



Leakage-Free Evaluation of Capsule Network Heads for Brain Tumor MRI Classification

Maulana Kanigara*

Universitas Dian Nuswantoro,
Indonesia

Guruh Fajar Shidik²

Universitas Dian Nuswantoro,
Indonesia

***Corresponding author:**

Maulana Kanigara, Universitas Dian Nuswantoro,
Indonesia. ✉ kanigara11@gmail.com

Article Info:

Article history:

Received: September 19, 2026

Revised: October 19, 2026

Accepted: October 21, 2026

Available Online: October 29, 2026

Keywords:

Brain Tumor Classification;
Capsule Network;
Data Leakage;
Magnetic Resonance Imaging;
Reproducible Evaluation;



This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by
Politeknik Siber Cerdika International

Abstract

Background: Deep learning studies using the widely used Brain Tumor MRI Dataset routinely report accuracies above 95%. However, the dataset combines overlapping public collections, and capsule networks are often promoted for such data without controlled comparisons.

Objective: This study quantified duplicate-induced data leakage in the dataset and investigated whether capsule heads improve classification under leakage-free evaluation.

Methods: Exact and near-duplicate slices were identified using file hashing and correlation analysis of head-cropped thumbnails, with the correlation threshold calibrated against label-conflict rates. Ten models using three backbones were trained under a unified protocol with five random seeds on a consumer-grade GPU. Capsule, dense, and pooling heads were compared. The models were also evaluated using the original, potentially leaky split, and their rotation robustness and inference costs were assessed.

Results: Near-duplicates linked 1,707 of 5,712 training images (29.9%) to images in the test set, and data leakage inflated test accuracy by 3.22–6.33 percentage points through both memorization and the larger effective training set. With a lightweight backbone trained from scratch, the capsule head improved macro-F1 by 7.16 percentage points over the dense head while using 91.3% fewer parameters, and by 6.63 percentage points over the pooling head. These improvements were consistent across all five random seeds. However, the capsule head provided no improvement with ImageNet-pretrained backbones and did not consistently reduce the loss in accuracy under rotation.

Conclusion: Capsule heads appear to offer the greatest benefit when features are learned from limited data. Leakage-free evaluation across multiple random seeds is essential for credible brain tumor MRI benchmarks. These conclusions apply to the tested public dataset and experimental protocol and do not constitute clinical validation.

To cite this article: Kanigara, M., & Shidik, G. F. (2026). Leakage-free evaluation of capsule network heads for brain tumor MRI classification. *Equivalent: Jurnal Ilmiah Sosial Teknik*, 8(5), 2042-2056. <https://doi.org/10.59261/jequi.v8i5.478>

INTRODUCTION

Brain and other central nervous system tumors are among the most consequential neoplasms because they directly affect neurological function, and their management depends on the early and accurate differentiation of tumor types. The 2021 World Health Organization (WHO) classification of central nervous system tumors integrates histological and molecular criteria Louis et al. (2021), yet magnetic resonance imaging (MRI) remains a first-line modality for detecting lesions and establishing differential diagnoses among common entities such as gliomas, meningiomas, and pituitary tumors. Population-based registries show that these tumors remain a persistent health burden that varies by histological type and age (Ostrom et al., 2021). Interpreting MRI studies, however, depends on specialist radiologists, who are unevenly distributed. Indonesia, for example, has approximately 2,161 registered radiologists, or 1.2 per 100,000 inhabitants, concentrated in Java and Sumatra (Yunus, 2024). In such settings, automated

2042 | Equivalent: Jurnal Ilmiah Sosial Teknik

decision-support systems that run on affordable hardware are attractive, but only if their reported accuracy reflects performance on genuinely new patients. For an AI system to support clinical decision-making, benchmark performance must reflect generalization to previously unseen patients rather than similarity to images already represented in the training data. This requirement directly links the clinical need for reliable classification to the methodological need for leakage-resistant evaluation. In Indonesia, limited radiologist availability increases the appeal of computational decision support while also heightening the importance of conservative validation. Accordingly, the present study contributes to evaluation methodology for resource-constrained settings rather than claiming that the evaluated systems are ready for clinical deployment.

Deep learning has become a dominant approach to medical image analysis Litjens et al. (2017), and brain tumor classification using two-dimensional MRI slices is one of its most frequently studied tasks. Early work using the Figshare collection of contrast-enhanced T1-weighted MRI images Cheng et al. (2015) demonstrated that transfer learning with ImageNet-pretrained convolutional neural networks (CNNs) could distinguish among gliomas, meningiomas, and pituitary tumors with high accuracy (Deepak & Ameer, 2019; Swati et al., 2019). The field has since increasingly adopted the Brain Tumor MRI Dataset distributed on Kaggle Nickparvar (2021), which combines the Figshare collection with the SARTAJ Bhuvaji (2020) and Br35H datasets and includes an additional no-tumor class. Its accessibility has made it a de facto benchmark, and recent studies routinely report accuracies exceeding 95%. For instance, Lu et al. (2026) compared 13 baseline models using a cleaned version of the dataset. Consequently, reported accuracy gains above 95% do not necessarily demonstrate improved generalization when the independence of training and test sets has not been verified; some of the apparent improvement may instead result from overlap among source collections or duplicated image slices.

Capsule networks have been repeatedly proposed for this task. A capsule represents an entity using a vector whose length encodes the entity's activation, commonly interpreted as the probability that the entity is present, while its orientation encodes pose information. Dynamic routing by agreement is used to establish part-whole relationships (Sabour et al., 2017). Afshar et al. (2018) argued that these properties could make capsule networks robust to rotation and affine transformations and less dependent on large training datasets, potentially making them suitable for medical imaging. They subsequently incorporated coarse tumor boundaries into capsule-based models (Afshar et al., 2019). Hybrid architectures that combine convolutional backbones with capsule layers have continued to appear in the literature (Jabbar et al., 2023). However, the robustness argument has been challenged. In controlled ablation experiments, Gu et al. (2021) found that capsule networks were not more robust to affine transformations than comparable CNNs and that dynamic routing could even reduce robustness.

Many brain tumor classification studies also modify the backbone, classification head, loss function, and training procedure simultaneously while reporting results from a single training run. Consequently, the contribution of the capsule component remains uncertain, particularly because performance variability caused by random initialization alone can exceed the differences being claimed (Bouthillier, 2021). Previous capsule-based studies also differ in dataset composition, preprocessing, backbone architecture, loss function, and training procedure. Because these factors often change simultaneously, their reported performance gains cannot be attributed uniquely to the capsule head. The present study therefore holds the backbone and training conditions constant within each head-matched comparison, enabling a more direct examination of the incremental contribution of the classification head.

