Detection of brain area with Focal Cortical Dysplasia

Kashuk Ekaterina, Butylin Andrey
GitHub link: https://github.com/AndrewBioChem/NEUROML_FCD

1. Background

Focal cortical dysplasia (FCD) is one of the most common causes of drug-resistant epilepsy. However, its diagnosis remains extremely difficult, particularly in mild or "MRI-negative" cases. Traditional visual assessment methods, such as T1w, T2w, and FLAIR, strongly depend on the radiologist's experience and often fail to detect subtle microstructural changes.

2. Literature review:

- **Chen et al. A radiomics nomogram based on multiparametric MRI for diagnosing focal cortical dysplasia and initially identifying laterality. 2024.**

43 histologically confirmed FCD patients were examined. Authors developed a radiomics nomogram of multiparametric MRI using logistic regression. The inclusion of features of three MR contrasts enhanced ROC-AUC and the authors described correlations among selected radiomic signatures with disease duration and lateralization information at first presentation.

- **Zijun et al. Multi-level Fusion of FDG PET and MRI for Automated Epileptic Lesion Detection. 2024.**

The paper proposes a multi-level fusion of FDG-PET and T1 MRI radiomics (93 features) comparing separate and merged-image strategies to exploit PET's focal metabolic localization and structural detail of MRI. The fused approach improved automated lesion localization.

3. Task

The main task was to detect FCD regions on MRI-derived radiomics data using ML classifiers. The objective was to develop a model with **zero false negatives patients** (no missed patients) during cross-validation (CV), and to minimize false positives (FP).

4. Data

The dataset contained **region-wise radiomics features** derived from several MRI modalities (T1, T2, FLAIR). Each region was labeled as either *FCD-positive* or *control*. MRI-derived radiomics dataset included 279 features extracted from T1-weighted,

T2-weighted and FLAIR images across 168 subjects. The dataset contained both healthy and FCD-affected 66 brain regions per subject. Each patient had several labeled regions.

5. Explorative Data Analysis

The composition of the dataset, feature distributions, and class balance were explored. Statistically significant radiomic differences between FCD and control tissue were identified through statistical testing (Mann-Whitney U with FDR correction). The effect size approximated group differences in size and direction. Top features were ranked by adjusted p-values and effect sizes.

Dataset information:

- Total samples: 11,088
    - Healthy samples: 10,660
    - FCD samples: 428
- FCD prevalence: 3.9%
- Unique subjects: 168
- Unique brain areas: 66
- Total MRI features: 279

Top features by significance and effect size:

1. **original_gldm_DependenceVariance**: p_adj = 9.89e-17, effect = 0.253 (Higher in FCD)
2. **original_gldm_DependenceNonUniformityNormalized**: p_adj = 1.01e-14, effect = -0.236 (Lower in FCD)
3. **original_gldm_LargeDependenceHighGrayLevelEmphasis_T1**: p_adj = 9.88e-14, effect = 0.227 (Higher in FCD)
4. **original_gldm_LargeDependenceEmphasis**: p_adj = 1.25e-13, effect = 0.225 (Higher in FCD)

