

Complete Patient-Level Leakage Among Traceable Images in a Widely Used Brain Tumor MRI Benchmark

Jaiveer Bassi

jaiveerbassi@yahoo.com

Independent Researcher

Article

Keywords: data leakage, patient-level splitting, brain tumor MRI, benchmark audit, reproducibility, transfer learning

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11078963/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Complete Patient-Level Leakage Among Traceable Images in a Widely Used Brain Tumor MRI Benchmark

Jaiveer Bassi ^{*}
Independent Researcher, USA
jaiveerbassi@yahoo.com

^{*}Corresponding author: Jaiveer Bassi, Independent Researcher, USA (jaiveerbassi@yahoo.com)

Abstract

Reported accuracies on a widely used four-class brain tumor MRI benchmark routinely exceed 98%, but the released split discards patient identity. We content-matched tumor images to their figshare source and recovered identifiers for 4586 images from 233 patients. Of 906 tumor test images, 761 were traceable; every one belonged to a patient represented in training. Separately, 216 of 1311 test images were pixel-identical to training images. Removing duplicates did not reduce accuracy (99.22% to 99.42%), whereas five-fold patient-disjoint cross-validation yielded 95.19% overall and 93.68% on tumor classes. On the same 761 images, accuracy fell from 99.12% when the patient appeared in training to 93.60% when absent, increasing the error rate 7.3-fold. A matched random-fold control attributed 2.73 of the 4.03 percentage-point gap to patient grouping rather than the change in evaluation design; the class without patient identifiers changed negligibly despite relaxed proxy grouping. These results show that the released split measures recognition of additional slices from previously seen patients, not generalization to unseen patients. We release the recovered patient map, fixed folds, predictions, and analysis code.

Keywords: data leakage; patient-level splitting; brain tumor MRI; benchmark audit; reproducibility; transfer learning

1 Introduction

Deep neural networks are widely evaluated on public medical-image collections because those collections make comparison inexpensive and reproducible. Their apparent precision is meaningful only when the evaluation unit matches the intended use. For two-dimensional slices from a volumetric scan, splitting images independently can place adjacent slices from the same patient in both training and testing. The resulting test set is not independent: it measures recognition of another view of a patient already encountered during training rather than generalization to a new patient. Slice-level leakage has materially inflated performance in brain MRI and other medical-imaging tasks[1, 2], while broader reviews identify non-independent samples and duplicate images as recurring causes of unreliable machine-learning results[3, 4].

The mechanism is specific to how volumetric imaging is distributed for two-dimensional classification. A single diagnostic examination yields many contiguous slices of the same anatomy, acquired on one scanner, in one session, under one protocol. Neighbouring slices therefore share tumor appearance, skull shape, field of view, intensity profile, and whatever artefacts the acquisition introduced. When such slices are shuffled and then divided by image rather than by subject, a model can reach a correct answer by recognising the examination rather than the pathology, and the test set silently rewards that shortcut. Because the shortcut is available to every model trained on the split, leaderboard-style comparisons between architectures remain internally consistent while every entry on the leaderboard overstates what the task requires. This failure mode is not exotic: it is one of the most frequently catalogued forms of leakage in applied machine learning[3], and it is difficult to detect from the outside precisely because the resulting numbers look excellent rather than anomalous.

Aggregated public collections make the problem harder to see. Compilations assembled from several upstream sources are convenient, and their convenience is why they spread; but aggregation frequently discards the metadata that made grouped evaluation possible in the first place. Once subject identifiers are dropped and files are renamed into class folders, a downstream user has no way to honour subject boundaries, and no warning that the boundaries were ever violated. The released split then acquires an authority it has not earned: it is treated as the benchmark’s official partition, reused across papers for comparability, and cited as evidence that a method generalises. Establishing whether that authority is warranted requires recovering the discarded metadata, which is ordinarily impossible. It is possible here only because part of the collection descends from a source that still publishes patient identifiers, which makes this benchmark an unusually tractable test case for a problem that is otherwise invisible.

This study audits a popular four-class brain tumor MRI collection distributed through Mendeley Data and Kaggle[5, 6]. The collection contains 7023 pre-split images labelled glioma, meningioma, pituitary tumor, or no tumor, and has supported many reported accuracies near 99%. Its tumor images are attributed partly to the figshare collection of Cheng *et al.*[7], which retains patient identifiers, but the aggregated four-class release does not. Consequently, the released folders appear to

prevent a patient-level audit.

