UNDER REVIEW

RESCIENCE C

Replication / Medical imaging

[¬Re] Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm

Adithya Balakumar[1], **Spursh Desphande**[2], and **Wenhao Lu**[1]

[1]Millburn High School, Millburn, USA – [2]Crescenta Valley High School, La Crescenta, USA

Edited by
(Editor)

Received
–

Published
–

DOI
–

Abstract Buda, Saha and Mazurowski (2019) trained a U-Net to segment lower-grade gliomas in 110 patients from The Cancer Genome Atlas and reported a mean per-patient Dice of 0.82 and a median of 0.85. Their central claim is not the segmentation number but what follows from it: shape features measured on the automatically produced masks remain associated with genomic subtype about as strongly as the same features measured on manual masks, with an AUC of 0.80 against 0.78 for separating RNASeq cluster R2 from the rest. We reimplemented the pipeline from the published description and evaluated it on the same 110 patients under patient-level five-fold cross-validation. Segmentation came close on the median (0.808) and short on the mean (0.707), the gap being driven by six patients the model missed almost completely. On the manual masks the radiogenomic association held: bounding ellipsoid volume ratio against RNASeq cluster gave an uncorrected Fisher p of 0.0005 and the manual-mask discrimination AUC was 0.764, close to the reported 0.78. On our predicted masks it disappeared. No feature-by-subtype pair survived Bonferroni correction and the AUC fell to 0.455. Discarding the badly segmented patients did not recover it, because bounding ellipsoid volume ratio computed on predicted masks is uncorrelated with the same feature computed on manual masks even when segmentation is good ($r = -0.01$ for patients above 0.5 Dice). We report this as a failed replication of the paper's central claim and argue that the shape features are considerably more fragile under segmentation error than the original numbers suggest.

A replication of M. Buda, A. Saha and M. A. Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Computers in Biology and Medicine 109 (2019), 218–225.

Correspondence should be addressed to Wenhao Lu (wenhao@nxthorizon.org)
The authors have declared that no competing interests exists.
Code is available at https://github.com/Alscend-Research/lgg-imaging-repro.
Data is available at https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation.

1 Introduction

Buda, Saha and Mazurowski[1] trained a U-Net[2] to segment lower-grade gliomas in 110 patients from The Cancer Genome Atlas and reported a mean per-patient Dice coefficient of 0.82 with a median of 0.85. Taken alone that number is unremarkable, and it is not what the paper is about. The claim that matters is the one built on top of it: three shape features measured on the *automatically* produced tumor masks stay associated with genomic subtype about as strongly as the same features measured on masks a radiologist drew by hand. For the specific task of separating RNASeq cluster R2 from the remaining clusters using the inverse bounding ellipsoid volume ratio, the paper reports an AUC of 0.80 from the model masks against 0.78 from the manual ones.

The association itself was established two years earlier by Mazurowski et al.[3] on the same cohort using manual segmentations. The 2019 paper's contribution is that the manual step can be dropped. That is the step worth checking, and it is also the step most exposed to segmentation quality, since every shape feature is a functional of the mask.

We reimplemented the pipeline from the published description, without using the authors' code, and ran it on the same 110 patients under patient-level five-fold cross-validation. Three things came out of it. The segmentation landed close to the target on the median (0.808 against 0.85) and well short on the mean (0.707 against 0.82), with the gap coming almost entirely from six patients the model missed. The radiogenomic association survived on the manual masks, at strengths close to what the original papers report. And on our predicted masks it vanished: no feature-by-subtype pair survived multiple-comparison correction, and the discrimination AUC fell to 0.455, which is chance.

We label this a failed replication of the paper's central claim. The rest of this article is mostly about the third result, because we think the reason it fails is more interesting than the fact that it does.

2 Methods

2.1 Data

We used the public Kaggle release of the dataset[4], which packages the TCGA-LGG collection[5,6] used in the original work: 110 patients, 3,929 axial slices at 256×256, three input channels (pre-contrast, FLAIR, post-contrast) and one binary FLAIR-abnormality mask per slice. Of those slices, 1,373 contain tumor and 2,556 are empty. Genomic and demographic labels ship with the release in `data.csv`, one row per patient, which is why the radiogenomic half of this replication costs no GPU time at all.