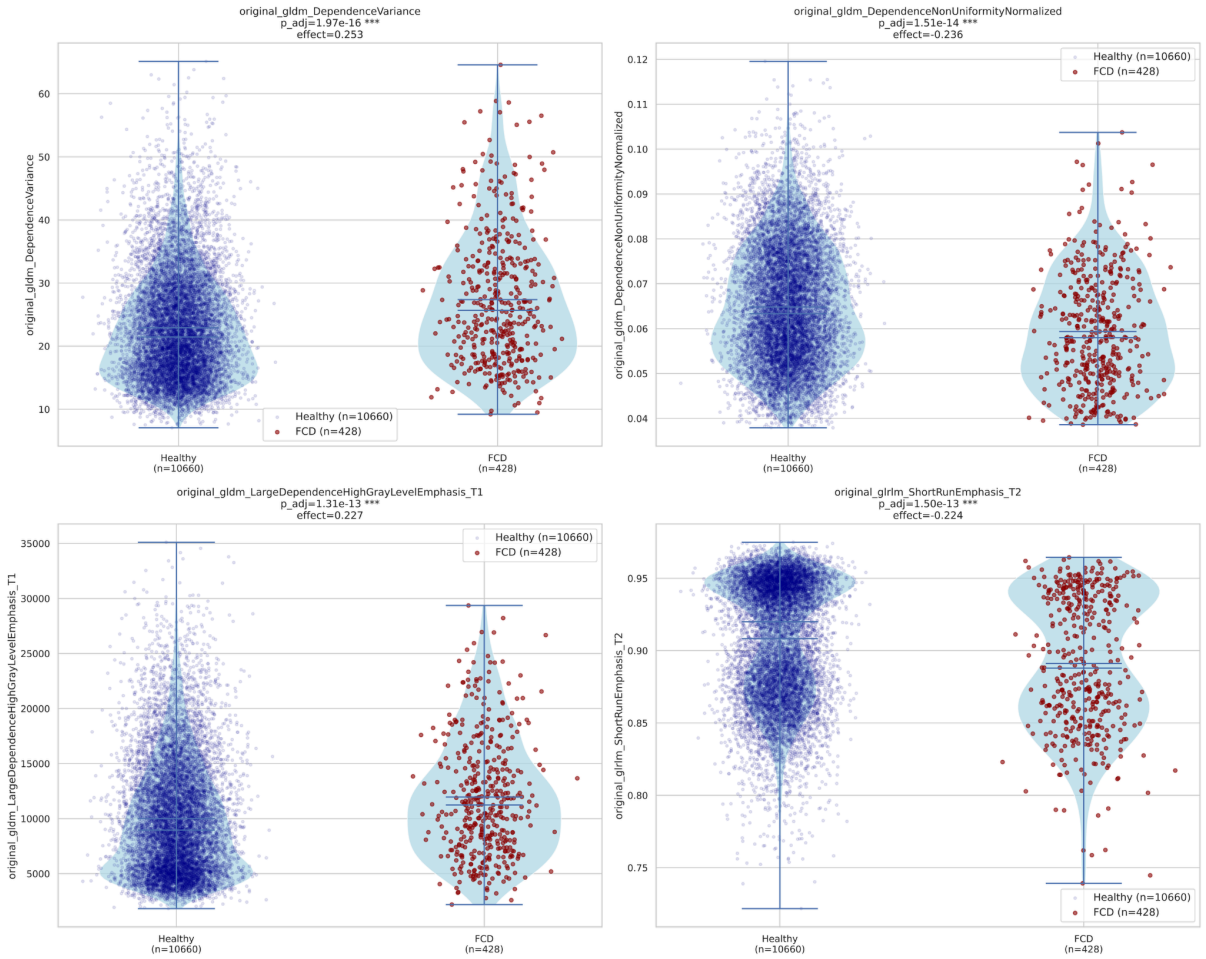


Fig. 1. *Distributions of the top 4 radiomics features differentiating FCD from healthy tissue. Violin plots illustrate feature value distributions across all samples, with overlaid individual data points for healthy (blue) and FCD (red) regions. Each subplot displays adjusted p-value (FDR-corrected) and effect size (rank-biserial correlation).*

6. Reasoning

Given the small number of positive samples per patient and strong imbalance, we have chosen models with interpretable behavior and robust regularization:

1. **Patient-level separation:** no patient data leakage across folds
2. **False Negatives per patient equal 0** during CV
3. **Minimization of False Positives:** select model with lowest amount of FP given zero patient FN.

7. Training and optimization

We tested the following models: Logistic Regression, SVM (linear), KNN, Random Forest, Gradient Boosting and XGBoost.

Cross-validation and model selection were organized to be strictly patient-wise. We held out 30% of subjects as a final test set (no overlap with training set) to evaluate generalization; the remaining subjects were used for model selection with a GroupKFold (`n_splits=5`) so that every fold contained only whole patients and no region from the same subject appeared in both training and validation sets.

Within each fold we fitted a `StandardScaler` on the training split only (to avoid data leakage) and used that fitted scaler to transform both the training and validation sets. Because the dataset is highly imbalanced (many more healthy regions than FCD), we used `RandomUnderSampler` on the training portion of each fold to rebalance classes for model fitting. The sampler was trained and applied only to the training partition (not to test) to avoid data leakage. We also used SCOPE but it was computationally expensive and demonstrated results similar to `RandomUnderSampler.`

For clinical sensitivity we optimized a decision threshold per fold using the rule `patient_FN == 0` (every positive patient must have ≥1 detected region on that training fold) and among thresholds that satisfy that constraint, we selected the one that minimized fold false positives: the per-fold thresholds were averaged and used for the final evaluation.

We used the following hyperparameters:

- Logistic Regression `C = [0.01, 0.1, 1.0]` with both `penalty = ['l1','l2']` (solver `saga` used for L1),
- SVM with `C = [0.01, 0.1, 1.0]` and `kernel ∈ {'linear','rbf'}` plus `gamma ∈ {'scale','auto'}` for RBF,
- KNN with `n_neighbors ∈ [10,30,50]` and `p ∈ {1,2}` for Manhattan/Euclidean distances and `weights='distance'` option tested
- Gradient Boosting with `loss ∈ {'log_loss', 'exponential'}` `learning_rate = [0.1, 1, 10, 100]`, `n_estimators = [1, 10, 50, 100]`, `subsample = [0.01, 0.1]` and `criterion ∈ {'friedman_mse', 'squared_error'}`
- Random Forest with `n_estimators = [3, 5, 7, 10, 20, 30, 50, 100, 200]`, `criterion ∈ ['gini', 'entropy', 'log_loss']`, `max_depth = [None, 5, 20, 100]`, `max_features ∈ ['sqrt', 'log2', None]`, `class_weight ∈ ['balanced', 'balanced_subsample']`
- XGBoost with `n_estimators ∈ [100,300]`, `max_depth ∈ [3,5]`, and `learning_rate ∈ [0.01,0.1]`

All grid searches were evaluated using the grouped folds and the clinical selection rule (patient_FN==0, then minimize average FP across folds) to prioritize patient-level sensitivity.

8. Results: Model-by-model summaries

Models' hyperparameters were selected with patient-wise GroupKFold cross-validation. During CV we optimized decision thresholds per fold to enforce the patient-level rule (patient_FN == 0) and then averaged those fold thresholds to obtain the final threshold. After selecting the best hyperparameters, each model was retrained on the entire training set (using the same preprocessing used in CV: scaler fit on full train and sampling applied only during model fitting) to produce the final classifier. That final classifier and the CV-derived threshold were applied to the held-out test set (30% of patients) to obtain the test results.