A second, more fundamental problem concerns data leakage. Leakage between training and evaluation datasets is a leading cause of irreproducible results in machine-learning research (Kapoor & Narayanan, 2023). In brain MRI classification, splitting images at the slice level rather than the subject level has been associated with reported accuracy inflation of 29%–55% (Yagis et al., 2021). Such errors form part of a broader pattern of methodological shortcomings in medical imaging research Varoquaux & Cheplygina (2022) and are related to shortcut learning, in which models exploit spurious regularities rather than the intended signal (Geirhos et al., 2020). Near-duplicate images are a recognized source of this problem, even in general computer vision benchmarks (Barz & Denzler, 2020). Because the Kaggle dataset aggregates collections obtained

from the web, duplicated and reprocessed image slices are plausible. Two recent studies have addressed related concerns: the BRISC dataset excluded exact and near-duplicate images while introducing a new expert-annotated resource Fateh et al. (2026) and Lu et al. (2026) constructed a leakage-free benchmark comprising 3,522 unique scans.

Although these efforts yielded cleaner datasets, they did not quantify the extent to which leakage inflates the evaluation metrics of the same models trained under the original protocol. Nor did they distinguish the effects of memorizing duplicate or near-duplicate images from those of having a larger effective training set, or establish an objective criterion for selecting the near-duplicate similarity threshold. Similarly, capsule-based performance claims have not been systematically reassessed using a leakage-free, multi-seed experimental design that isolates the classification head from the backbone and measures computational costs on consumer-grade hardware. Such measurements are relevant both to resource-constrained institutions and to the environmental costs of model development (Schwartz et al., 2020). Unlike earlier work that primarily cleaned or reconstructed the benchmark, the present study quantifies the methodological consequences of leakage by training the same models under leakage-prone and leakage-free protocols and distinguishing the contribution of duplicate-related exposure from the effect of a larger training set.

This study addresses these gaps through two related research questions: how much duplicate-induced leakage exists in the Brain Tumor MRI Dataset, and to what extent does it inflate classification accuracy? Additionally, does a capsule head, compared with dense and global average pooling heads attached to identical backbones, improve accuracy, rotation robustness, or computational efficiency when evaluation is leakage-free? The study's novelty lies in combining an objectively calibrated duplicate audit, a controlled comparison between leakage-prone and leakage-free evaluation protocols that decomposes the resulting performance inflation, and a head-level capsule ablation across three backbones using five random seeds, statistical testing, and computational measurements on a single consumer-grade GPU.

The findings are intended to inform researchers benchmarking models on this dataset and practitioners considering lightweight models for settings with limited radiology capacity. The objectives are operationalized as follows: (1) quantify duplicate-induced leakage and its association with changes in test accuracy under matched leakage-prone and leakage-free protocols; and (2) estimate the incremental effects of capsule, dense, and global average pooling heads on accuracy, macro-F1 score, rotation response, and computational efficiency while holding the backbone constant. The practical contribution is therefore methodological guidance for reproducible model evaluation and resource-aware model selection. Because no external or prospective clinical validation was performed, the findings do not establish the evaluated models as diagnostic systems suitable for patient care.

METHOD

This quantitative experimental study manipulated the data split, classification head, and backbone while holding all other training conditions constant. The study used the first publicly released version of the Brain Tumor MRI Dataset Nickparvar (2021), comprising 7,023 two-dimensional MRI slices across four classes, with 5,712 images in the official Training folder and 1,311 images in the official Testing folder. The data are publicly available and de-identified. Because the study did not involve human or animal participants, ethical approval was not required. Figure 1 summarizes the workflow. The unit of analysis varied by stage: individual MRI slices served as prediction units; duplicate groups served as the units for preventing cross-subset data leakage; seed-level model metrics constituted the paired observations for classification-head comparisons; and paired per-image predictions were used for McNemar's tests. The official Testing folder was retained as the common evaluation benchmark within this public dataset, allowing clean and leaky protocols to be compared using identical test images.

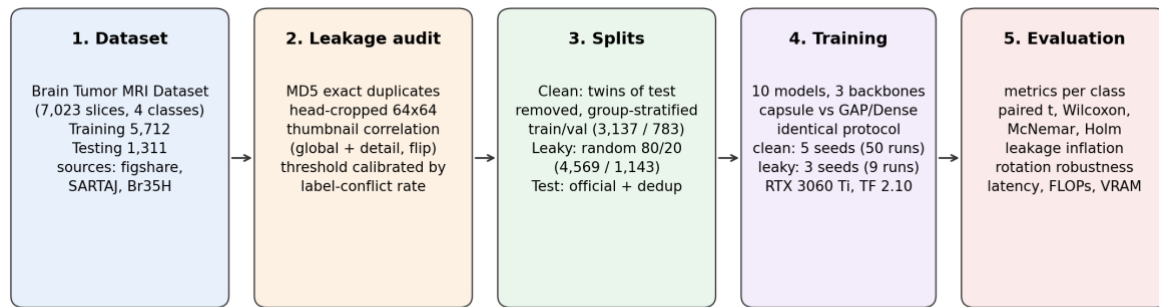


Figure 1. Research workflow, from the leakage audit to training and evaluation

Duplicate auditing was performed before training. Exact duplicates were identified using identical MD5 file hashes. To identify near-duplicates, each image was cropped to the head region, resized to 64×64 pixels, and normalized. Two global Pearson correlation coefficients were calculated for each candidate pair: one using the complete thumbnails and another using the thumbnails after subtraction of their respective Gaussian-blurred versions to emphasize edges and local structures. Each comparison was performed in both the original and horizontally mirrored orientations. Candidate pairs with a detail correlation of at least τ and a global correlation of at least 0.95 were classified as near-duplicates and merged into groups using the union-find algorithm.

Genuine duplicates were expected to have highly consistent labels, whereas unrelated image pairs would be expected to disagree in their labels with probability $1 - \Sigma p_c^2 = 0.748$. Accordingly, τ was selected by examining label-conflict rates across similarity thresholds ranging from 0.80 to 1.00, the corresponding removal rates, and visual inspections of candidate pairs. Here, global correlation refers to the Pearson correlation coefficient calculated across all pixels of the normalized 64×64 thumbnails, whereas detail correlation refers to the Pearson correlation coefficient calculated after subtracting a Gaussian-blurred version of each thumbnail. Threshold calibration combined label-conflict rates, removal rates, and visual review to reduce the risk of incorrectly classifying anatomically similar but distinct images as duplicates solely because of high global correlation. The union-find algorithm converted pairwise similarity links into connected duplicate groups, which could contain more than two images.

