Decision Stability Under Segmentation Uncertainty in CT Radiomics

1st Jakub Ciszewski
Warsaw University of Technology
Warsaw, Poland
jakub.ciszewski.dokt@pw.edu.pl
ORCID: 0009-0007-9930-0538

2nd Radosław Roszczyk
Warsaw University of Technology
Warsaw, Poland
radoslaw.roszczyk@pw.edu.pl
ORCID: 0000-0003-0721-6739

3rd Krzysztof Siwek
Warsaw University of Technology
Warsaw, Poland
krzysztof.siwek@pw.edu.pl
ORCID: 0000-0003-2642-2319

***Abstract*—Segmentation uncertainty propagates through a computational imaging pipeline, from region delineation through feature extraction to classifier output, and is conventionally quantified only at the intermediate stage: a descriptor is deemed reliable when its intraclass correlation coefficient (ICC) survives perturbation of the segmentation. A deployed model, however, does not report a descriptor, it reports a class. We investigate whether numerical reproducibility of the intermediate features implies reproducibility of the final computational decision. Using 53 patients from the public Adrenal-ACC-Ki67-Seg collection, we generate six deterministic contour variants per tumour by physical-distance morphology, add ten random shape-changing deformations per patient, and extract 107 features from each of 901 masks. Feature stability is high: 73 of 107 features reach ICC ≥ 0.85, with a median of 0.942. Decision stability is not. Under patient level inference with stratified bootstrap intervals, the reference model changes the predicted class for 7.5% $[5.3, 9.9]$ of alternative contours under margin shifts and 13.4% $[10.3, 16.7]$ under shape deformations, affecting 72% and 83% of patients at least once across the ensemble and one patient in three under a single alternative contour, while the ICC threshold that improves discrimination most produces no stability gain at all. Aggregating predictions over the contour ensemble reduces the flip rate by a factor of 16 to 24, to 0.3% $[0.0, 1.0]$ and 0.8% $[0.3, 1.5]$, at unchanged discrimination, with three contours recovering 78% of the achievable gain. We further show that the metric is gameable: a classifier that stops discriminating drives the flip rate towards zero, which happened three times by accident in this study. This is a controlled methodological study under *simulated* segmentation uncertainty in a single small cohort, not a validation of a clinical Ki-67 model.**

***Index Terms*—radiomics, segmentation uncertainty, decision stability, test-time augmentation, reproducibility, adrenocortical carcinoma**

I. Introduction

A radiomic pipeline is a chain of numerical stages: a region of interest is delineated, resampled and discretised, several hundred descriptors are computed from it, and a classifier maps them to a decision. Uncertainty entering at the first stage propagates through all of them, and delineation is the least reproducible step of the chain [1]. The community response has been a large body of work on *feature* robustness: test-retest imaging, multi-observer segmentation and simulated perturbation, each summarised by an intraclass correlation coefficient (ICC) and used to discard unstable features before modelling [2], [3]. A systematic review of 481 radiomic studies found the ICC to be the dominant reliability metric in the field, applied most often to segmentation variability, with thresholds of 0.75, 0.85 and 0.90 in common use [4], [5]. Reporting checklists have codified the practice: CLEAR asks for a reproducibility assessment, METRICS scores it, and the second Radiomics Quality Score treats robustness as a readiness criterion [6]–[8].

This validates an intermediate quantity of the computation rather than its output. A deployed model does not report a feature value, it reports a probability or a class. Individually stable features can still yield unstable decisions whenever patients sit close to the decision boundary, and small, class-imbalanced cohorts make that regime common rather than exceptional. The converse failure is equally available and, to our knowledge, unremarked: a model that has collapsed to predicting the majority class is perfectly stable and perfectly useless.

Adrenocortical carcinoma (ACC) makes the problem concrete. It is a rare tumour, the Ki-67 proliferation index above 10% marks a high risk of recurrence and is used to select adjuvant treatment [17], [18], and a publicly available collection with paired segmentations and Ki-67 annotation contains 53 patients [20]. Everything that makes decision instability likely is present at once: a small cohort, class imbalance, heterogeneous multi-vendor acquisition and a clinically fixed operating threshold. We deliberately propose no new feature, classifier or perturbation scheme. This is a computational reliability study of an otherwise conventional image-processing and classification pipeline, examined along the axis that decides whether its output can be acted on for an individual patient.

A. Contributions

- We quantify feature robustness and *decision* robustness inside one perturbation framework and show that they dissociate: 68% of features pass ICC ≥ 0.85, yet a single simulated alternative contour is expected to change the predicted class for one patient in three (17.9 of 53).
- We show that ICC-based filtering does not buy decision stability at the thresholds most commonly used, and we reconcile this with the opposite result in the literature by measuring both quantities on the same data.