Table 1. *Performance comparison of six machine learning models for FCD detection. For each model, the best hyperparameters are reported along with average train-set specificity, region-level recall, and patient-level recall. Test-set performance includes specificity, region-level recall, patient-level recall, and AUC.*

Model	**Best parameters**	**Average specificity on train set**	**Average recall (region) on train test**	**Average patient recall on train set**	**Specificity on test set**	**Recall (region) on test set**	**Patient recall on test set**	**AUC**
Logistic Regression	{'C':0.01,'penalty':'l2'}	0.4059	0.9058	1	0.380	0.904	0.941	0.727
SVM (linear)	{'C':0.01,'kernel':'linear','gamma':'scale'}	0.4004	0.9217	1	0.364	0.898	0.922	0.724
KNN	{'n_neighbors':50,'weights':'distance','p':2}	0.3825	0.8623	1	0.418	0.752	0.922	0.629
Random Forest	{'n_estimators':200,'criterion':'entropy','max_features':'sqrt','class_weight':'balanced_subsample'}	0.3872	0.9061	1	0.379	0.873	0.922	0.72
Gradient Boosting	{'n_estimators':300,'max_depth':3,'learning_rate':0.01}	0.4244	0.8515	1	0.404	0.752	0.863	0.65
XGBoost	{'n_estimators':100,'max_depth':3,'learning_rate':0.05}	0.9071	0.3444	1	0.315	0.866	0.922	0.701