Table 1 gives the breakdown by contributing institution. This is a United States National Cancer Institute collection from five US institutions and should not be described as anything else. One institution (EZ) contributes a single patient, so we report it but exclude it from any per-site comparison.

Roughly 108 of the 110 patients also appear in mainline BraTS training data[7,8]. Any encoder pretrained on BraTS would therefore have seen the test patients, so we trained from random initialization throughout. This also matches the original setup.

2.2 Segmentation model and training

The model is a four-level U-Net reimplemented from the architecture description: two convolution-plus-ReLU blocks per level with batch normalization, max-pooling in the encoder, transposed convolutions in the decoder, skip connections at every level. No

Table 1. Dataset composition by contributing institution. Institution codes are the two letters following TCGA_ in the filenames.

Institution	Patients	Slices	Tumor slices
DU	45	1,878	624
HT	34	1,029	404
CS	16	358	121
FG	14	640	218
EZ	1	24	6
Total	110	3,929	1,373

pretrained backbone. The loss is a sum of soft Dice and binary cross-entropy. Training ran for 60 epochs per fold with mixed precision, seed 42, on all 3,929 slices.
We kept the empty slices. Dropping them is a common convenience on this dataset and it changes the problem, because a model never shown an empty slice has no reason to learn to predict nothing. Since 65% of the slices are empty, the effect on any slice-averaged metric is large.
Training the five folds took 3.17 GPU-hours in total, roughly 38 seconds per epoch and 38 minutes per fold, on a single NVIDIA Tesla T4 (16 GB) in a Kaggle notebook. The session provided two T4s but the pipeline used only one. Figure 1 shows the training curves. All five folds converge in the same way and none shows the divergence or collapse that would suggest a broken run.

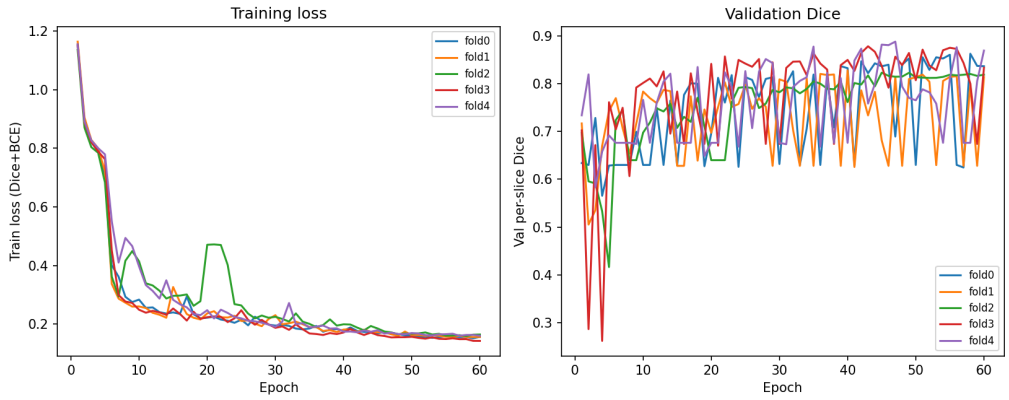

Figure 1. Training loss and validation Dice per epoch for each of the five folds. The validation curve is computed per slice during training and is not the headline metric; it is here only to show that the runs behaved.

2.3 Evaluation protocol

Three choices here do most of the work, and getting any of them wrong produces a number that looks like the paper's without meaning the same thing.

Splits — Folds are assigned at the patient level, never the slice level. A test in the repository builds the splits and asserts that train and validation patient sets are disjoint within every fold, that validation sets are pairwise disjoint across folds, and that together they cover all 110 patients. It runs before training. The realized split is stored as JSON so a reviewer can check it or reuse it.

Aggregation — The headline metric is per-patient Dice computed on the aggregated volume: intersections and set sizes are summed over a patient's slices and the coefficient is formed once, then averaged over patients. Under five-fold cross-validation each patient is predicted exactly once, so all 110 contribute and the result is directly comparable to the reported 0.82 and 0.85. Where prediction and ground truth are both empty we score 1.0; where exactly one is empty we score 0.0.