Under the clean protocol, training images belonging to any duplicate group that contained a test image or exhibited conflicting labels were removed. Exact duplicates within the Training folder were collapsed, and a stratified, group-based five-fold partitioning procedure was used to construct the validation set while ensuring that no duplicate group crossed the training-validation boundary. Performance was evaluated on both the official test set and a deduplicated test set containing one image per duplicate group. Under the leaky protocol, a conventional random, stratified 80:20 split was applied to the full Training folder. After cleaning, 3,920 images remained available for model development. These images were partitioned by duplicate group into 3,137 training images and 783 validation images, yielding an approximately 80:20 split. The five grouped folds were used to establish a stratified validation partition without allowing duplicate groups to cross between the training and validation subsets. The official 1,311-image Testing folder was never used for model fitting. The deduplicated test set ($n = 1,039$) was used exclusively for additional sensitivity analysis.

Ten model configurations based on three backbone architectures were compared, with their model sizes reported in Table 2. A lightweight VGG-style backbone comprising four convolutional blocks, with 32–256 filters and a 14×14 output feature map, was trained from scratch. VGG16 Simonyan & Zisserman (2015), truncated after its fourth convolutional block (VGG16-B4), and EfficientNet-B0 Tan & Le (2019) were initialized with ImageNet-pretrained weights (Deng et al., 2009). Each backbone was paired with a capsule head and a non-capsule classification head, using global average pooling (GAP) or, for the lightweight backbone, an additional flattened representation followed by a 512-unit dense layer. Full VGG16, ResNet50 He et al. (2016), and the convolutional neural network (CNN) from the authors' earlier draft were included as reference models. The capsule head comprised 32 types of 8-dimensional primary

capsules, which were routed through three iterations of dynamic routing to four 16-dimensional class capsules. The lengths of the class-capsule output vectors served as class scores (Sabour et al., 2017). Capsule models were trained using the margin loss function:

$$L_k = T_k \max(0, 0.9 - \|\mathbf{v}_k\|)^2 + 0.5 (1 - T_k) \max(0, \|\mathbf{v}_k\| - 0.1)^2,$$

where $T_k = 1$ for the true class and $T_k = 0$ otherwise. Non-capsule classification heads used softmax outputs and categorical cross-entropy.

All models followed the same training protocol: grayscale inputs resized to 224×224 pixels; data augmentation involving rotations within $\pm 15^\circ$, translations and zooming within $\pm 10\%$, and horizontal flipping; the Adam optimizer Kingma & Ba (2015), with a learning rate of 10^{-4} ; a batch size of 16; balanced class weights; and a maximum of 60 training epochs. Early stopping was applied after 10 consecutive epochs without improvement in validation loss, and the learning rate was halved after four consecutive epochs without improvement. Pretrained backbones were fully fine-tuned. Under the clean protocol, each model was trained using five random seeds, yielding 50 runs. Under the leaky protocol, Caps-Light, EffNet-B0, and the first-draft CNN were each trained using three random seeds, yielding nine runs. Two additional seeds and two model configurations (VGG16-B4 + GAP and Caps-EffNet-B0) were introduced following an interim review, and all runs are reported.

Training was performed on an NVIDIA GeForce RTX 3060 Ti GPU with 8 GB of memory, using TensorFlow 2.10.1 and deterministic operations. Repeated runs with the same seed produced bit-identical predictions. The additional seeds and the VGG16-B4 + GAP and Caps-EffNet-B0 comparisons were introduced after interim inspection to strengthen the head-matched analysis. Because these additions were not fully prespecified, they are reported transparently and interpreted as supplementary, exploratory evidence rather than as independent, preregistered confirmation. All clean-protocol results reported in Tables 2 and 3 are based on five seeds per model, whereas the leaky-protocol comparison in Table 4 is based on three seeds per model.

Model performance was evaluated using accuracy, macro-averaged F1-score, precision, recall, one-versus-rest area under the receiver operating characteristic curve (ROC AUC), the Matthews correlation coefficient (MCC), and absolute confusion matrices. Capsule models were compared with their head-matched counterparts using paired t-tests with 95% confidence intervals and Wilcoxon signed-rank tests across seeds, with Holm correction for multiple comparisons (Holm, 1979). McNemar's tests were also performed on paired per-image predictions within each seed (Dietterich, 1998). Robustness to image rotation (0° – 90°) and computational efficiency were assessed using floating-point operations (FLOPs), inference latency, throughput, and GPU memory consumption. The code is available from the corresponding author upon reasonable request. The paired t-test used the five seed-level macro-F1 differences as paired observations to estimate the mean difference and its confidence interval. The Wilcoxon signed-rank test was retained as a distribution-robust sensitivity analysis; however, with only five paired observations ($n = 5$), its attainable two-sided p-values are coarse, and statistical power is limited. In contrast, McNemar's test used paired per-image predictions within each seed to compare classification errors on the same test images. Holm adjustment was applied across the five prespecified head comparisons reported in Table 3, whereas the rotation analyses remained exploratory and were not adjusted for multiple comparisons.

RESULTS AND DISCUSSION

Results

Sample Characteristics and Leakage Audit

The dataset was moderately imbalanced, and the audit showed that the training and test folders were far from independent (Table 1). There were 503 exact-duplicate pairs, of which 215 crossed the two folders; none had conflicting labels. At the calibrated threshold, a further 8,669 near-duplicate pairs were identified, 2,997 of which occurred between training and test images. These pairs formed 1,915 duplicate groups, the largest containing 41 slices, and 733 groups spanned both folders. Figure 2 illustrates the forms these duplicates took, ranging from mirrored and re-contrasted copies to adjacent slices showing the same lesion. Pair counts are not equivalent to image counts because a connected duplicate group can contain several images and therefore generate multiple pairwise links. Applying group-level removal to the training folder excluded 1,707 images belonging to groups that overlapped with the test folder, 4 images in label-conflict

groups, and 81 exact duplicates within the training folder, yielding a total of 1,792 removed training images.

Table 1
Dataset composition, leakage audit, and resulting splits

Subset	Glioma	Meningioma	No tumor	Pituitary	Total
Official Training folder	1,321	1,339	1,595	1,457	5,712
Official Testing folder	300	306	405	300	1,311
Removed from training (clean)	131	290	1,082	289	1,792
Removed (% of class)	9.9	21.7	67.8	19.8	31.4
Clean training subset	949	835	412	941	3,137
Clean validation subset	241	214	101	227	783
Deduplicated test set	278	284	203	274	1,039

Note. Removals comprise 1,707 images sharing a duplicate group with the test set, 4 images in label-conflict groups, and 81 exact duplicates within the Training folder. The leaky protocol used 4,569 training and 1,143 validation images from a random stratified split.