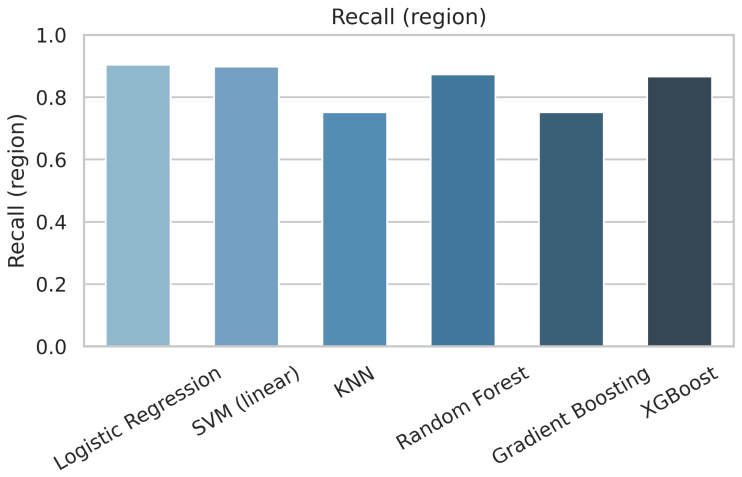


Fig.2. *Region-level recall on test set across ML models.*

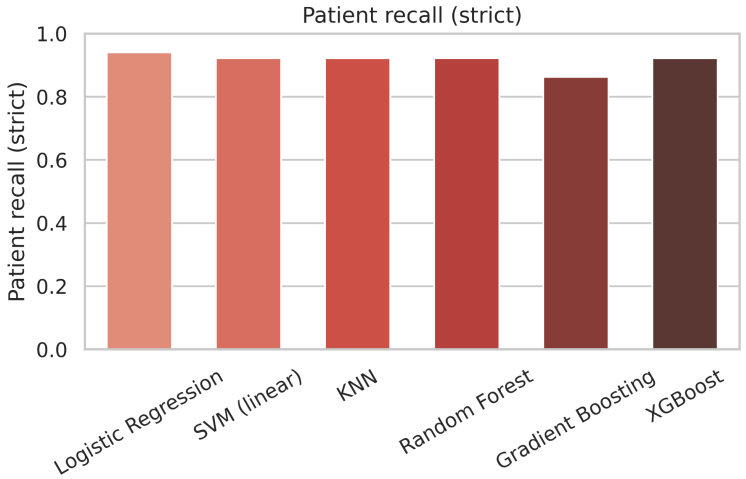


Fig.3. *Patient-level recall on test set across ML models.*

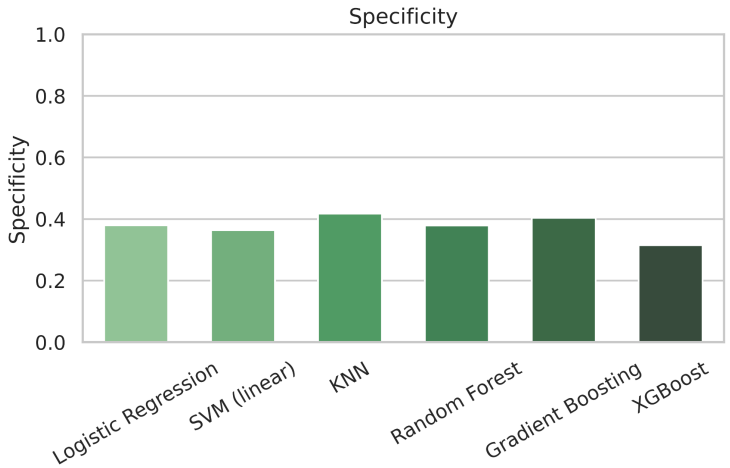


Fig.4. *Specificity on test set across ML models.*

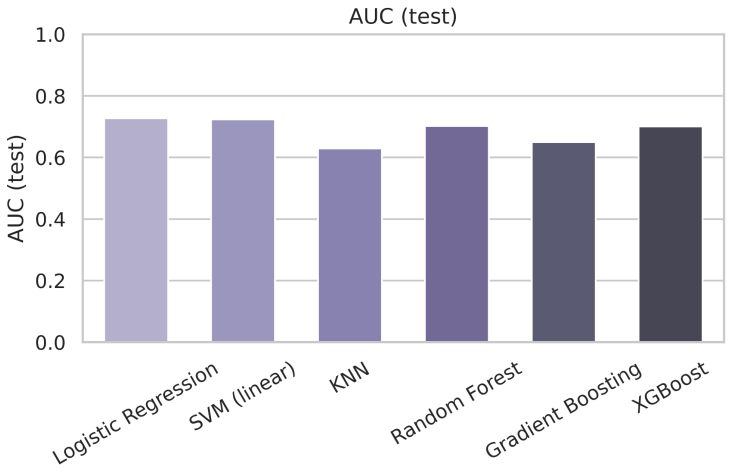


Fig.5. *AUC score on test set across ML models.*

Across six models, patient-wise cross-validation and test evaluation showed that all models achieved perfect average patient-level recall on training folds, confirming sensitivity to FCD-positive subjects. **Logistic Regression** and **linear SVM** provided high region-level recall (~0.90) and strong patient-level recall on test (0.941 and 0.922) with moderate test specificity (0.380–0.364) and AUC ~0.72–0.73. **KNN** and **Gradient Boosting** were less region-sensitive (recall ~0.75–0.85) but maintained patient recall (0.922 and 0.863) with slightly higher test specificity (~0.418–0.404). Ensemble tree models such as **Random Forest** and **XGBoost** showed balanced region recall (~0.87 and 0.866), patient recall (0.922), and test specificity (~0.379–0.315), with AUCs of 0.72 and 0.701.

Overall, **Logistic Regression, SVM, and Random Forest** demonstrated the best combination of sensitivity, patient-level recall, region-level recall, and overall discriminative performance.

9. Important features

Feature importance was determined using two approaches: model coefficients and permutation importance.

For **linear models** such as Logistic Regression and linear SVM, the magnitude of the standardized coefficients was used to assess the contribution of each feature to the model's decision boundary. Features with larger absolute coefficient values were considered more influential in distinguishing between healthy and FCD regions.

For **nonlinear and distance-based models** (KNN), where no intrinsic importance measure exists, **permutation importance** was used. In this approach, each feature was randomly shuffled, and the resulting drop in model performance (AUC) assessed its importance.

For **tree-based ensemble models** (Random Forest, Gradient Boosting, and XGBoost) importance was computed from each model's **internal split-based metrics** (using the `.feature_importances_`). This method quantifies how frequently and effectively a feature reduces impurity (classification error) across all trees in the ensemble.

Logistic Regression:

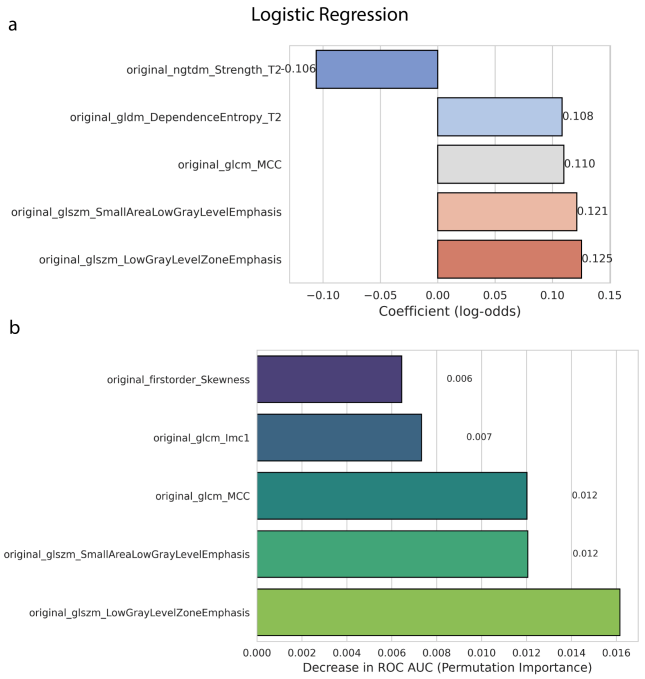


Fig.6. *Top radiomic features for Logistic Regression.* (a) Ranked by model coefficients. (b) Ranked by permutation importance (impact on AUC when shuffled).

KNN:

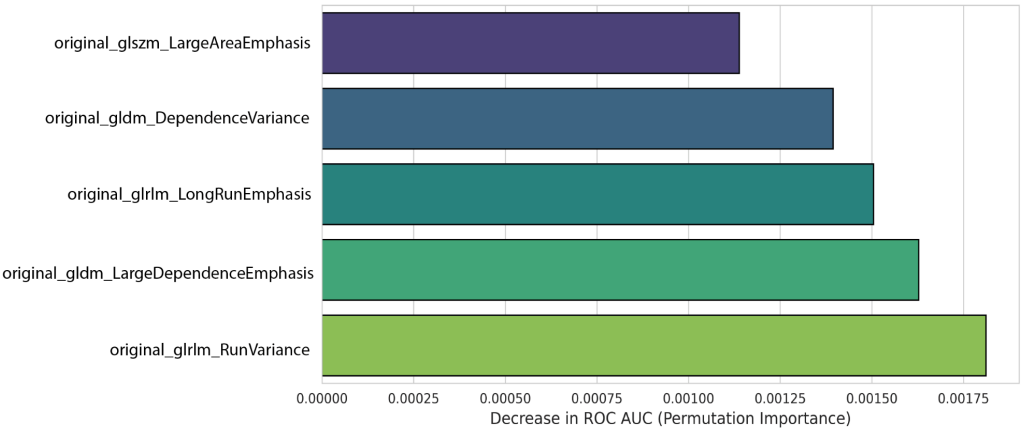


Fig.7. *Top radiomic features for KNN ranked by permutation importance (impact on AUC when shuffled).*