We recover the discarded identifiers by crop-normalized, orientation-invariant content matching against figshare. This enables three linked tests. First, we quantify exact duplicate and patient overlap in the released split. Second, we compare the same ResNet-18 pipeline under released, deduplicated, and five-fold patient-disjoint evaluation. Third, we run a matched random-fold control over the identical pooled population to separate the effect of patient grouping from the effect of changing the evaluation design. The contribution is therefore not a new classifier. It is a provenance-assisted audit, a corrected evaluation, and reusable metadata that make patient-disjoint testing possible for this benchmark.

2 Results

2.1 The released split contains the same images and patients on both sides

The released folders contain 5712 training and 1311 test images. Hashing decoded pixels identified 216 test images (16.5%) that were pixel-identical to a training image; 103 were also byte-identical. Across both folders, the nominal 7023 files represented 6597 distinct pixel contents. Exact duplication was strongly class-dependent: no glioma test image was duplicated, compared with 106 meningioma, 105 no-tumor, and five pituitary test images.

The patient audit was more consequential. Crop-normalized matching across all eight dihedral image orientations recovered a figshare identifier for 4586 of 5023 tumor images (91.3%) and all 233 source patients. Of the 906 tumor images in the released test folder, 761 (84%) were traceable: 287 glioma, 178 meningioma, and 296 pituitary images. Every traceable test image belonged to a patient represented in the training folder. Overall, 206 of 233 patients (88.4%) appeared in both folders. The claim is limited to the traced population and is weakest in meningioma, for which coverage was 58%; it does not assume anything about the remaining 145 tumor test images. Table 1 summarizes the audit, and Figure 1 shows exact cross-split duplicates.

Table 1: Audit of the released dataset and split. Patient-overlap percentages use only tumor images with a recovered figshare identifier.

Audit quantity	Result
Released folders	5712 train; 1311 test
Distinct pixel contents	6597
Pixel-identical test images in training	216/1311 (16.5%)
Recovered figshare identifiers	4586/5023 (91.3%)
Traceable tumor test images	761/906 (84.0%)
Traceable tumor test images with patient represented in training	761/761 (100%)
Patients appearing in both folders	206/233 (88.4%)

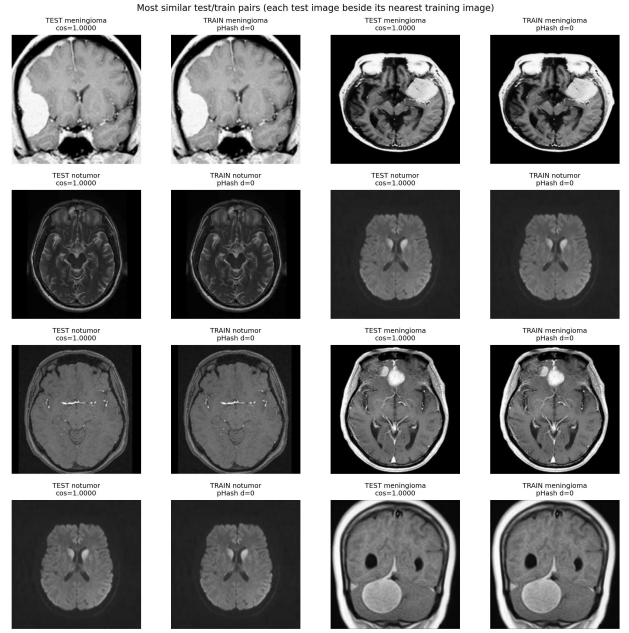


Figure 1: Released test images (left within each pair) and their pixel-identical training counterparts.

2.2 Patient-disjoint evaluation lowers accuracy

We assigned each recovered patient to exactly one of five folds and trained three aligned random seeds per fold. Unmatched tumor images remained training-only. Because no patient identifiers exist for the no-tumor class, we grouped that class by near-duplicate cluster; the tumor-only result is reported separately as the evaluation with complete patient grouping. Training and validation were also group-disjoint. Every eligible image was tested exactly once per seed, and no fold contained patient or exact-pixel overlap across its training and test boundary.

Pooled patient-disjoint accuracy was 95.19% across 6292 images (seed replicates 94.91%, 95.42%, and 95.23%; 95% patient-clustered bootstrap interval 93.58–96.59%). On the 4586 tumor images, accuracy was 93.68% and macro-F1 was 0.935. Across the three complete seed replicates, meningioma was weakest (mean F1 0.898), followed by glioma (0.943) and pituitary tumor (0.959; Figure 2a).