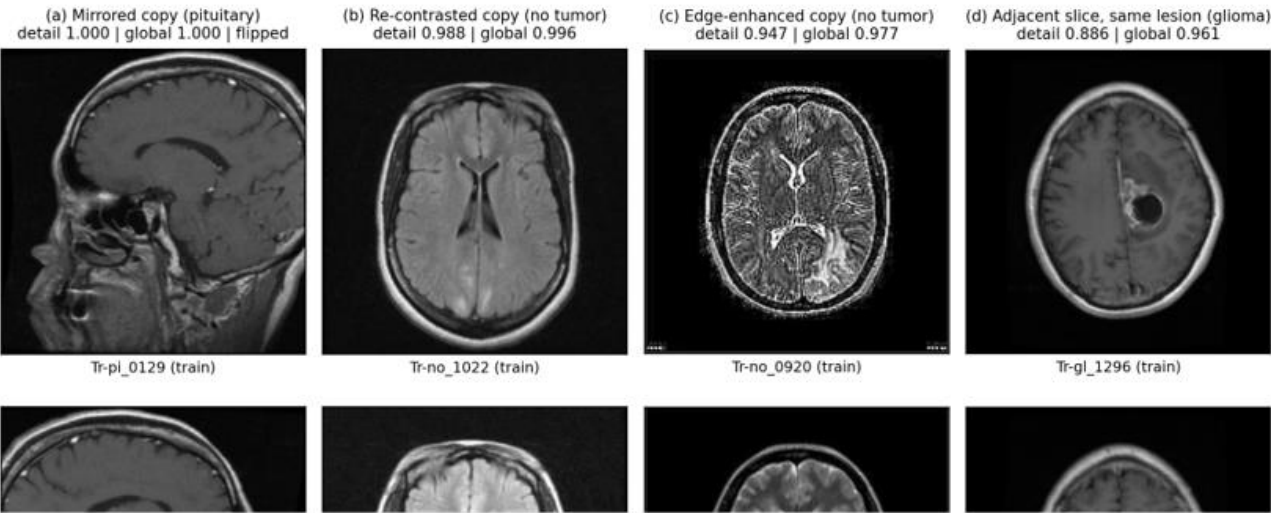


Figure 2. Examples of training–test near-duplicates detected by the audit: (a) mirrored copy, (b) re-contrasted copy, (c) edge-enhanced copy, and (d) adjacent slice showing the same lesion. Values are thumbnail correlations

Figure 3 summarizes the threshold calibration. The label-conflict rate was at most 0.001 in the two highest similarity bands and zero in every band from 0.80 to 0.97, far below the rate of 0.748 expected for unrelated pairs. This finding indicates that the detected pairs were genuine duplicates rather than chance matches. The few conflicts in the highest similarity bands involved identical images labeled as either meningioma or pituitary tumor, indicating label noise in the source data. Lowering the threshold from 0.95 to 0.85 increased the proportion of training images removed from 24.2% to 30.0%, after which the curve flattened, reaching 30.8% at a threshold of 0.80. Inspection of the 0.87–0.90 band revealed re-contrasted copies in the no-tumor class and adjacent slices from the same patient in the glioma and meningioma classes. The threshold was therefore fixed at 0.85. Leakage was strongly class-dependent: 67.8% of no-tumor training images were removed, compared with 9.9% of glioma images, and 202 of the 405 no-tumor test images (49.9%) were redundant copies within the test set. Under the leaky protocol, 1,374 of 4,569 training images (30.1%) shared a duplicate group with the test set, and 620 of 1,143 validation images (54.2%) shared a duplicate group with the training subset.



Figure 3. Calibration of the near-duplicate threshold: label-conflict rate per detail-similarity band compared with the rate expected for unrelated pairs (left) and percentage of training images removed at each threshold, overall and per class (right)

Classification Performance under the Clean Protocol

Table 2 reports performance under the clean protocol. EffNet-B0 achieved the highest macro-F1 score on the official test set ($95.68 \pm 0.58\%$), followed by ResNet50 ($95.26 \pm 1.01\%$) and Caps-EffNet-B0 ($94.79 \pm 1.42\%$). Both VGG16-B4 models achieved macro-F1 scores of approximately 92.8%. Among the models trained from scratch, Caps-Light ($84.79 \pm 1.58\%$) clearly outperformed Light + GAP ($78.16 \pm 2.85\%$) and Light + Dense ($77.63 \pm 3.49\%$). Two models exhibited instability across seeds: full VGG16 collapsed to a single predicted class in two of five runs ($57.03 \pm 43.74\%$), whereas VGG16-B4 + GAP never collapsed. The first-draft CNN achieved a macro-F1 score of only $56.64 \pm 24.53\%$. Results on the deduplicated test set differed from those on the official test set by -0.8 to $+0.4$ percentage points in macro-F1, and the ranking changed only among models whose mean scores differed by less than 0.5 percentage points. Figure 4 relates macro-F1 to model size and shows that the capsule models did not define the accuracy-size frontier, which was established by EffNet-B0. The stronger performance of the pretrained EfficientNet-B0 and ResNet50 models was consistent across the principal accuracy and macro-F1 summaries and was compatible with the benefits of transfer learning. However, this finding should not be interpreted as proof that pretraining is universally superior because backbone capacity, optimization stability, and initialization differed simultaneously across these model families.

Table 2

Classification performance under the clean protocol (mean $\pm$ SD over five seeds)

Model	Params (M)	Accuracy (%)	Macro-F1 (%)	Macro-AUC (%)	MCC	Macro-F1, dedup. (%)
Caps-Light	2.35	85.71 ± 1.42	84.79 ± 1.58	96.60 ± 0.38	0.811 ± 0.019	84.13 ± 1.43
Light + Dense	26.87	79.13 ± 2.82	77.63 ± 3.49	94.63 ± 1.15	0.725 ± 0.035	76.93 ± 3.49
Light + GAP	1.18	78.66 ± 3.16	78.16 ± 2.85	94.34 ± 0.88	0.719 ± 0.038	78.43 ± 2.07
First-draft CNN	59.16	61.94 ± 21.14	56.64 ± 24.53	91.10 ± 7.40	0.526 ± 0.262	55.81 ± 23.14
Caps-VGG16-B4	9.40	92.97 ± 0.31	92.86 ± 0.35	98.85 ± 0.20	0.907 ± 0.004	92.24 ± 0.40
VGG16-B4 + GAP	7.64	92.89 ± 0.84	92.78 ± 0.83	98.67 ± 0.33	0.905 ± 0.011	92.25 ± 0.72
VGG16	14.72	62.75 ± 36.43	57.03 ± 43.74	78.55 ± 26.10	0.513 ± 0.472	57.43 ± 43.01
Caps-EffNet-B0	5.38	94.86 ± 1.36	94.79 ± 1.42	98.95 ± 0.37	0.931 ± 0.018	94.28 ± 1.45
EffNet-B0	4.05	95.71 ± 0.56	95.68 ± 0.58	99.48 ± 0.21	0.943 ± 0.007	95.30 ± 0.68

Model	Params (M)	Accuracy (%)	Macro-F1 (%)	Macro-AUC (%)	MCC	Macro-F1, dedup. (%)
ResNet50	23.60	95.30 ± 0.98	95.26 ± 1.01	99.50 ± 0.10	0.937 ± 0.013	94.69 ± 1.08

Note. Official test set (n = 1,311) unless marked; dedup. = deduplicated test set (n = 1,039). The VGG16 values include two degenerate runs that predicted a single class.

