification Models and Feature Selection

Three supervised classifiers were evaluated:

1) Logistic regression: with balanced class weights and L_2regularisation (liblinear solver).

2) Support Vector Machine (RBF kernel): with probability estimates enabled and balanced class weights.

3) Random Forest with 500 trees, balanced class weights, and full CPU parallelism.

To address the high-dimensional feature space relative to the sample size, univariate feature selection was performed using mutual information:

- SelectKBest(mutual_info_classif).

The number of selected features $k$was tuned from:

$$k \in \{10,20,50,100,200,\text{all}\}$$

with automatic adjustment to ensure $k \leq$ number of available features.

G. Nested Cross-Validation and Hyperparameter Optimisation

Model evaluation employed nested cross-validation to avoid optimistic bias from hyperparameter tuning. Splits were stratified using the combined {Group}_{Sex} label to preserve distributions of both study group and sex.

- Outer loop: Stratified K-fold ($k = 5$) repeated 10 times (50 outer test evaluations per model).
- Inner loop: Stratified K-fold ($k = 5$) used for hyperparameter selection.

Hyperparameters were tuned using GridSearchCV with ROC-AUC as the optimisation criterion. Search spaces include:

- Logistic Regression: $C \in \{0.1,1,10,100\}$
- SVM-RBF: $C \in \{0.1,1,10,10, \gamma \in \{\text{scale}, 0.01,0.1,1\}$
- Random Forest: max_depth ∈ {None, 5, 10, 20}, min_samples_split ∈ {2, 5, 10}

A fixed seed (seed = 42) controlled the cross-validation shuffling and mutual-information estimation to support reproducibility.

H. Evaluation Metrics and Reporting

Model performance was assessed on the outer test folds of the repeated nested cross-validation procedure to ensure that all reported metrics reflect generalisation to unseen data. For each outer fold, classifiers produced class membership probabilities for the Huffaz class (H), which were used to compute the receiver operating characteristic area under the curve (ROC-AUC) as the primary discrimination metric. AUC was selected because it is threshold-independent and is widely used for binary classification evaluation, particularly under potential class imbalance. To summarise performance across resampling runs, fold-level metrics were aggregated over 50 outer-fold evaluations per model (5 folds × 10 repeats). For each classifier, we reported the mean and standard deviation of outer-fold AUC values.

Note on statistical dependence. Because the outer-fold AUCs are produced from repeated cross-validation, they are not strictly independent (training sets overlap across folds and repeats). Therefore, p-values from fold-wise tests against AUC = 0.5 should be interpreted as supportive evidence rather than definitive confirmatory inference. Consistent with this limitation, we emphasize effect size (mean AUC), variability (standard deviation), and stability across repeats, and we frame the findings as exploratory pending confirmation in an independent validation cohort.

To test whether each model performed significantly better than random guessing, we conducted one-sided hypothesis tests against the null AUC = 0.5, using the set of outer-fold AUC values. Two complementary tests were applied:

1) Wilcoxon signed-rank test on $(\text{AUC} - 0.5)$, assessing whether the median improvement over chance was greater than zero.

2) One-sample t-test on AUC values, assessing whether the mean AUC exceeded 0.5.

III. Results and Discussion

A. Classification Performance (Repeated Nested Cross-Validation)

This performance was evaluated using repeated nested cross-validation (outer: 5 folds × 10 repeats; n = 50 outer-test AUCs per model). This design follows recommended practice to reduce optimistic bias when feature selection and hyperparameter tuning are performed [9].

Across the three classifiers trained on PyRadiomics VOI features (first-order + texture families), the mean AUC values were:

- Random Forest: 0.6704 ± 0.1792
- Logistic Regression: 0.5948 ± 0.2153
- SVM (RBF): 0.4356 ± 0.1927

Overall, Random Forest achieved the strongest discrimination, followed by Logistic Regression, while SVM_RBF underperformed relative to chance.

These results suggest that, under the current VOI definition and preprocessing assumptions, the extracted radiomics descriptors contain group-associated imaging signatures that provide above-chance discrimination in this cohort. However, given the modest mean AUC values and substantial fold-to-fold variability, these signals should not be interpreted as definitive evidence of neuroplastic change, but rather as exploratory markers that warrant validation and robustness testing.

B. Statistical Significance vs. Chance Level (AUC = 0.5)

The fold-level hypothesis tests were designed to assess whether each classifier's discrimination ability exceeded random guessing, using AUC = 0.5 as the null benchmark. Two complementary one-sided tests were applied. First, the Wilcoxon signed-rank test was performed on the set of values $(\mathrm{AUC}-0.5)$, testing whether the median improvement over chance was greater than zero. This non-parametric test is useful when the distribution of fold AUCs may not be normal and provides robustness to outliers. Second, a one-sample t-test was applied directly to the AUC values to test whether the mean AUC was greater than 0.5. While the t-test assumes approximately normal sampling distribution of the mean, the paired use of a non-parametric and parametric test provides a more interpretable and transparent assessment of "better-than-chance" performance.