Extras — IoU and 95th-percentile Hausdorff distance appear in our tables but are not reproduction targets, since the original reports neither. HD95 is undefined when a prediction contains no foreground, which happened for three patients, so it is computed over 107. We report per-slice Dice as well, clearly separated, for reasons that Section 3.1 makes obvious.

2.4 Shape features and radiogenomic analysis

We reconstructed each patient's 3D mask from the cross-validated slice predictions and implemented the three shape features from their published descriptions[1,3]: bounding ellipsoid volume ratio (BEVR), angular standard deviation and margin fluctuation. The precise definitions trace back to the earlier paper and leave some room for interpretation; where a definition was ambiguous we implemented the most literal reading and recorded the choice. This matters, and we return to it in the discussion.

Each feature was discretized at its median and tested against each of the seven genomic cluster labels that ship with the dataset: `RNASeqCluster`, `MethylationCluster`, `miRNACluster`, `CNCluster`, `COCCluster`, `RPPACluster` and `OncosignCluster`. The test is Fisher's exact test, one cluster against the rest, which gives 104 tests per mask set. We report p-values both raw and Bonferroni-corrected across those 104 tests. Patients with a missing label are dropped from that test only, which is why the sample size varies between 89 and 110 across rows.

The whole analysis was run twice, once on the predicted masks and once on the manual ones, which is what makes the comparison the original paper draws possible.

2.5 Deviations from the original

Cross-validation — The original used 22 non-overlapping subsets of five patients. We used five folds of 22. Both are patient-level and both predict every patient exactly once; ours trains on 88 patients per fold rather than 105. Less training data per fold should cost some accuracy, and we cannot rule out that this explains part of the segmentation gap.

Feature definitions — Reimplemented from prose, not from the authors' code. No per-patient reference feature values are published, so we cannot verify our implementation independently of the segmentation.

Correction procedure — The original reports p-values without fully specifying the correction. We apply Bonferroni over all 104 tests within a mask set, which is conservative, and we report raw p-values alongside so the comparison can be made either way.

Training schedule — Optimizer settings, augmentation and epoch count are not fully specified in the paper. Ours are in the repository configuration file.

3 Results

3.1 Segmentation

Table 2 gives the headline comparison. The median is close to the target and the mean is not, which already says something about the shape of the distribution.

Table 2. Per-patient Dice over all 110 patients, against the values reported by Buda et al. Additional metrics are shown for completeness and are not reproduction targets.

Metric	Original	This work	Δ
Mean Dice (per patient)	0.82	0.707	−0.113
Median Dice (per patient)	0.85	0.808	−0.042
Not reported by the original			
Standard deviation of per-patient Dice	–	0.246	
Mean IoU (per patient)	–	0.591	
Median IoU (per patient)	–	0.678	
Mean HD95 (per patient, $n = 107$)	–	9.95	
Median HD95 (per patient, $n = 107$)	–	3.74	

Figure 2 shows why. Most patients are segmented well: 56 of 110 reach 0.8 or better and 37 reach 0.85 or better, so the bulk of the distribution sits where the original says it should. The mean is dragged down by a tail. Eighteen patients fall below 0.5, seven below 0.1, and six score exactly zero. Three of those six produced no foreground anywhere in the volume; the other three produced foreground that missed the tumor entirely. Removing the seven worst cases lifts the mean to 0.754, which is still 0.066 short, so the tail is most of the gap but not all of it.