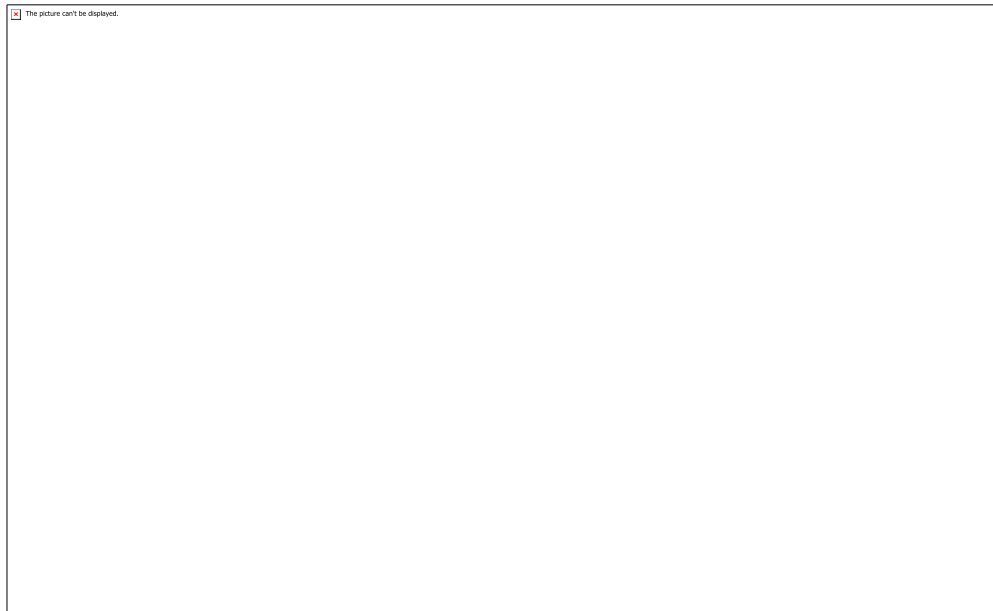


Figure 4. Macro-F1 on the official test set versus number of parameters (mean ± SD over five seeds); stars denote capsule heads

Effect of the Capsule Head

Table 3 isolates the effect of the classification head. On the light backbone, the capsule head outperformed both non-capsule heads in every seed. Relative to the dense head, macro-F1 increased by 7.16 percentage points (95% CI [2.03, 12.30]; paired t-test $p = .018$), and McNemar's test was significant in all five seeds, although the Holm-adjusted p value was .072. Relative to the pooling head, macro-F1 increased by 6.63 percentage points (95% CI [4.44, 8.83]; Holm-adjusted $p = .006$). These gains were achieved with 91.3% fewer parameters than the dense head, although Caps-Light had approximately twice as many parameters as Light + GAP. The pattern persisted on the deduplicated test set, with gains of 7.20 and 5.70 percentage points, respectively. On the ImageNet-pretrained backbones, the capsule head provided no measurable benefit: Caps-VGG16-B4 and VGG16-B4 + GAP differed by 0.08 percentage points ($p = .882$), and Caps-EffNet-B0 scored 0.88 percentage points below EffNet-B0 (95% CI [-2.78, 1.02]; $p = .266$), outperforming it in only two of five seeds. Caps-VGG16-B4 was, however, less variable across seeds than its counterpart (SD = 0.35 vs. 0.83 percentage points). The 7.16- and 6.63-percentage-point macro-F1 effects on the light backbone were accompanied by wide uncertainty intervals because only five seeds were available. These results demonstrate a reproducible within-experiment advantage for the capsule head on this specific backbone trained from scratch, but they do not isolate which internal capsule property (routing, vector representation, or effective capacity) caused the gain.

Table 3

Head-matched comparisons of macro-F1 on the official test set (clean protocol, five seeds)

Comparison	Δ macro-F1 (95% CI)	Wins	t-test p	Holm p	McNemar p < .05
Caps-Light vs. Light + Dense	+7.16 (+2.03, +12.30)	5/5	.018	.072	5/5
Caps-Light vs. Light + GAP	+6.63 (+4.44, +8.83)	5/5	.001	.006	5/5
Caps-VGG16-B4 vs. VGG16-B4 + GAP	+0.08 (-1.32, +1.47)	3/5	.882	.882	1/5
Caps-EffNet-B0 vs. EffNet-B0	-0.88 (-2.78, +1.02)	2/5	.266	.532	2/5
Caps-Light vs. First-draft CNN	+28.15 (-1.39, +57.69)	5/5	.057	.172	5/5

Note. Δ = capsule model minus partner, in percentage points. Holm correction across the five comparisons. For all comparisons the Wilcoxon signed-rank p was at least .0625, the minimum attainable with five pairs.

Error Patterns

Meningioma was the most difficult class for every model. Its mean recall was 60.26% for Caps-Light, 80.07% for Caps-VGG16-B4, 86.93% for Caps-EffNet-B0, and 88.76% for EffNet-B0, whereas pituitary tumors were recognized with recall above 98% by all three capsule models. The absolute confusion matrix for Caps-Light (Figure 5) shows that, on average, 80.4 of 306 meningioma images (26.3%) were classified as no tumor and 30.6 (10.0%) as pituitary tumor. It also shows that 41.0 of 300 glioma images (13.7%) were classified as meningioma. In contrast, fewer than one glioma or pituitary image per run was classified as no tumor. Thus, almost all tumor-to-no-tumor errors originated from meningioma, which also explains the lower precision of the no-tumor class for Caps-Light (82.87%).