For Logistic Regression, the mean AUC across the 50 outer-fold evaluations was 0.5948 with a relatively large standard deviation (0.2153), indicating that performance varied notably across cross-validation splits. Nevertheless, both tests yielded consistent evidence that the model exceeded chance: Wilcoxon $p = 0.00132$ and t-test $p = 0.00154$ (one-sided). This implies that, despite variability, the pipeline repeatedly produced AUC values above 0.5 more often than would be expected under random guessing. In practical terms, Logistic Regression appears to capture a modest but reliable linear separability in the radiomics feature space. The high variance suggests sensitivity to which subjects appear in each training/test fold, which is expected in small-sample radiomics; however, the consistent p-values across two different tests support that the observed effect is not driven solely by a few favourable folds.

In contrast, SVM with RBF kernel (SVM_RBF) showed a mean AUC of 0.4356 (std 0.1927), which is below the chance level of 0.5. Correspondingly, both one-sided tests for improvement over chance returned non-significant results (Wilcoxon $p = 0.989$, t-test $p = 0.989$). This indicates there is no evidence that the SVM_RBF model performed better than random guessing in this setting; instead, the direction of the effect suggests systematic underperformance. A mean AUC below 0.5 can happen when the learned decision function is effectively "reversed" relative to the chosen positive class, but in a properly configured pipeline with consistent label encoding, persistent below-chance performance more commonly reflects instability or poor generalisation for that model class under the current data regime. With only 47 subjects and a high-dimensional radiomics feature set, an RBF kernel can overfit subtle fold-specific patterns during inner-loop tuning and fail to transfer to outer-loop test folds, leading to degraded AUC.

The strongest results were obtained with the Random Forest classifier, which achieved a mean AUC of 0.6704 (std 0.1792). Both statistical tests provide very strong evidence that Random Forest performance exceeded chance: Wilcoxon $p = 3.78\times10^{-7}$ and t-test $p = 8.94\times10^{-9}$. This consistency across parametric and non-parametric testing suggests that improvement over chance is not only present, but also robust across the majority of folds. The Random Forest model likely benefits from its ability to model non-linear relationships and interactions among radiomic features, which is particularly relevant when combining multiple texture families across multiple VOIs. Although variability remains (std ≈ 0.18), the shift of the AUC distribution above 0.5 appears sufficiently large that even with fold-to-fold fluctuations, the overall performance advantage persists.

Statistical vs. practical significance. While the hypothesis tests against AUC = 0.5 provide evidence that some models exceed chance in this resampling framework, the practical significance depends on the magnitude and stability of performance. The relatively large standard deviations indicate performance instability across splits, consistent with high-dimensional, low-sample radiomics settings. Therefore, the current results are best viewed as evidence of a potentially learnable signal rather than a clinically actionable classifier.

C. Why Random Forest Performed Best?

The Random Forest's advantage is consistent with the nature of radiomics: many features are correlated and may interact nonlinearly. Tree ensembles can exploit non-linear relationships and feature interactions with comparatively little feature engineering. In contrast, Logistic Regression is linear and will only separate classes well if the discriminative signal is approximately linear in the selected feature space.

The relatively high standard deviations (≈0.18–0.22) indicate that performance depends noticeably on which subjects fall into each fold, as expected given the small cohort size (47 subjects) and high-dimensional radiomics space. This variability is also widely recognised as a challenge in radiomics pipelines and motivates careful model validation and feature selection.

D. Why SVM_RBF Underperformed?

The RBF SVM produced a mean AUC < 0.5 despite scaling and hyperparameter search. Likely contributors include:

1) Small-sample instability: RBF SVM can be sensitive to fold composition when sample size is limited; the decision boundary may overfit the inner-CV splits and fail to generalise.

2) Feature distribution characteristics: Radiomics features often include heavy tails and inter-feature dependencies; even with standard scaling, the effective geometry in feature space may not suit an RBF kernel under the tested grid.

3) High-dimensional, low-n regime: With many candidate features per VOI and relatively few subjects, kernel methods can become unstable unless regularisation and selection are very carefully constrained.

Given these results, SVM_RBF is not recommended as the primary model for this dataset without further stabilisation (e.g., more conservative feature selection, nested feature stability constraints, or a simpler kernel).

IV. Limitations

This study has several limitations that affect the reliability and generalizability of the findings. First, the cohort size (n = 47) is small relative to the dimensionality of radiomics features, which increases the risk of performance instability and optimistic estimates despite using nested cross-validation. This is reflected in the relatively large fold-to-fold standard deviations in AUC, suggesting that results are sensitive to which participants appear in each split.

Second, no independent external validation dataset was available; therefore, the results should be interpreted as exploratory rather than confirmatory, and the reported performance should not be treated as evidence of clinical utility.