SVM:

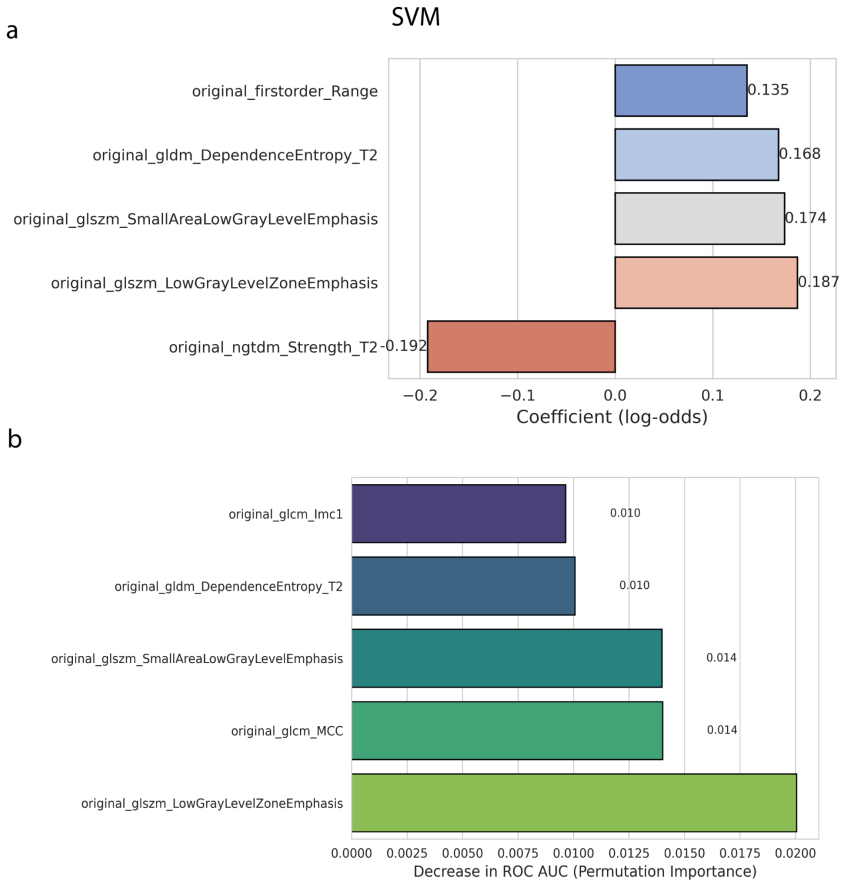
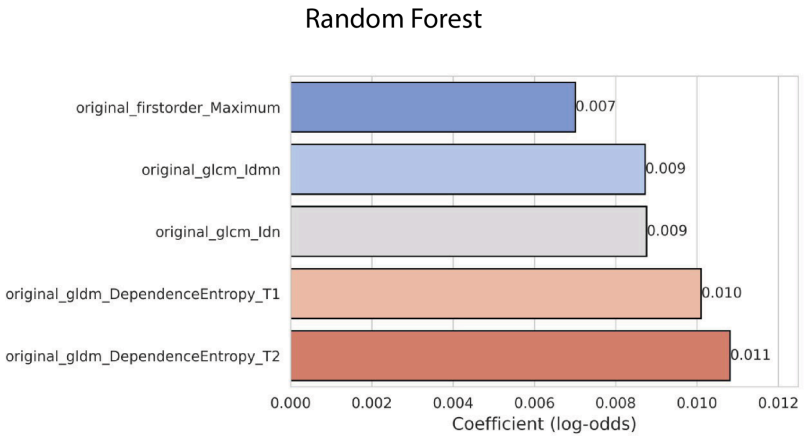


Fig.8. *Top radiomic features for SVM.* (a) Ranked by model coefficients. (b) Ranked by permutation importance (impact on AUC when shuffled).

Random Forest:

a



b

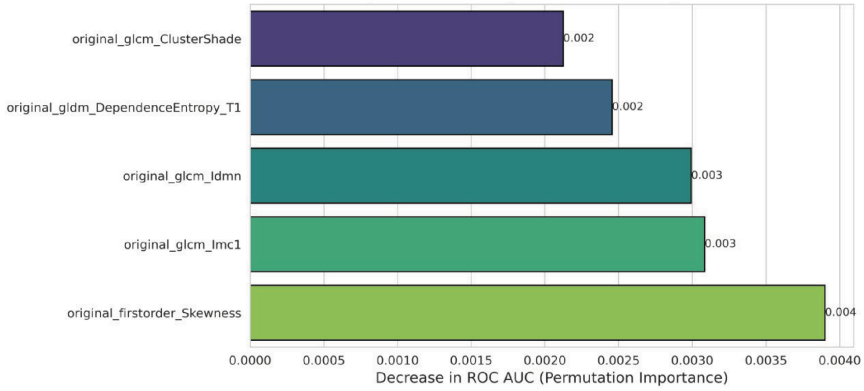


Fig.8. *Top radiomic features for Random Forest.* (a) Ranked by internal feature importance. (b) Ranked by permutation importance (impact on AUC when shuffled).