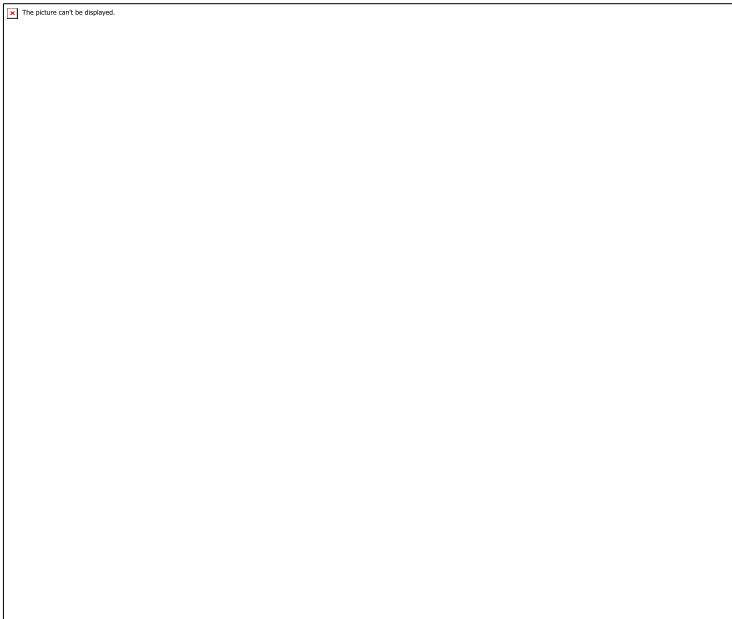


Figure 5. Confusion matrix of Caps-Light on the official test set under the clean protocol: mean absolute counts over five seeds with row percentages

Inflation Caused by Leakage

Training under the leaky protocol increased test accuracy for all three models examined (Table 4 and Figure 6a): by 6.11 percentage points for Caps-Light, 3.22 percentage points for EffNet-B0, and 6.33 percentage points for the first-draft CNN, whose results remained highly variable. EffNet-B0 achieved $98.93 \pm 0.23\%$ accuracy under the leaky protocol, compared with $95.71 \pm 0.56\%$ under the clean protocol. Validation accuracy was inflated to a lesser extent, by 1.23 percentage points for Caps-Light and 1.02 percentage points for EffNet-B0. To locate the source of the inflation, the test images were divided into those with a near-duplicate in the leaky training subset ($n = 859$, 65.5%) and those without one ($n = 452$). Models trained under the clean protocol, which had encountered neither subset during training, performed similarly on both, with accuracy differences of +1.48 percentage points for Caps-Light and +0.41 percentage points for EffNet-B0, indicating comparable difficulty (Figure 6b). Under the leaky protocol, the advantage on images with a training counterpart increased to +4.73 and +1.97 percentage points, respectively. However, accuracy on images without a counterpart also increased, from 84.73% to 88.72% for Caps-Light and from 95.44% to 97.64% for EffNet-B0.

Table 4
Test accuracy (%) under the clean and leaky protocols

Model	Clean	Leaky	Inflation	Twin, clean	No twin, clean	Twin, leaky	No twin, leaky
Caps-Light	85.71 ± 1.42	91.81 ± 2.79	+6.11	86.22	84.73	93.44	88.72
EffNet-B0	95.71 ± 0.56	98.93 ± 0.23	+3.22	95.86	95.44	99.61	97.64
First-draft CNN	61.94 ± 21.14	68.27 ± 16.17	+6.33	63.38	59.20	67.60	69.54

Note. Official test set. Twin = test image with a near-duplicate in the leaky training subset (n = 859); no twin, n = 452. Clean protocol: five seeds; leaky protocol: three seeds.

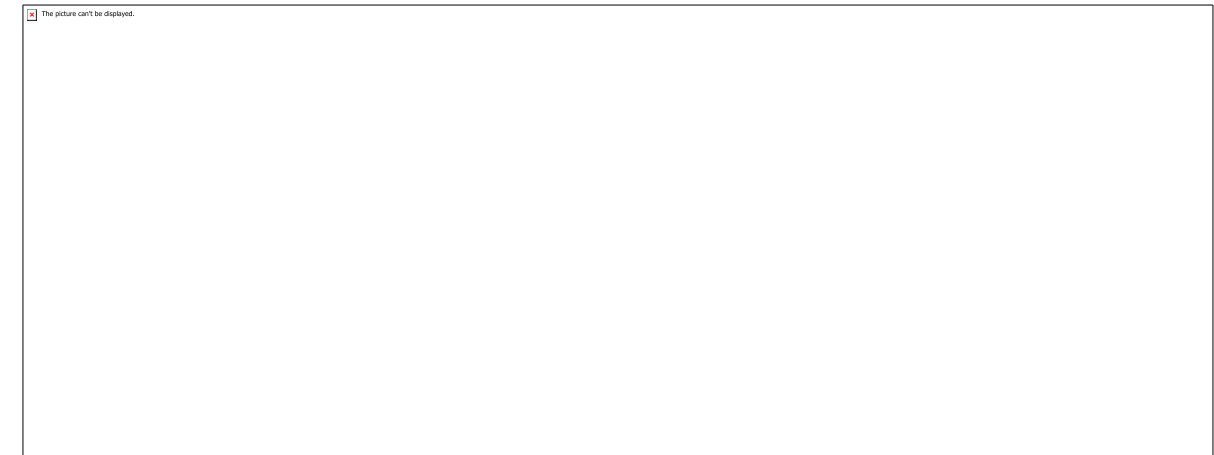


Figure 6. Effect of leakage: test accuracy under the leaky and clean protocols (a) and accuracy on test images with and without a near-duplicate in the leaky training subset (b)

Rotation Robustness and Computational Efficiency

Accuracy decreased with rotation for every model, with the steepest declines occurring beyond the $\pm 15^\circ$ range covered by augmentation (Figure 7). On the light backbone, Caps-Light remained more accurate than both comparators at every angle, by 9.45 percentage points over Light + Dense and 5.51 percentage points over Light + GAP at 90° . However, its accuracy decline from 0° to 90° (10.16 percentage points) was comparable to that of Light + GAP (10.13 percentage points; difference in decline = -0.04 percentage points, $p = .97$) and was not significantly smaller than that of Light + Dense (12.84 percentage points; $p = .15$). On VGG16-B4, the capsule head lost less accuracy than the pooling head at 45° (by 2.62 percentage points, $p = .005$) and 60° (by 3.39 percentage points, $p = .030$), but the difference was not significant at 90° ($p = .24$), where VGG16-B4 + GAP varied widely across seeds (SD = 8.03 percentage points). On EfficientNet-B0, the capsule head lost more accuracy than the pooling head at 60° (by 1.25 percentage points, $p = .008$) and 90° (by 2.25 percentage points, $p = .040$), with accuracy declining from 94.38% to 84.04% for the capsule model, compared with 95.40% to 87.31% for EffNet-B0. These 40 exploratory comparisons were not corrected for multiplicity.