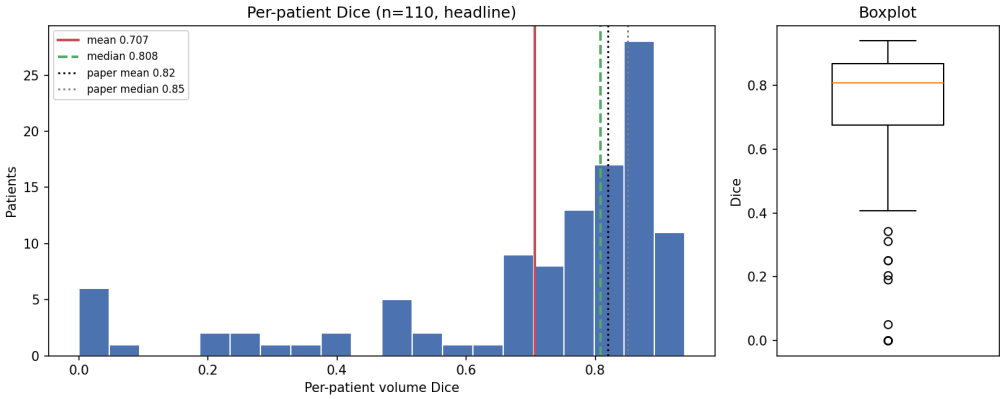

Figure 2. Distribution of per-patient Dice across the 110 patients, with the original paper's mean and median marked. The body of the distribution sits near the target; the left tail does not.

Per-slice Dice deserves a warning. Averaged over all 3,929 slices it comes to 0.858, which is above the paper's headline number and would look like a successful reproduction to anyone not reading carefully. It is inflated by the 2,556 empty slices the model correctly leaves empty. Restricted to tumor-bearing slices it drops to 0.637. Neither number is comparable to a per-patient volume-aggregated Dice, and the gap between them is a good illustration of why the choice of aggregation has to be stated explicitly[9].
Figure 3 makes the range concrete with three patients, at Dice 0.50, 0.82 and 0.94, mirroring the low, moderate and high cases in the original paper's Figure 4. At the high end the predicted and manual contours are nearly interchangeable; at the low end the

model has found roughly the right region but the boundary is wrong in ways a shape feature will feel.

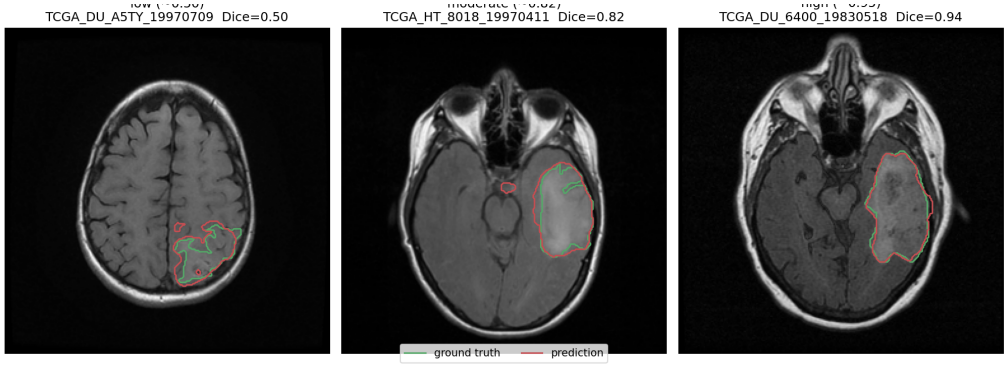

Figure 3. Example predictions at low (∼0.50), moderate (∼0.82) and high (∼0.94) per-patient Dice. Ground-truth boundary and predicted boundary are overlaid on the FLAIR channel for a representative slice from each patient.

3.2 Where the model fails

Grouping per-patient Dice by institution (Figure 4) gives 0.762 for HT ($n = 34$), 0.705 for DU ($n = 45$), 0.650 for FG ($n = 14$) and 0.631 for CS ($n = 16$). EZ contributes one patient at 0.888 and is excluded from the comparison. The spread between sites is smaller than the spread within them, and the two weakest sites are also the two smallest, so we would not read much into the ordering.

We also broke the per-patient scores down by the demographic and grade fields that ship with the dataset (Figure 5). Differences between subgroups are small next to a standard deviation of 0.246, and several subgroups have fewer than ten patients. We include the figure because the fields are there and someone will ask, not because we think it supports a conclusion.