The cleanest comparison uses the 761 traceable tumor test images scored under both regimes. Released-split models, which trained on each patient’s other slices, achieved 99.12%; patient-disjoint models achieved 93.60%. The absolute difference was 5.52 percentage points and the error rate was 7.3 times higher without patient overlap (Figure 3a).

The headline comparison still changes both grouping and evaluation design. We therefore built a five-fold random partition over the same 6292 images, preserving each patient-fold’s class counts, model pipeline, group-disjoint validation, and three seeds. Pixel-identical images were kept together so duplicate leakage could not confound the comparison. Random-fold cross-validation reached 97.92%, 2.73 points above patient-disjoint cross-validation (95% patient-clustered

bootstrap interval 1.75–3.89; Table 2). On the 4586 tumor images, the corresponding accuracies were 97.47% and 93.68%, a 3.79-point difference (95% interval 2.42–5.36). Thus, 68% of the 4.03-point overall gap between released and patient-disjoint evaluation was attributable to the grouping rule and 32% to the remaining design changes.

The per-class pattern supplies a control specific to patient identity. Random grouping improved all three classes with recovered patient identifiers, most strongly meningioma (+7.33 points). No-tumor accuracy changed by only -0.12 points even though its grouping was also relaxed: near-duplicate clusters were indivisible in the patient arm, whereas the random arm kept only exact-pixel groups together (Figure 3b). It is therefore not an unchanged-grouping null; rather, the negligible change despite relaxed proxy grouping strengthens the contrast with the patient-identified classes. Patient-fold means spanned 5.86 points, compared with 0.70 points for random folds (Figures 2b and 3b). A random-effects decomposition estimated patient-partition variation at 2.18 points and seed variation at 0.83 points; for random folds, the partition component reached the zero boundary while seed variation remained 0.87 points. The zero is a truncated method-of-moments estimate, not proof of literal zero variance.

Table 2: Classification performance under released, deduplicated, matched random-fold, and patient-disjoint evaluation. Difference rows report random minus patient accuracy with patient-clustered paired 95% bootstrap intervals.

Evaluation	Test n	Accuracy or difference
Released holdout	1311	99.22% $\pm$ 0.16
Deduplicated holdout	1092	99.42% $\pm$ 0.30
Random-fold cross-validation	6292	97.92%
Patient-disjoint cross-validation	6292	95.19%
Random – patient	6292	2.73 pt (1.75–3.89)
Random-fold, tumor classes	4586	97.47%
Patient-disjoint, tumor classes	4586	93.68%
Random – patient, tumor	4586	3.79 pt (2.42–5.36)

2.3 Removing exact duplicates does not repair the split

We removed every test image that was pixel-identical to training data and collapsed duplicate groups within each folder. Across ten seeds, accuracy changed from 99.22% to 99.42%, an increase of 0.20 points rather than the decrease predicted by a duplicate-memorization account. The 95% confidence interval for the difference was -0.03 to +0.43 points, and a Welch test gave $p = 0.081$ (Table 3). Simply rescored the original seed-42 model on the reduced test set produced the same pattern. More importantly, deduplication left all 756 remaining traceable test images and all 206 shared patients represented in training. It therefore removed repeated files without restoring independence.

Duplicated test images were not easier for the released-split models: their pooled error rate was 1.39%, compared with 0.66% for non-duplicated images. Because ten models scored

the same fixed images, pooled predictions are not independent; paired model-level and image-permutation tests did not reach conventional significance ($p = 0.117$ and $p = 0.078$, respectively). The direction is nonetheless inconsistent with exact-copy memorization as the explanation for near-ceiling accuracy.

Table 3: Ten-seed released and deduplicated evaluations.