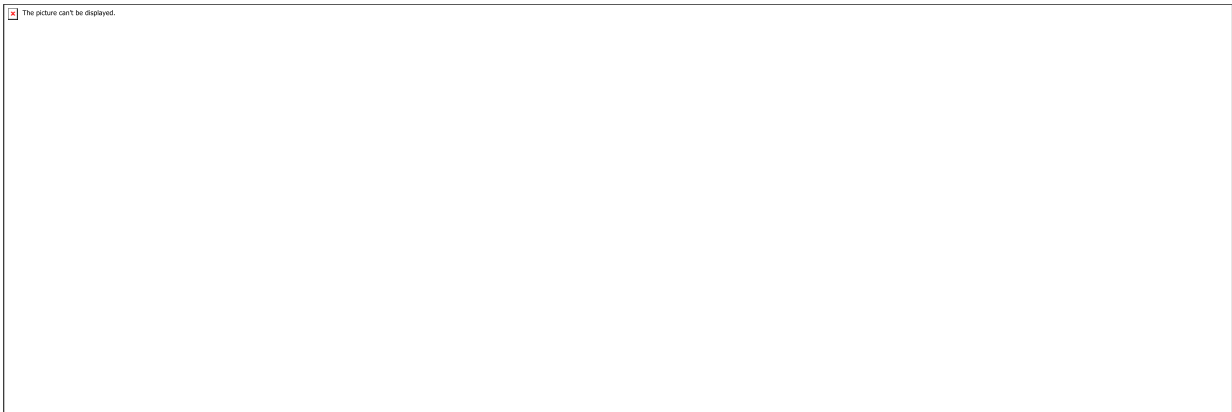


Figure 7. Accuracy on the rotated deduplicated test set for each head-matched group (mean $\pm$ SD over five seeds); the shaded area marks the rotation range used for augmentation

Table 5 summarizes computational cost. Relative to the dense head, the capsule head substantially reduced the parameter count and file size (2.35 vs. 26.87 million parameters; 9.1 vs. 102.6 MB) but left the FLOP count essentially unchanged (5.16 vs. 5.17 G), because the backbone dominated computation. Dynamic routing increased GPU latency: 2.09 ms per image for Caps-Light, compared with 1.76 ms for the dense head and 1.39 ms for the pooling head. On the CPU, Caps-Light was faster than Light + Dense (15.8 vs. 20.7 ms) and close to Light + GAP (14.7 ms). Every model trained within the available memory at a batch size of 16, with peak training memory ranging from 1,608 to 5,549 MB. FLOPs were a poor proxy for latency: EffNet-B0 required 0.79 G FLOPs, approximately 39 times fewer than VGG16, yet its GPU latency was 10.70 ms, compared with 4.25 ms for VGG16. However, EffNet-B0 was faster on the CPU (21.8 vs. 52.4 ms).

Table 5

Computational efficiency on the RTX 3060 Ti workstation

Model	Params (M)	Size (MB)	FLOPs (G)	GPU latency (ms)	GPU throughput (img/s)	CPU latency (ms)	Train memory (MB)	Time/epoch (s)
Caps-Light	2.35	9.1	5.16	2.09	1,608	15.8	1,608	16.3
Light + Dense	26.87	102.6	5.17	1.76	1,681	20.7	2,060	16.5
Light + GAP	1.18	4.6	5.12	1.39	1,741	14.7	1,653	15.8
First-draft CNN	59.16	225.8	27.93	5.10	418	73.1	5,549	47.4
Caps-VGG16-B4	9.40	35.9	28.02	3.99	626	48.4	3,923	36.4
VGG16-B4 + GAP	7.64	29.2	27.94	3.41	641	46.3	3,907	35.8
VGG16	14.72	56.2	30.71	4.25	584	52.4	3,957	39.5
Caps-EffNet-B0	5.38	21.0	0.85	10.34	901	23.1	2,513	23.6
EffNet-B0	4.05	16.0	0.79	10.70	951	21.8	2,433	23.0
ResNet50	23.60	90.4	7.72	9.61	619	44.6	4,061	32.2

Note. Latency for one 224×224 image after warm-up (mean of 300 GPU or 50 CPU calls); throughput at batch size 16; train memory = peak GPU memory during training at batch size 16.

Discussion

The audit indicates that the Brain Tumor MRI Dataset, as commonly used, does not provide an independent test of generalization. Nearly one-third of the training folder was linked to the test folder, and the linkage was concentrated in the no-tumor class, whose images contained many reprocessed copies. The comparison between the leaky and clean protocols shows that this overlap has measurable consequences. The observed accuracy inflation of 3.2–6.3 percentage points is smaller than the 29%–55% reported for slice-level splitting in neurodegenerative disease classification, because most test images in the present dataset remain genuinely distinct (Yagis et al., 2021). Nevertheless, the inflation is substantial relative to the differences on which architectural claims are usually based. It was also sufficient to raise the accuracy of a strong baseline from 95.7% to 98.9%, a range commonly reported for this dataset.

The analysis of test images with and without a training counterpart adds a nuance that deduplication alone cannot reveal. Part of the inflation is attributable to memorization, as reflected in the additional gain on images with a counterpart. Another part reflects the 46% larger training subset available under the leaky protocol, as reflected in the gain on images without a counterpart. Clean evaluation therefore lowers reported accuracy for two distinct reasons, and comparisons between cleaned results and leaky results in the literature conflate these effects. Validation leakage compounds the problem because more than half of the leaky validation images had a training counterpart, making model selection and early stopping optimistic as well.

The comparison with Lu et al. (2026), who retained 3,522 unique scans, shows that leakage-free evaluation is not defined by a single standard: different similarity criteria remove different amounts of data, and results obtained using differently cleaned versions of the dataset are not directly comparable. The label-conflict calibration proposed here offers a transparent and reproducible way to justify a similarity threshold because it relies on an observable property of genuine duplicates rather than an arbitrary cutoff. The same analysis revealed identical images

carrying different tumor labels, a form of label noise that no classifier can resolve and that should be addressed before a benchmark is constructed.

The head-matched ablation qualifies the view that capsule networks are generally advantageous for brain tumor MRI classification. When features had to be learned from approximately 3,100 training images, the capsule head produced a consistent gain of 6.6–7.2 percentage points over both conventional heads. This finding supports the aspect of (Afshar et al. (2018) argument concerning learning from limited data: routing part-level evidence into class capsules appears to provide a useful inductive bias when the backbone itself is weak. The same head provided no benefit on VGG16-B4 or EfficientNet-B0. One possible explanation is that ImageNet pretraining supplies more transferable and better-conditioned features, reducing the incremental influence of the classification head.

This interpretation is consistent with evidence that transfer learning in medical imaging benefits substantially from feature reuse and well-conditioned initialization (Raghu et al., 2019). The capsule head is therefore an option for settings in which pretrained weights cannot be used or a small model trained from scratch is required, rather than a general upgrade. Caps-Light, with a macro-F1 score of 84.8%, remained approximately 11 percentage points below EffNet-B0, despite EffNet-B0 having only 1.7 times as many parameters. Thus, the best accuracy per parameter in this study came from transfer learning, not from capsules. This explanation is theoretical and was not directly tested through feature-ablation or representation analyses in the present experiment.