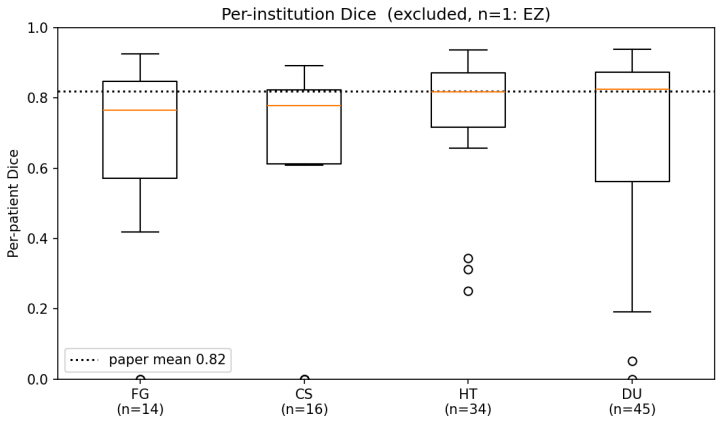

Figure 4. Per-patient Dice grouped by contributing institution. EZ ($n = 1$) is shown for completeness and excluded from comparisons.

3.3 Shape features do not survive the segmentation

Before testing any association we checked something simpler: does a shape feature computed on a predicted mask agree with the same feature computed on the manual mask

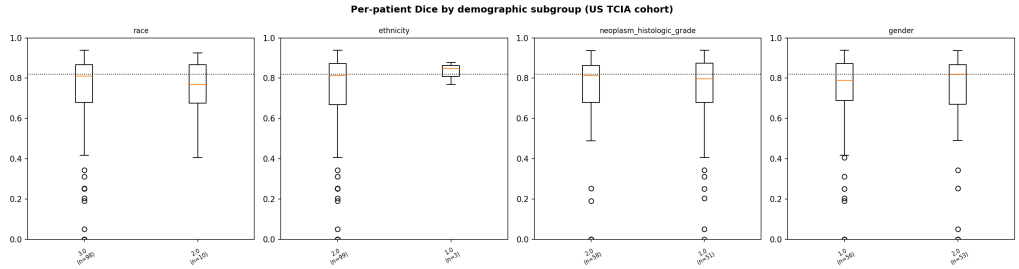

Figure 5. Per-patient Dice by demographic subgroup and histologic grade. Small subgroups are labeled with their n. No subgroup difference approaches the within-group spread.

for the same patient? Table 3 and Figure 6 answer that.

Table 3. Pearson correlation between shape features computed on predicted and on manual masks, for all patients with a computable feature and for the subset segmented at Dice ≥ 0.5.

Feature	All patients		Dice ≥ 0.5	
	n	r	n	r
Angular standard deviation	107	$+0.34$	92	$+0.44$
Margin fluctuation	107	$+0.31$	92	$+0.31$
Bounding ellipsoid volume ratio	107	-0.08	92	-0.01

Angular standard deviation and margin fluctuation carry over weakly and improve, or at least do not degrade, when the badly segmented patients are removed. BEVR does not carry over at all. Its correlation with the manual value is indistinguishable from zero, and restricting to well-segmented patients does not help. That is the single most informative number in this replication, because BEVR is the feature the original paper's headline result runs on.

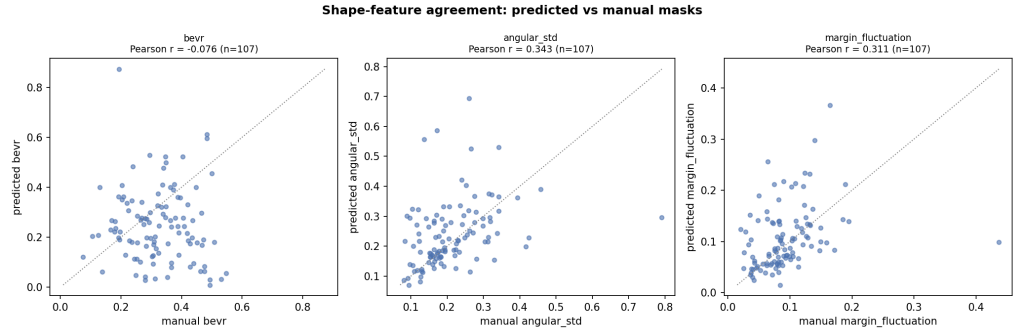

Figure 6. Per-patient agreement between shape features measured on manual masks (x-axis) and on predicted masks (y-axis). Angular standard deviation and margin fluctuation show a weak positive relationship. Bounding ellipsoid volume ratio shows none.