	Released	Deduplicated
Test images	1311	1092
Mean accuracy	99.22%	99.42%
Standard deviation	0.16 pt	0.30 pt
Range	98.93–99.47	98.81–99.73

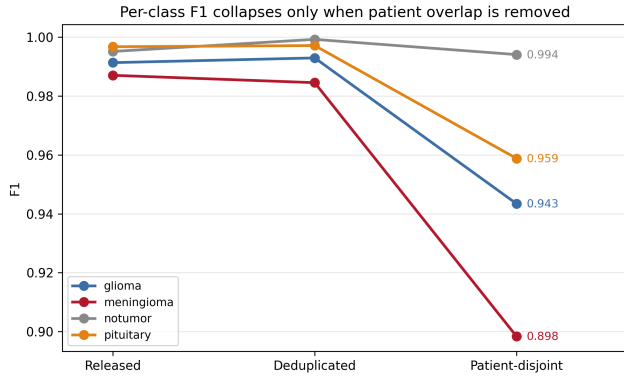
2.4 Calibration and saliency do not establish clinical reliability

Temperature scaling changed little on the released split, where expected calibration error (ECE) was already 0.0043, but the patient-disjoint fold was less well calibrated (ECE 0.0119; Table 4). On that fold, eleven errors exceeded 0.90 confidence before scaling and two after scaling, even though aggregate ECE worsened to 0.0324. This single-fold result shows that a favorable aggregate calibration score on the leaked split does not transfer automatically to unseen patients; it does not establish that temperature scaling solves the problem.

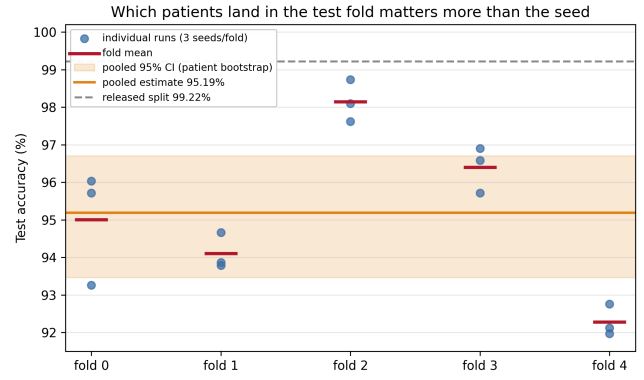
Grad-CAM maps[10] from released and patient-disjoint models were generally diffuse at the final convolutional layer’s 7×7 resolution (Figure 4). A patient-disjoint no-tumor example contained a visible text watermark and attention extending toward image corners. Because the examples were selected illustrations rather than a systematic analysis, they neither prove lesion localization nor rule out source shortcuts.

Table 4: Calibration for seed 42 before and after temperature scaling. The patient-disjoint result uses fold 0 and group-disjoint validation.

	Released		Patient-disjoint	
	Before	After	Before	After
ECE	0.0043	0.0048	0.0119	0.0324
NLL	0.021	0.020	0.109	0.123
Errors > 0.90	4	1	11	2

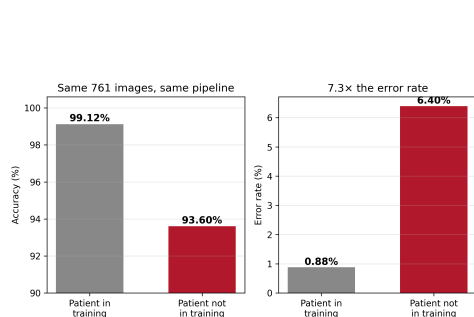


(a) Per-class F1 across released, deduplicated, and patient-disjoint evaluation. Patient-disjoint annotations are arithmetic means of the three complete seed replicates.

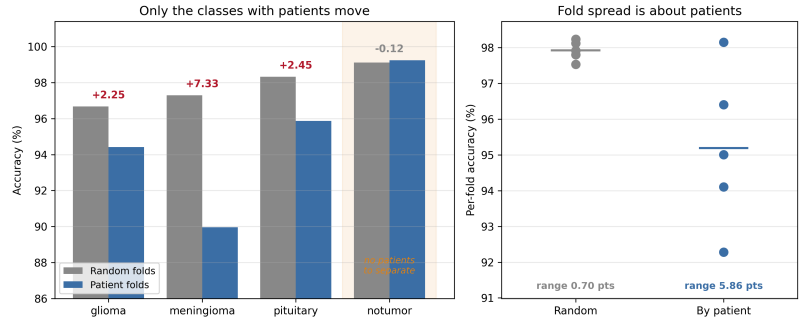


(b) Seed and fold variation under patient-disjoint evaluation.

Figure 2: Performance heterogeneity becomes visible only after patient grouping.



(a) The same 761 images scored with their patients seen or unseen during training.



(b) Matched random-fold and patient-fold results.

Figure 3: Controls isolating patient overlap and patient grouping.

