FLAIR Dependence Confounds Cross-Modal Consistency

A Summary of Findings and Analysis Pipeline: BraTS-Africa Glioma Segmentation (Paper 2)

Ofile Seneo Mfetane et al. / SPARK Academy 2026 (CAMERA Africa) / MIRASOL Workshop, MICCAI 2026

Overview

This report summarises the full Paper 2 pipeline: from the Week 3 SPARK Academy development notebook (leave-one-out cross-modal consistency as a label-free reliability signal for glioma segmentation) through to the final statistical analysis reported in the MIRASOL 2026 submission. The notebook builds the SegResNet segmentation backbone, generates leave-one-out (LOO) predictions, and produces the consistency features and figures; the paper adds the rigorous statistical testing (Spearman, bootstrap, Steiger's test, LOOCV classification) that determines whether the notebook's headline correlation actually holds up.

Research Question

Can disagreement between leave-one-out predictions, generated by removing one MRI modality at a time from an otherwise complete four-modality input, serve as a label-free proxy for segmentation reliability on BraTS-Africa?

Method Summary

Component	Detail
Dataset	BraTS-Africa, 95 cases (76 train / 19 validation, seed = 42); T1n, T1c, T2w, FLAIR at 240x240x155
Backbone	SegResNet (MONAI), ~4.7M params, 4-channel in/out (background, NETC, SNFH, ET)
Training	DiceFocalLoss (class weights 0.1/3.0/1.0/4.0), AdamW (lr=1e-4), cosine annealing, 20-epoch warmup, 300 epoch
LOO design	5 inference passes per case: full-modality + 4 single-channel-zeroed passes (no-T1n, no-T1c, no-T2w, no-FLAIR)
Corrected flaw	An earlier single-modality-only masking approach was rejected as catastrophically out-of-distribution; LOO (3-of-4
Consistency features	Mean LOO agreement, minimum LOO agreement, volume instability (CV), LOO instability (1 minus mean agreem
Triage rule	Accept (min-agree >= 0.75) / Radiologist Review (0.45-0.75) / Repeat Acquisition (< 0.45)
Statistics	Pearson + Spearman correlation, 5,000-resample bootstrap CIs, Steiger's dependent-correlations test, LOOCV lo

Segmentation Performance

The SegResNet checkpoint reached a best validation mean subregion DSC of 0.787 (epoch 250; NETC = 0.687, SNFH = 0.831, ET = 0.844), well above zero-shot nnU-Net transfer (DSC = 0.310). Per-case DSC ranged from 0.210 to 0.966 across the 19 validation cases, with 15/19 (79%) classified Accept and 4/19 (21%) classified Requires Review under the 0.75 threshold. Recurring failure modes were absent ET predictions (weak gadolinium enhancement), omission of small NETC regions, and over-extension of SNFH boundaries.

Central Finding: FLAIR Dependence as a Confound

Removing FLAIR was the single most disruptive LOO pass in 14 of 19 cases (74%), occurring across the full performance range, including the best-performing case (DSC = 0.966). This indicates the model's sensitivity to FLAIR removal reflects architectural dependence, SNFH (the largest subregion) is defined by FLAIR signal, rather than case-specific scan quality. As a direct consequence, the standard LOO consistency signal showed no significant relationship with segmentation quality (Pearson r = 0.240, p = 0.322).

Correlation of Consistency Features with Segmentation Quality (n = 19)

Metric	Pearson r	p	Bootstrap 95% CI	Spearman rho	p
Standard, min LOO agreement	0.240	0.322	[-0.137, 0.561]	0.342	0.152
Standard, mean LOO agreement	0.648	0.003	[-0.295, 0.912]	0.330	0.168
T1-family mean agreement (post hoc)	0.660	0.002	[-0.332, 0.927]	0.202	0.408
T1-family min agreement (post hoc)	0.521	0.022	[-0.269, 0.844]	0.142	0.562

Excluding FLAIR and computing a T1-family consistency metric raised the Pearson correlation to r = 0.660, but this gain does not survive scrutiny: the Spearman correlation is near zero (rho = 0.202), every bootstrap CI spans zero, and Steiger's test comparing the T1-family metric to standard minimum agreement is non-significant (z = 1.644, p = 0.100). The Pearson gain is driven by a single catastrophic-failure case (DSC = 0.210), not a robust monotonic trend.

Binary Reliability Classification (LOOCV Logistic Regression, n = 19)

Model	LOOCV AUC	Bootstrap Mean	95% CI	Accuracy
Standard LOO features	0.350	0.354	[0.000, 0.667]	0.474
Extended with T1-family features	0.350	0.354	[0.000, 0.667]	0.526

Both classifiers scored below the 0.500 random baseline. The extended model had high recall (0.75) but low precision (0.27) for the requires-review class, indicating unstable rather than deployment-ready discrimination.