3.4 Radiogenomic associations

Table 4 compares the associations. On the manual masks they are there. Margin fluctuation against RNASeq cluster gives a raw p of 1.1×10^{-4} and survives Bonferroni correction at $p = 0.012$. BEVR against RNASeq cluster gives a raw p of 5.4×10^{-4}, which lands almost exactly on the $p < 0.0005$ that Mazurowski et al.[3] report for that pair, and

misses our conservative correction at $p = 0.056$. Angular standard deviation and margin fluctuation against the cluster-of- clusters label both reach $p = 0.039$ corrected.
On the predicted masks there is nothing. The strongest raw result across all 104 tests is margin fluctuation against methylation cluster at $p = 0.004$, which corrects to 0.42. BEVR against RNASeq cluster, the pair the paper leads with, gives a raw p of 0.157. Figure 7 shows the full grid for both mask sets side by side.

Table 4. Fisher exact tests for the two associations the original paper highlights, plus the strongest result we found on each mask set. Bonferroni correction is over the 104 tests run within each mask set.

Association	Masks	Raw p	Bonferroni p	Verdict
RNASeq × BEVR (original: $p < 0.0002$)	manual	0.00054	0.056	near
RNASeq × BEVR	predicted	0.157	1.000	miss
RNASeq × margin fluct. (original: $p < 0.005$)	manual	0.00011	0.012	pass
RNASeq × margin fluct.	predicted	0.059	1.000	miss
Strongest overall	manual	0.00011	0.012	
Strongest overall	predicted	0.004	0.419	

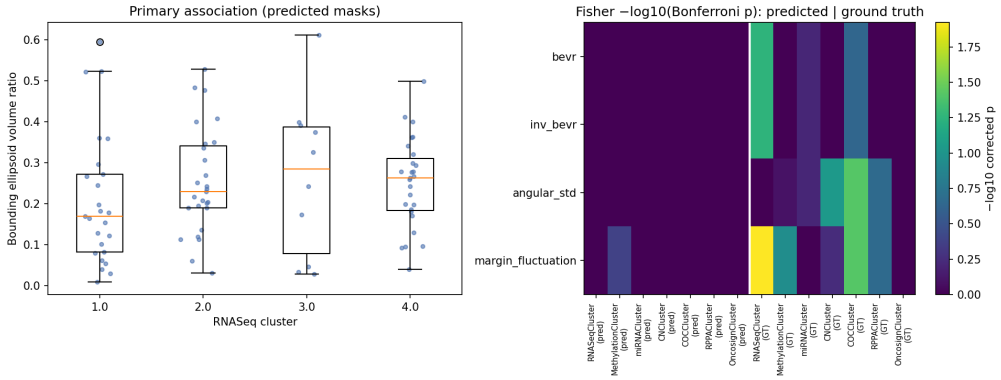

Figure 7. Radiogenomic associations. Left, the primary feature by RNASeq cluster. Right, $-\log_{10} p$ across every feature-by-subtype pair for predicted and manual masks. The manual panel has structure; the predicted panel does not.

3.5 Discrimination

The paper's validation of its own claim is a discrimination task: separate RNASeq cluster R2 from the rest using the inverse bounding ellipsoid volume ratio, reporting an AUC of 0.80 from model masks and 0.78 from manual masks. Our manual-mask AUC is 0.764 over 92 patients with an RNASeq label, close enough to 0.78 that we consider that part reproduced. Our predicted-mask AUC is 0.455 over 89 patients, which is chance. Figure 8 shows both curves.
Restricting to patients segmented at Dice ≥ 0.5 moves the manual AUC to 0.727 and the predicted AUC to 0.522. Throwing away the failures does not rescue the predicted-mask result, which is consistent with the BEVR agreement in Table 3 and rules out the simplest explanation.