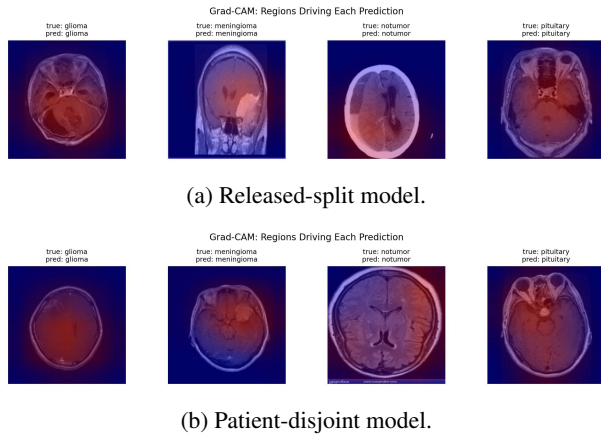


Figure 4: Grad-CAM illustrations for one image per class. These qualitative examples are not localization evidence.

3 Discussion

The released split cannot measure generalization to an unseen patient within its traceable tumor population: all 761 iden-

tifiable tumor test images come from patients represented in training. The practical consequence is visible in two complementary comparisons. On identical images, withholding the patient increases the error rate 7.3-fold. Under matched cross-validation, changing only the fold-assignment rule reduces accuracy by 2.73 points overall and 3.79 points for tumor classes (Table 2). No-tumor lacks patient identifiers and underwent a different, smaller relaxation from near-duplicate to exact-pixel grouping; its negligible -0.12-point change nevertheless strengthens the interpretation that patient identity, rather than fold size, drives the tumor-class difference.

The size of the effect deserves emphasis because accuracy compresses it. A fall from 99.12% to 93.60% on the same 761 images sounds moderate when read as a difference of five percentage points, but the quantity that matters operationally is the error rate, which rises 7.3-fold. A reader comparing published methods on this benchmark is working in a range where differences of a few tenths of a point are reported as progress; the leakage penalty is an order of magnitude larger than that resolution. Two methods separated by 0.3 points on the released split cannot be meaningfully ranked, because the split’s own measurement error with respect to the intended question

is far greater than the gap being interpreted. This is consistent with earlier findings that slice-level leakage inflates apparent performance in brain MRI and in neuroimaging classification more broadly[1, 2], and it extends them by separating the contribution of grouping from the contribution of evaluation design within a single dataset.

Exact duplication is a separate, smaller data-quality defect. Removing it does not reduce accuracy and cannot repair the patient overlap. This distinction matters because duplicate audits are easier than provenance recovery and may create false reassurance: a split can contain no identical files yet remain invalid when different slices from one patient cross the boundary. Benchmark curators should preserve subject identifiers or publish immutable subject-grouped folds, and users should group both test and validation partitions. Model selection against patients present in training is contaminated even when final evaluation is patient-disjoint.

Three practices would have prevented the situation documented here, and each is inexpensive. First, aggregation pipelines should carry an opaque subject key through every transformation, even when the subject’s identity is of no interest to the task; the key costs one column and preserves the option of grouped evaluation for everyone downstream. Second, a redistributed collection should state its provenance per image rather than per collection, so that a user can tell which upstream source a file came from and whether that source retained identifiers. Third, when subject identity genuinely cannot be recovered, the release should say so explicitly instead of shipping a split whose silence implies independence. A collection that documents its own limits invites appropriate use; one that does not is read as endorsing whatever evaluation its folder structure suggests.

For readers of this literature, the practical implication is that a reported accuracy on a pre-split public collection carries an implicit claim about the split that is rarely examined. We suggest three questions that can be asked without access to the original data: whether the evaluation unit is the subject or the image, whether the validation partition used for model selection was grouped in the same way as the test partition, and whether the reported variance reflects repeated partitions or only repeated seeds. Our results bear directly on the third: seed variation alone understates uncertainty substantially, since the estimated partition standard deviation was 2.6 times the seed standard deviation. A study reporting a tight interval across seeds on one grouped split is reporting the stability of its optimiser, not the stability of its estimate.

The patient-disjoint result is not a clinical-performance estimate. It comes from a single aggregated public collection, uses two-dimensional slices, and does not test transport to a new hospital, scanner, demographic group, or acquisition protocol. The no-tumor class has no recovered subject identity and is grouped only by near-duplicate cluster; residual subject overlap is therefore unknown. We assign the 437 unmatched tumor images to training only, which prevents them from contaminating the evaluated tumor population but narrows that population to the recovered figshare component. Reporting 93.68% tumor-only accuracy makes this boundary explicit.