Gradient Boosting:

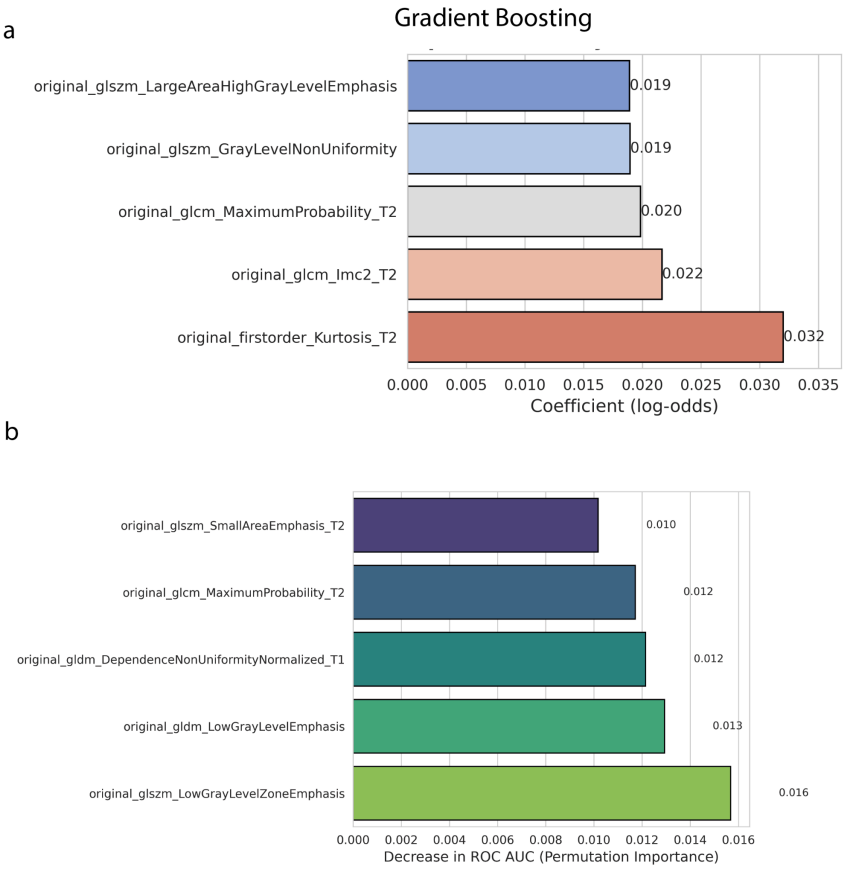


Fig.9. *Top radiomic features for Gradient Boosting.* (a) Ranked by internal feature importance. (b) Ranked by permutation importance (impact on AUC when shuffled).

XGBoost:

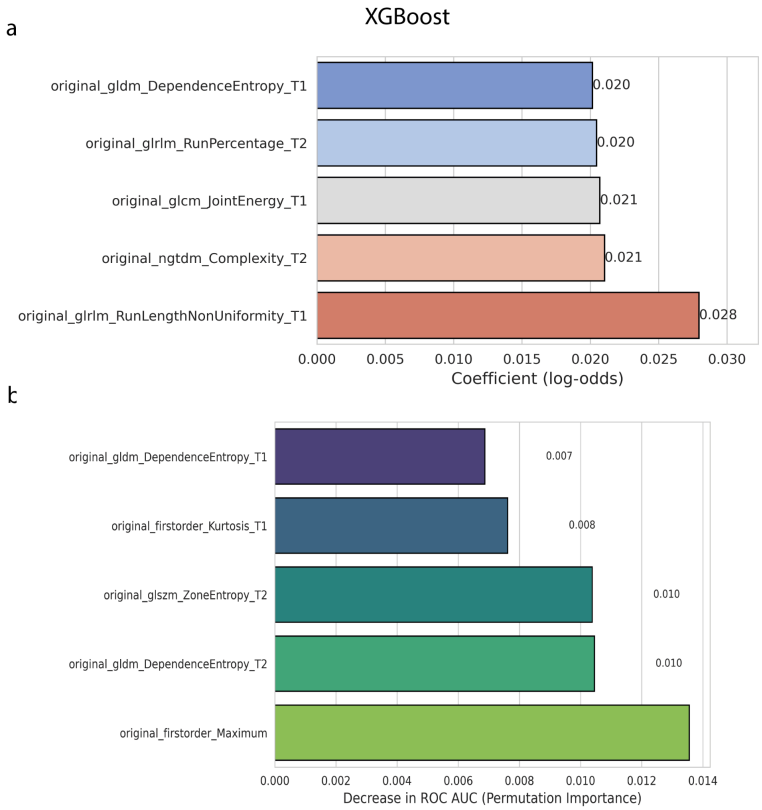


Fig.10. *Top radiomic features for XGboost.* (a) Ranked by internal feature importance. (b) Ranked by permutation importance (impact on AUC when shuffled).

Table 2. Key radiomic features identified across models.
Most informative features were consistently detected across several models using model coefficients, permutation importance, and built-in importance scores from tree-based models.

Feature	Appears in (models / lists)	Description
original_glszm_LowGrayLevelZoneEmphasis	LR (coef, perm), SVM (coef/perm), GB perm	measures the predominance of low-intensity homogeneous zones, reflecting subtle hypointense patches in dysplastic tissue
original_glszm_SmallAreaLowGrayLevelEmphasis	LR (coef, perm), SVM (coef/perm), GB perm	emphasizes small low-intensity areas, indicating microregions of cortical irregularity or gliosis
original_glcm_MCC	LR (coef, perm), SVM perm	assesses the balance of grey-level co-occurrences, with changes suggesting disrupted spatial organization in voxel intensities
original_gldm_DependenceEntropy_T2	LR (coef, perm), XGB perm, RF imp	quantifies randomness of neighborhood dependencies on T2, capturing textural heterogeneity in dysplastic cortex

original_ngtdm_Strength_T2	LR coef, SVM coef	measures the contrast between voxels and their neighbors, where lower values indicate smoother, less distinct cortical transitions
original_gldm_DependenceVariance	KNN perm	captures variability in dependence cluster sizes, reflecting patchy and irregular tissue architecture
original_glrlm_RunLengthNonUniformity_T1	XGBoost gain, SelectKBest	quantifies variability in consecutive voxel run lengths, indicating heterogeneous texture patterns
original_firstorder_Maximum	RF imp, XGB perm	represents the highest voxel intensity, highlighting focal signal changes associated with cortical abnormalities

We further visualized these features to verify and illustrate the univariate differences suggested by statistical tests: the plots show group separation, medians, spread and outliers, and include FDR-adjusted p-values, effect sizes and sample counts for context.