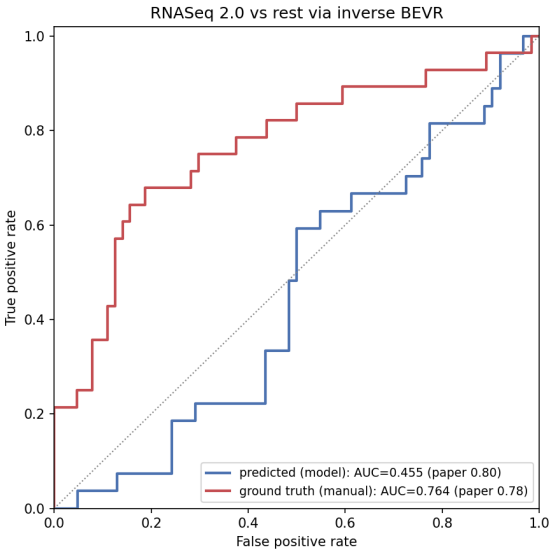

Figure 8. ROC curves for RNASeq cluster R2 against the rest using inverse bounding ellipsoid volume ratio. Manual masks reach 0.764, close to the 0.78 reported. Predicted masks reach 0.455 against a reported 0.80.

4 Discussion

4.1 What reproduced

The biology reproduced. Shape features measured on the manual masks of these 110 patients are associated with genomic subtype, at strengths close to what both original papers report, and the manual-mask discrimination AUC of 0.764 sits within noise of the published 0.78. We reimplemented the features independently, so this is a real check on the underlying finding rather than a restatement of it.

The segmentation half reproduced partially. A median of 0.808 against 0.85 is a reasonable outcome for an independent reimplementation trained on 88 patients per fold instead of 105. The mean is another matter. Six complete failures out of 110 move a mean by more than a tenth of a point while leaving a median almost untouched, and Figure 3 shows the kind of boundary error involved: visible at a glance, invisible in a summary statistic.

4.2 What did not, and the most likely reason

The claim that automatic masks preserve the association did not reproduce, and the gap is not subtle: 0.455 against 0.80.

There are two candidate explanations and we can separate them partly, though not cleanly. The first is that our segmentation is simply worse than theirs, so our features are noisier. The second is that the features are fragile with respect to segmentation error in a way the original numbers do not reveal. The evidence favors the second, or at least says the first is not sufficient. If poor segmentation were the whole story, restricting the analysis to well-segmented patients should recover some signal. It does not. Among patients at Dice ≥ 0.5, BEVR computed on predicted masks correlates with BEVR computed on manual masks at $r = -0.01$.

A plausible mechanism: BEVR is a ratio of tumor volume to the volume of the smallest enclosing ellipsoid, so it depends on the extreme points of the mask rather than its bulk. A few stray voxels at the edge of the volume, or a small false-positive component in

a slice far from the tumor, barely move Dice and can move the bounding ellipsoid a great deal. Our masks are also assembled from independently predicted 2D slices, so slice-to-slice inconsistency inflates the enclosing ellipsoid at almost no cost in overlap. Angular standard deviation and margin fluctuation are boundary statistics computed within slices, which is consistent with their surviving the round trip better.
We want to be careful about how far to push this. We did not reach the original segmentation quality, so we cannot demonstrate that a model at 0.82 mean Dice would also fail. What we can say is that the failure is not explained by our worst patients, and that the feature carrying the paper's headline result is the one that transfers worst.

4.3 What would settle it

Two things, one for replicators and one for the original authors.
For anyone repeating this: reach 0.82 using the 22-fold protocol and rerun the radiogenomic analysis unchanged. If the association returns, the fragility is a function of segmentation quality and there is a threshold worth characterizing. If it does not, the feature implementation is the problem.
For the original authors, or anyone publishing a radiogenomic pipeline of this kind: publishing the per-patient feature values, for both manual and automatic masks, would cost nothing and would let a replicator test the feature implementation separately from the segmentation. As it stands the two failure modes are confounded, and neither we nor a reviewer can tell them apart from the published material. This is a general point about pipeline papers[10] rather than a criticism of this one.