Partition variation also limits comparisons between models. Patient-fold means spanned nearly six points and the estimated partition standard deviation was 2.6 times the seed standard deviation. Differences of one or two points between methods evaluated on different single splits are therefore not interpretable without repeated grouped partitions. This concern applies to our earlier single-partition result and to published studies on this benchmark; Supplementary Table S5 documents that correction and the other analyses superseded during this work. It does not imply that their calculations are incorrect on their stated splits. It means those splits answer a different and easier question.

Several further limitations remain. The provenance audit verified the figshare component but not the reported SARTAJ or Br35H components. The content-matching threshold was calibrated on realistically transformed positive controls, yet 8.7% of tumor images remained unmatched. The classifier ignores three-dimensional context, which may contribute to glioma–meningioma confusion. Calibration and Grad-CAM analyses use one fold and a small number of qualitative examples, respectively. A stronger next study would combine external, institution-disjoint validation; preserved subject and acquisition metadata; volumetric models; systematic shortcut tests with background masking; and uncertainty estimates assessed across all patient folds.

The artifacts we release are intended to make the corrected evaluation the path of least resistance rather than an additional burden. The patient map records, for each matched image, its figshare identifier, the accepted similarity score, and the orientation under which it matched, so that a user can audit or tighten our acceptance threshold rather than trusting it. The fold assignments are cached and immutable, so two studies that adopt them are comparing methods on identical partitions instead of on independent draws whose spread we have shown to be substantial. The per-image predictions allow every headline quantity in this paper to be recomputed without retraining, which is also what allowed us to verify them. None of this requires redistributing the images themselves, and none of it depends on our particular classifier: a future study may replace ResNet-18 entirely and still inherit a patient-disjoint evaluation of this benchmark.

For medical AI, overstated benchmark accuracy can shape research priorities and perceived readiness. On the patient-disjoint tumor evaluation, the model rarely called a tumor healthy (7 of 4586 images) but assigned the wrong tumor type to 6.2%, with glioma recall 0.944. These are slice-level counts, not patient-level clinical endpoints. The model should not be used for diagnosis or treatment decisions, and the released artifacts are intended for benchmark auditing and reproducible research only.

4 Methods

4.1 Dataset and provenance recovery

We used the 7023-image four-class collection from Mendeley Data[5], also mirrored on Kaggle[6]; its released-folder

composition appears in Supplementary Table S1. Images were used under their published folder labels. To recover subject identity, we downloaded the 3064-slice figshare source[7], rendered each slice according to its accompanying conversion description, isolated the non-background brain bounding box, and resized both source and target images to a common square. We compared ImageNet-pretrained ResNet-18 embeddings across rotations and reflections. A transformed-positive control that applied cropping, resizing, and JPEG re-encoding to known matches produced a median cosine similarity of 0.9988 and minimum of 0.9884; we accepted matches at cosine ≥ 0.98 . The complete mapping, scores, orientations, and three label disagreements are released in `figshare_patient_map.json`. Supplementary Methods and Supplementary Table S2 provide additional sensitivity details.

4.2 Duplicate and redundancy audit

We computed SHA hashes over raw bytes and over decoded pixels resized to a fixed resolution. Cross-folder pixel hashes identified re-encoded copies that were not byte-identical. For a secondary redundancy analysis, we greedily grouped images with embedding cosine similarity ≥ 0.99 while preventing two cluster representatives from merging. Exact-duplicate and redundancy counts by class appear in Supplementary Table S3. Exact-duplicate counts support the primary contamination claim; near-duplicate clusters were used only as a conservative grouping proxy for the no-tumor class.

4.3 Model and training

All experiments used ImageNet-pretrained ResNet-18[8] with its final layer replaced by a four-class linear head. Grayscale images were replicated across three channels, resized to 224×224 , and normalized by ImageNet means and standard deviations. Training augmentation applied horizontal flips, rotations up to 10 degrees, and brightness and contrast jitter up to 10%. We optimized cross-entropy with Adam at learning rate 10^{-4} , batch size 32, no weight decay, and a maximum of 30 epochs. A plateau scheduler halved the learning rate after two stagnant validation epochs; early stopping used five epochs without validation-loss improvement. Patient-fold runs used group-disjoint validation by recovered patient or no-tumor cluster. Software versions and all hyperparameters appear in the repository and Supplementary Table S4.