Interpretation

Standard cross-modal LOO consistency does not provide a usable label-free reliability signal for this SegResNet configuration on BraTS-Africa. The clearest, well-substantiated result is the FLAIR dependence confound itself, not the T1-family workaround. Naive consistency conflates learned, architecturally correct modality reliance with genuine scan-quality degradation, and cannot tell the two apart without further correction.

Three design principles follow for future consistency-based reliability work:

- Modality dependence must be characterised for a given model before consistency is trusted as a reliability signal.
- Continuous reliability scoring is preferable to binary triage while discriminative power remains this limited.
- Per-subregion, modality-aware consistency metrics are likely more clinically interpretable than a single whole-tumour scalar.

The most-disruptive-modality indicator remains directly actionable regardless of the scalar's weakness: no-T1c dominance implicates contrast timing or dosing for ET review, no-FLAIR dominance flags SNFH boundary risk, and no-T1n dominance flags NETC fragmentation risk from INU artefacts.

Limitations

- All analyses are underpowered at n = 19; every bootstrap CI spans zero, so results are hypothesis-generating, not confirmatory.
- The T1-family metric was defined post hoc after observing the FLAIR confound in this same validation set, carrying overfitting risk; it needs independent-cohort validation.
- LOO masking zeros channels in artificially complete acquisitions; this may not generalise to real missing or corrupted modalities.
- Findings are restricted to one architecture and configuration (SegResNet with the stated loss and class weights); generalisation to nnU-Net, UNETR, or SwinUNETR is untested.
- Triage thresholds (Accept / Review / Repeat Acquisition) are clinically unvalidated.

Notebook Pipeline: Deliverables and Outputs

The Week 3 SPARK Academy notebook ("Cross-Modal Consistency as a Post-Inference Reliability Signal") implements the full pipeline behind the paper: dataset preprocessing, SegResNet training under two hyperparameter configurations, LOO inference, consistency feature extraction, triage assignment, and figure and table export.

Hyperparameter configurations tested

Setting	Config A (Baseline)	Config B (Extended, production)
Max epochs	150	300
LR warmup	10 epochs	20 epochs
Patches/volume	2	4
ET class weight	3.0	4.0

Dropout	0.1	0.2
Augmentation	Flip, rotate	+ Gaussian noise, contrast jitter
Best validation Dice	~0.71	~0.797

Config B improved mean validation Dice by roughly 0.09 over Config A and substantially reduced ET collapse cases, consistent with its higher ET class weight and doubled patch sampling; it was carried forward as the production checkpoint.

A documented methodological correction

An earlier version of the pipeline tested single-modality-only inference (zeroing three channels, keeping one). This was rejected as methodologically unsound: a four-channel-trained SegResNet given only one active channel is severely out-of-distribution, so the resulting predictions reflect model confusion rather than a meaningful modality-specific reliability signal. The corrected leave-one-out design (three of four channels active) keeps inputs much closer to the training distribution while still probing per-modality dependence.

Exported figures and tables

File	Content
fig1_sample_multimodal.png	Sample multimodal BraTS-Africa case
fig2_training_curves.png	Config B training curves
fig3_best/borderline/worst_overlay.png	Segmentation overlays at three quality tiers
fig4_loo_agreement.png	LOO agreement bar chart and scatter
fig5_per_modality_disruption.png	Per-modality disruption distribution
fig6_triage_summary.png	Triage distribution and confusion matrix
fig7_loo_disruption_overlay.png	Most-disrupted case LOO overlay
fig8_workflow_diagram.png	Pipeline architecture and workflow diagram
table1_per_case_dice.csv	Per-case validation Dice
table2_loo_consistency_features.csv	LOO consistency features and triage labels
loo_feature_importance.csv	Random forest / logistic regression feature importance

Clinical framing (from the notebook's discussion)

The pipeline is framed for SSA clinical settings where a second radiologist, planning-system ground truth, or scanner-console QC metrics are often unavailable. LOO consistency requires only four additional inference passes (roughly 2 to 4 minutes on a GPU) and no extra labels or training, which the notebook argues is negligible next to the cost of an undetected failure entering a clinical pipeline, though the statistical results above show the current scalar signal is not yet reliable enough to act on alone.

Bottom Line

Standard cross-modal LOO consistency is confounded by FLAIR architectural dependence and does not reliably track segmentation quality on BraTS-Africa in its current form; no deployment claim is made. The

FLAIR-dependence finding itself, and the design principles it motivates, are the substantiated contribution of this work.