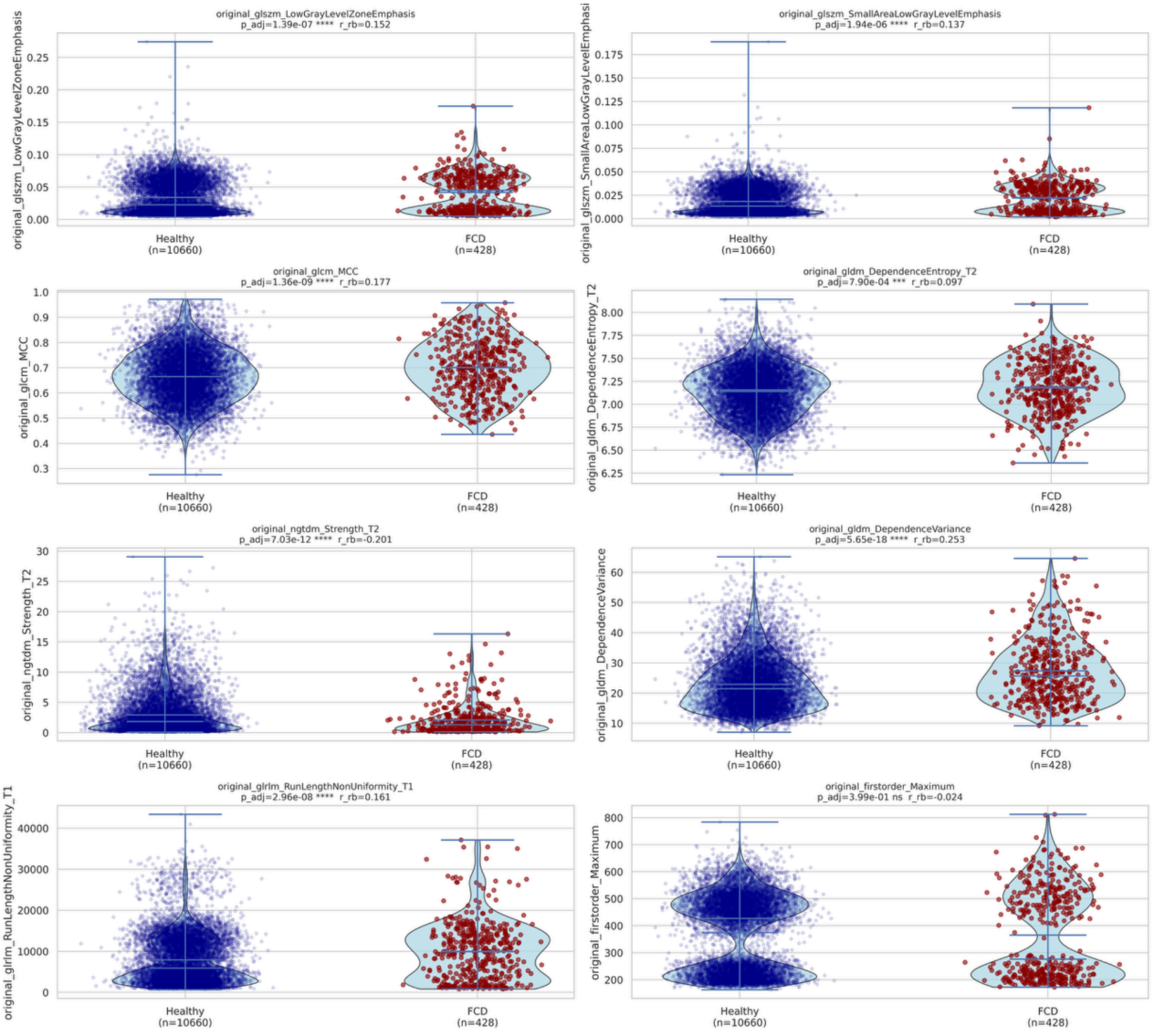


Fig.11. *Distributions of the eight top radiomics features in healthy vs FCD regions (violin plots with jittered points). Each subplot shows the FDR-adjusted p-value (Mann–Whitney U) and the rank-biserial effect size (r_rb); blue = healthy, red = FCD.*

10. Interpretation of the results

We can conclude that FCD regions are characterized primarily by microarchitectural heterogeneity and subtle textural contrast rather than by outlier single-voxel intensities.