4.4 Limitations

Five folds rather than 22, which costs training data per fold. A single seed and a single architecture, so we report no variance across restarts. Feature definitions inferred from prose. A conservative Bonferroni correction over 104 tests, which is why we give raw p-values as well. And a median split for the Fisher tests, where the original's discretization is not fully specified. Any of these could account for part of the segmentation gap. None of them plausibly account for an AUC of 0.455 next to a reported 0.80.

4.5 A note on the metric

If we had reported per-slice Dice averaged over all slices, this article would have opened with 0.858 against a target of 0.82 and declared the segmentation reproduced. That number is real, computed on the same predictions, and meaningless for this comparison, because two thirds of the slices contain no tumor. Anyone reproducing work on this dataset should state which aggregation they use before quoting a number[9,11].

5 Conclusion

We reproduced the underlying radiogenomic association on manual masks and came within 0.04 of the reported median segmentation Dice. We did not reproduce the claim the paper is built on, that automatically extracted shape features preserve the association with genomic subtype. Our predicted masks yield an AUC of 0.455 where 0.80 is reported, no association survives correction, and the key feature is uncorrelated with its manual counterpart even on patients where the segmentation is good. On the evidence available we report this as a failed replication, with the caveat that our segmentation did not match the original's mean Dice and we cannot fully exclude that as a contributing cause.

6 Model footprint

The paper does not report deployment characteristics, but they are cheap to measure and matter for anyone deciding whether to run this model, so we include them (Table 5). The network is small: 7.76 million parameters, under 30 MB on disk, and a peak inference footprint of 85 MB of GPU memory. A whole patient volume is segmented in 0.32 s on the Tesla T4 and in about 7 s on CPU, so the model is usable without a GPU if throughput is not a concern. Numbers are measured on the fold-0 checkpoint.

Table 5. Deployment characteristics of the trained U-Net, measured on the fold-0 checkpoint. Not reported by the original.

Quantity	Value
Parameters	7,763,041 (7.76 M)
On-disk size	29.7 MB
Peak inference VRAM	85 MB
GPU latency per slice	8.9 ms
GPU latency per volume	319 ms
CPU latency per slice	195 ms
CPU latency per volume	6,978 ms

7 Reproducing this article

All code, the fixed cross-validation splits, the per-patient metrics and every figure's underlying CSV are in the repository listed on the first page. The leakage test runs before training. The full pipeline is five commands and 3.17 GPU-hours; the radiogenomic analysis and every figure in this article are CPU-only and rerun in under a minute from stored predictions.

References

1. M. Buda, A. Saha, and M. A. Mazurowski. "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm." In: **Computers in Biology and Medicine** 109 (2019), pp. 218–225.
2. O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: **Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)**. Vol. 9351. Lecture Notes in Computer Science. Springer, 2015, pp. 234–241.
3. M. A. Mazurowski, K. Clark, N. M. Czarnek, P. Shamsesfandabadi, K. B. Peters, and A. Saha. "Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data." In: **Journal of Neuro-Oncology** 133.1 (2017), pp. 27–35.
4. M. Buda. **LGG MRI Segmentation**. https : / / www . kaggle . com / datasets / mateuszbuda / lgg - mri - segmentation. Kaggle dataset, 110 patients from the TCGA-LGG collection. 2019.
5. K. Clark et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository." In: **Journal of Digital Imaging** 26.6 (2013), pp. 1045–1057.
6. The Cancer Genome Atlas Research Network. "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas." In: **New England Journal of Medicine** 372.26 (2015), pp. 2481–2498.
7. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)." In: **IEEE Transactions on Medical Imaging** 34.10 (2015), pp. 1993–2024.
8. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features." In: **Scientific Data** 4 (2017), p. 170117.

9. L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, et al. “Metrics reloaded: recommendations for image analysis validation.” In: **Nature Methods** 21.2 (2024), pp. 195–212.
10. G. Varoquaux and V. Cheplygina. “Machine learning for medical imaging: methodological failures and recommendations for the future.” In: **npj Digital Medicine** 5 (2022), p. 48.
11. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.” In: **Nature Methods** 18.2 (2021), pp. 203–211.