4.4 Evaluation designs

The released condition trained on the 5712-image training folder and tested on all 1311 released test images. The deduplicated condition removed test images duplicated in training and collapsed within-folder exact-duplicate groups. The patient-disjoint condition assigned each of 233 recovered patients to exactly one of five test folds using class-stratified longest-processing-time allocation, balancing folds to within two slices per class. Two patients spanning labels

were assigned by dominant label. Unmatched tumor images were training-only. No-tumor images were deduplicated and grouped by near-duplicate cluster. We ran seeds 0, 1, and 42 in every fold.

For the matched random control, we partitioned the same 6292 eligible images into five test folds with per-class sizes identical to the patient folds but ignored patient identity. Exact-pixel groups remained intact. One cached partition (`rand_folds.json`) was used for all three seeds, and each image appeared in one test fold in both arms. The comparison therefore changes the assignment rule without changing the pooled evaluation population.

4.5 Statistical analysis

All tests were two-tailed and used a significance level of $\alpha = 0.05$; exact p values are reported throughout. Accuracy and macro-F1 were pooled over each complete five-fold replicate. Patient-level bootstrap intervals resampled the 233 recovered patients with replacement; no-tumor clusters were resampled as grouping units when included. The matched random-versus-patient difference was computed per image and bootstrapped by patient or cluster. Variance components used a one-way random-effects method-of-moments decomposition of the 5×3 fold-by-seed accuracy table; negative partition estimates were truncated at zero and described as boundary estimates. Released-versus-deduplicated seed means were compared with Welch’s t -test on the ten per-seed accuracies of each condition, with an interval for the mean difference. Welch’s test was chosen because the two conditions are independent samples with unequal variances (0.16 and 0.30 percentage points) and unequal effective sample sizes; because each group contains only ten seeds, we verified that the per-seed accuracies were consistent with normality using the Shapiro–Wilk test ($p = 0.91$ released, $p = 0.23$ deduplicated) and confirmed the conclusion with a distribution-free two-tailed Mann–Whitney U test ($U = 26$, $p = 0.075$). Duplicate-error analyses treated either model or image as the resampling unit, not each correlated prediction as independent; the image-level comparison used a two-tailed permutation test that shuffled the duplicate label across the 1311 test images over 10,000 resamples, which makes no distributional assumption, and the model-level comparison used a paired t -test over the ten per-seed error-rate pairs ($t = 1.73$; a Wilcoxon signed-rank test agreed, $p = 0.23$). Duplicated images were identified by re-hashing the decoded pixels of the released folders, so the 216 exact duplicates of the audit and the error comparison use one definition; the script `duplicate_error_test.py` regenerates both. Bootstrap intervals are two-sided percentile intervals from 10,000 resamples and are not adjusted for multiplicity; the pre-specified primary comparison is the matched random-versus-patient difference, and the remaining tests are reported as exact p values for description rather than as a family of confirmatory decisions.

Calibration used expected calibration error, negative log-likelihood, and temperature scaling fitted to validation logits following Guo *et al.*[9]. Grad-CAM targeted ResNet-18 layer