Third, radiomics features can be sensitive to preprocessing choices such as discretization bin width and intensity normalization. We used a fixed bin width (binWidth = 0.01) and disabled normalization to make the pipeline explicit; however, alternative settings could change feature distributions and model performance. Fourth, the statistical tests against AUC = 0.5 rely on cross-validation-derived AUC samples that are not strictly independent due to overlapping training data across folds and repeats; thus, p-values are supportive rather than definitive. Finally, the VOI set was restricted to selected Brodmann regions and composites motivated by prior work; additional ROIs, confound control (e.g., age, education, scanner/site effects if applicable), and multimodal features may alter the observed signal. Future work should prioritise larger cohorts, external validation, sensitivity analyses of radiomics settings, and probability calibration reporting.

V. Conclusion

This study evaluated a VOI-based radiomics pipeline for distinguishing Huffaz from non-Huffaz participants using SPM pre-processed T1-weighted MRI and PyRadiomics first-order and texture features. Under repeated nested cross-validation, Random Forest achieved the highest mean discrimination (AUC = 0.6704 ± 0.1792), while Logistic Regression showed modest above-chance performance, and SVM-RBF did not generalize in this setting. Given the small cohort size and substantial variability across folds, the findings should be interpreted as exploratory evidence of group-associated imaging signatures rather than definitive proof of neuroplastic change or a clinically deployable classifier. Future work will focus on independent external validation, robustness analyses across radiomics preprocessing settings (e.g., discretization/normalization), expanded ROI coverage, and calibration assessment to determine whether predicted probabilities can be interpreted reliably.

Declaration on Generative AI

ChatGPT 5.2 was used to assist with language editing and to improve clarity during manuscript preparation. The authors take full responsibility for the content, confirm that all scientific interpretations and conclusions are their own, and approve the final version of the manuscript.

Acknowledgment

This work was supported by the Ministry of Higher Education of Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2024/WAS12/UIAM/01/3).

References

[1] U. Kumar, A. Singh, and P. Paddakannaya, “Extensive long-term verbal memory training is associated with brain plasticity,” *Sci. Rep.*, vol. 11, Art. no. 9712, May 2021, doi: 10.1038/s41598-021-89248-7.

[2] I. Jamaludin, M. Z. C. Azemin, M. I. M. Tamrin, and A. H. Sapuan, “Volume of Interest-Based Fractal Analysis of Huffaz’s Brain,” *Fractal Fract.*, vol. 6, no. 7, Art. no. 396, July 2022, doi: 10.3390/fractalfract6070396.

[3] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[4] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and *et al*., “The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping,” *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.

[5] M. R. Tomaszewski and R. J. Gillies, “The Biological Meaning of Radiomic Features,” *Radiology*, vol. 298, no. 3, pp. 505–516, Mar. 2021, doi: 10.1148/radiol.2021202553.

[6] J. D. Shur, S. J. Doran, S. Kumar, D. ap Dafydd, K. Downey, J. P. B. O’Connor, N. Papanikolaou, C. Messiou, D.-M. Koh, and M. R. Orton,, “Radiomics in Oncology: A Practical Guide,” *RadioGraphics*, vol. 41, no. 6, pp. 1717–1732, Oct. 2021, doi: 10.1148/rg.2021210037.

[7] E. P. Huang, J. P. B. O’Connor, L. M. McShane, M. L. Giger, and *et al*., “Criteria for the translation of radiomics into clinically useful tests,” *Nat. Rev. Clin. Oncol.*, vol. 20, no. 2, pp. 69–82, Feb. 2023, doi: 10.1038/s41571-022-00707-0.

[8] G. S. Collins, K. G. M. Moons, P. Dhiman, and et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, Apr. 2024, doi: 10.1136/bmj-2023-078378.

[9] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, Art. no. 91, Feb. 2006, doi: 10.1186/1471-2105-7-91.

[10] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiol. Artif. Intell.*, vol. 5, no. 4, Art. no. e220232, July 2023, doi: 10.1148/ryai.220232.

[11] A. Demircioğlu, "Applying oversampling before cross-validation will lead to high bias in radiomics," *Sci. Rep.*, vol. 14, no. 1, Art. no. 11563, May 2024, doi: 10.1038/s41598-024-62585-z.

[12] L. Zou, Y.-H. Choi, L. Guizzetti, D. Shu, J. Zou, and G. Zou, "Extending the DeLong algorithm for comparing areas under correlated receiver operating characteristic curves with missing data," *Stat. Med.*, vol. 43, no. 21, pp. 4148–4162, Sep. 2024, doi: 10.1002/sim.10172.

[13] W. Zhang, Y. Guo, and Q. Jin, "Radiomics and Its Feature Selection: A Review," *Symmetry*, vol. 15, no. 10, Art. no. 1834, Sep. 2023, doi: 10.3390/sym15101834