Specifically, GLSZM measures (**original_glszm_LowGrayLevelZoneEmphasis**, **original_glszm_SmallAreaLowGrayLevelEmphasis**) are significantly bigger in FCD (medians ≈0.045 vs. 0.023 and 0.022 vs. 0.013 respectively, p_adj < 0.001), which may correspond to more common low-intensity homogeneous small zones representing disturbed cortical lamination.

original_gldm_DependenceVariance shows the strongest influence (r_rb ≈ 0.253, p_adj < 1e-16) and indicates that areas of FCD contain more diverse local sets of equivalent voxels (for example, some of them are huge patches of homogeneous tissue and others are isolated patches) which reflects the patchy, disordered tissue structure present in dysplasia.

original_glcm_MCC is elevated in FCD (r_rb ≈ 0.177, p_adj < 1e-9), which demonstrates that the tissue has a strong, predictable spatial correlation between voxel intensities (the dysplasia is not randomly chaotic but contains large-scale, coherent abnormal patterns (e.g., abnormal neuronal tracts, blurred zones).

original_gldm_DependenceEntropy_T2 is slightly larger in FCD (r_rb ≈ 0.097, p_adj < 0.001), showing increased randomness of local similarity at T2.

original_ngtdm_Strength_T2 is smaller (r_rb ≈ −0.201, p_adj < 1e-11), showing less abrupt changes between neighbors and voxels, as anticipated by blurred cortex.
original_glrlm_RunLengthNonUniformity_T1 is larger (r_rb ≈ 0.161, p_adj < 1e-8), showing patchy tissue disruption where areas of severe disorganization (short runs) coexist with zones of aberrant homogeneity (long runs).

Following features were significantly different between FCD and healthy regions only by the Mann-Whitney U test:

- **original_gldm_DependenceVariance** (increased in FCD; effect ≈ 0.253, p_adj = 9.89e-17): captures increased local cluster size variability consistent with patchy cortical organization
- **original_gldm_DependenceNonUniformityNormalized** (decreased in FCD; effect ≈ −0.236, p_adj = 1.01e-14): captures reduced neighborhood dependency uniformity, consistent with heterogeneity of tissue
- **original_gldm_LargeDependenceHighGrayLevelEmphasis_T1** (increased in FCD; effect ≈ 0.227, p_adj = 9.88e-14): indicates large, high-intensity homogeneous patches on T1, perhaps corresponding to structural focal abnormalities or gliosis
- **original_gldm_LargeDependenceEmphasis_T2** (decreased in FCD; effect ≈ -0.225, p_adj = 1.25e-13): highlights FCD tissue is structurally more heterogeneous and disorganized, consistent with loss of lamination under FCD.

To conclude, FCD regions are marked by **micro-architectural heterogeneity and subtle textural contrasts** rather than single-voxel outliers. Key features such as **LowGrayLevelZoneEmphasis** and **SmallAreaLowGrayLevelEmphasis** are larger in FCD, reflecting small homogeneous patches. **DependenceVariance** shows the strongest effect, indicating variable local clusters, while **GLCM_MCC**, **DependenceEntropy_T2**, **Strength_T2**, and **RunLengthNonUniformity_T1** capture disrupted spatial patterns, increased randomness, and blurred cortex.

In our results most of the important features came from FLAIR-type contrasts, which is consistent with clinical practice (since radiologists most often detect FCD abnormalities on FLAIR, where cortical signal changes and gray-white blurring are most visible).

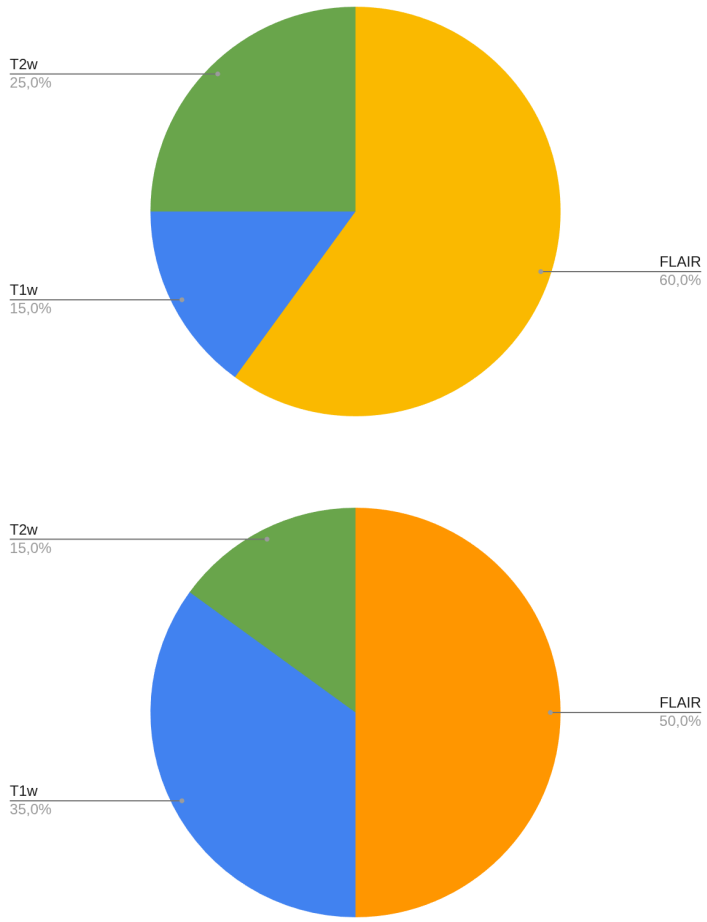


Fig.11. *Distributions of modalities (T1w, T2w, FLAIR) for the top 20 radiomics features for Logistic Regression (top) and Random Forest (bottom).*

11. What else could be done

Advanced models, including deep learning, could reveal complex patterns beyond classical radiomics. Feature interactions (for example, via SHAP) and clinical correlations could be explored.