458	4[10]. All primary quantities are regenerated from per-image	dataset audit and reported 99.16% on the released split. It	503
459	prediction files by the released analysis scripts.	contains none of the provenance recovery, leakage analysis,	504
		or patient-disjoint evaluation reported here and is superseded	505
		by this manuscript.	506
460	4.6 Ethics declarations		
461	This work is a secondary analysis of publicly available, de-	Author Contributions	507
462	identified imaging data and did not require institutional review	J.B. conceived the study, developed the software, performed	508
463	board approval; no experiments on humans, human samples,	the experiments and analyses, verified the results, prepared the	509
464	or animals were performed. The recovered “patient identi-	figures, and wrote and revised the manuscript.	510
465	fiers” are arbitrary within-dataset labels distributed with the		
466	source collection and do not identify individuals. No person-	Competing Interests	511
467	ally identifiable information appears in any text, figure, table,	The author declares no competing interests.	512
468	or Supplementary Information file.		
469	4.7 Use of AI assistance		
470	An AI coding assistant (Anthropic Claude) was used to re-	Funding	513
471	verse analysis code, execute experiments, draft and edit text,	This research received no specific grant from any funding	514
472	and review intermediate results. OpenAI Codex was subse-	agency in the public, commercial, or not-for-profit sectors.	515
473	quently used for structural and language editing to prepare the		
474	manuscript for submission. Large language models are not au-	References	516
475	thors of this work. The author executed or inspected the anal-	[1] Yagis, E. <i>et al.</i> Effect of data leakage in brain MRI	517
476	yses, checked the cited literature, verified the final manuscript	classification using 2D convolutional neural networks.	518
477	and artifacts, and takes full responsibility for their accuracy	<i>Sci. Rep.</i> 11 , 22544 (2021). https://doi.org/10.	519
478	and integrity.	1038/s41598-021-01681-w	520
479	Data Availability	[2] Wen, J. <i>et al.</i> Convolutional neural networks for clas-	521
480	The source images are third-party and are not redis-	sification of Alzheimer’s disease: overview and re-	522
481	tributed. The aggregated four-class collection is avail-	producible evaluation. <i>Med. Image Anal.</i> 63 , 101694	523
482	able from Mendeley Data[5] and Kaggle[6]; the figshare	(2020). https://doi.org/10.1016/j.media.	524
483	source used for provenance recovery is available under	2020.101694	525
484	CC BY 4.0[7]. Derived patient identifiers, fold assign-	[3] Kapoor, S. & Narayanan, A. Leakage and the re-	526
485	ments, file manifests, per-run metrics, and per-image pre-	producibility crisis in machine-learning-based science.	527
486	dictions are publicly available in the versioned GitHub	<i>Patterns</i> 4 , 100804 (2023). https://doi.org/10.	528
487	release v1.2.2 at https://github.com/imjbassi/	1016/j.patter.2023.100804	529
488	Brain-Tumor-MRI-Classification/releases/	[4] Roberts, M. <i>et al.</i> Common pitfalls and recommenda-	530
489	tag/v1.2.2 and the corresponding Zenodo archive at	tions for using machine learning to detect and prognos-	531
490	https://doi.org/10.5281/zenodo.22755070 .	ticate for COVID-19 using chest radiographs and CT	532
		scans. <i>Nat. Mach. Intell.</i> 3 , 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0	533
491	Code Availability		534
492	All code required to rebuild the split, train the models,	[5] Ghaffar, A. Brain Tumor Data. <i>Mendeley Data</i> ,	535
493	reproduce the statistical analyses, and regenerate the fig-	V1 (2024). https://doi.org/10.17632/	536
494	ures is available under the MIT License in GitHub re-	w4sw3s9f59.1	537
495	lease v1.2.2 at https://github.com/imjbassi/	[6] Nickparvar, M. Brain Tumor MRI Dataset.	538
496	Brain-Tumor-MRI-Classification/releases/	<i>Kaggle</i> (2021). https://www.kaggle.	539
497	tag/v1.2.2. The release includes pinned Python depen-	com/datasets/masoudnickparvar/	540
498	dencies and immutable patient and random-fold assignments.	brain-tumor-mri-dataset	541
499	Relation to a Previous Preprint	[7] Cheng, J. <i>et al.</i> Enhanced performance of brain tumor	542
500	An earlier preprint by the same author, <i>Brain Tu-</i>	classification via tumor region augmentation and parti-	543
501	<i>mor Classification with Pretrained CNNs in PyTorch</i>	tion. <i>PLoS ONE</i> 10 , e0140381 (2015). https://doi.	544
502	(DOI: 10.13140/RG.2.2.21638.28484), predates this	org/10.1371/journal.pone.0140381	545

- 546 [8] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learn-
547 ing for image recognition. In *Proc. IEEE Conf. Com-
548 puter Vision and Pattern Recognition* 770–778 (2016).
549 <https://doi.org/10.1109/CVPR.2016.90>
- 550 [9] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On
551 calibration of modern neural networks. In *Proc. 34th In-
552 ternational Conference on Machine Learning* 1321–1330
553 (2017).
- 554 [10] Selvaraju, R. R. *et al.* Grad-CAM: visual explana-
555 tions from deep networks via gradient-based localiza-
556 tion. In *Proc. IEEE International Conference on Com-
557 puter Vision* 618–626 (2017). [https://doi.org/
558 10.1109/ICCV.2017.74](https://doi.org/10.1109/ICCV.2017.74)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryinformation.pdf](#)