The rotation experiment does not support the claim that capsule heads confer rotation robustness. The higher accuracy of Caps-Light under rotation reflected its higher starting accuracy because its decline matched that of the pooling head. On EfficientNet-B0, the capsule head degraded more. Only on VGG16-B4 did the capsule head exhibit a smaller decline at intermediate angles, and that uncorrected result should be regarded as a hypothesis. This mixed picture is consistent with Gu et al. (2021), who showed that the transformation and routing components of capsule networks do not, by themselves, confer robustness to affine transformations. In brain MRI, clinical slices are acquired in standardized orientations, so large in-plane rotations are uncommon. Augmentation offers a cheaper and more direct route to the invariance that is actually needed. The rotation test uses synthetic in-plane transformations and should not be equated with the full range of orientation, motion, sequence, scanner, or acquisition variability encountered in clinical MRI. Its interpretation is restricted to the specified 0°–90° image-rotation stress test.

The efficiency measurements also refine the meaning of lightweight. The capsule head reduces the parameter count and storage requirements, which matters for memory-constrained deployment and model distribution, but it does not substantially reduce computation. On the GPU, it is slower than a pooling head because routing is iterative. The finding that EfficientNet-B0 was slower than VGG16 on the GPU despite requiring far fewer FLOPs supports the argument of Ma et al. (2018) that FLOPs are an indirect metric and that efficiency must be measured on the target hardware. On the RTX 3060 Ti used here, every model trained within the available memory and processed an image in approximately 10 ms or less, so a consumer GPU of this class is sufficient for developing and running all evaluated architectures. The relevant design choice is the balance among accuracy, stability across seeds, and latency on the actual deployment device, rather than parameter count alone. Reporting such measurements also supports more resource-conscious research practices (Schwartz et al., 2020).

Equally important, from a clinical perspective, is the classification of approximately one-quarter of meningioma images as no tumor by Caps-Light. This represents an experimentally concerning false-negative pattern, but it cannot be translated directly into a patient-level missed-tumor rate. Even EfficientNet-B0 misclassified approximately 11% of meningioma images. Because the no-tumor images were drawn largely from a different source collection than the tumor images, some of this confusion may reflect acquisition differences rather than anatomical features, suggesting a potential form of shortcut learning Geirhos et al. (2020) that requires evaluation using source-matched data. None of the evaluated models should therefore be regarded as ready for clinical use. Clinical risk cannot be estimated from these slice-level benchmark errors because the study lacks patient-level aggregation, source-matched external cohorts, prospective workflow testing, and comparison with clinician performance.

These findings matter beyond the local context in which the study was conducted. The dataset is used worldwide, including in countries where specialist capacity is limited and where decision-support tools trained on public data may be especially attractive. In Indonesia, the low density of radiologists and their concentration in a few regions make such tools appealing precisely where inflated benchmark results could cause the most harm (Yunus, 2024). Leakage-free, multi-seed evaluation is therefore not merely a methodological refinement but a prerequisite for the responsible transfer of such models into health systems in Indonesia and other low- and middle-income settings, in line with international calls for more rigorous validation in medical imaging research (Varoquaux & Cheplygina, 2022).

The contribution of this study to that discourse lies in three results not previously presented together for this dataset: a quantified and decomposed estimate of leakage-related inflation, an objective criterion for selecting near-duplicate thresholds, and a controlled demonstration that the benefit of capsule heads depends on whether pretrained features are available. Accordingly, the immediate contribution is to benchmark design and evaluation rigor. Any clinical implementation would require patient-level external validation, prospective testing, calibration across institutions and scanners, and assessment of clinical utility before deployment.

Several limitations should be acknowledged. First, the dataset provides no patient identifiers, so leakage was detected through image similarity. This approach cannot guarantee that all slices from a given patient were separated and may occasionally group images from different patients with very similar anatomy. Second, all experiments used a single aggregated dataset of two-dimensional slices without external validation, and the deduplicated test images still came from the same source collections as the training images. Third, two seeds and two models were added after an interim inspection of the results. Although all runs were reported, the additional comparisons did not follow a fully preregistered design. Fourth, five seeds limit statistical power; the smallest attainable two-sided Wilcoxon p value was .0625, and the rotation analyses were exploratory and uncorrected for multiple comparisons. Fifth, one training recipe was applied to all models, which favored comparability but may have disadvantaged some architectures, as suggested by the collapse of full VGG16 in two runs and by one ResNet50 run that reached the 60-epoch limit. Finally, efficiency was measured on a single workstation, so absolute latency values will differ on other hardware. These limitations constrain different aspects of validity: the absence of patient identifiers prevents guaranteed patient-level separation; the use of a single aggregated dataset limits external generalizability; post-interim additions increase the risk of analysis flexibility; five seeds limit the precision of seed-level significance tests; a common training recipe may not represent each architecture's optimum; and single-workstation timing limits the generalizability of hardware measurements.

CONCLUSION

This study shows that the accuracy reported on a widely used brain tumor MRI benchmark depends substantially on how its data are split. After an objectively calibrated audit removed training images linked to the test set, the evaluated models lost 3.2–6.3 percentage points of accuracy. This decline arose both from the removal of memorized near-duplicates and from the smaller training set, two effects that must be distinguished when results from different studies are compared. Under this leakage-free, multi-seed design, capsule heads were not a general improvement. They delivered a consistent accuracy gain with far fewer parameters than a dense head when a lightweight network was trained from scratch, but they provided no improvement on ImageNet-pretrained backbones, did not consistently reduce sensitivity to rotation, and saved memory rather than computation.

The study therefore supports a conditional interpretation of capsule heads and recommends leakage-free, multi-seed evaluation with hardware-level reporting for benchmarks with similar aggregated public-data characteristics. Future research should validate these findings using patient-level datasets with external test cohorts, determine whether confusion between meningioma and no tumor reflects differences between source collections, examine capsule heads combined with medical-domain or self-supervised pretraining, and evaluate the most accurate models prospectively in clinical environments with limited radiology capacity. The first objective was addressed by the observed leakage-related accuracy inflation of 3.22–6.33 percentage points under the tested protocols. The second objective was addressed conditionally: the capsule head

improved macro-F1 by 6.63–7.16 percentage points on the light backbone trained from scratch but did not improve performance in the ImageNet-pretrained VGG16-B4 or EfficientNet-B0 comparisons.

ACKNOWLEDGEMENT

The authors thank the creators of the figshare, SARTAJ, and Br35H collections and the maintainer of the Brain Tumor MRI Dataset for making their data publicly available.

AUTHOR CONTRIBUTION STATEMENT

Maulana Kanigara: conceptualization, methodology, software, investigation, formal analysis, visualization, and writing of the original draft. Guruh Fajar Shidik: conceptualization, supervision, validation, and writing (review and editing).

REFERENCES

- Afshar, P., Mohammadi, A., & Plataniotis, K. N. (2018). Brain Tumor Type Classification via Capsule Networks. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 3129–3133. <https://doi.org/10.1109/ICIP.2018.8451379>
- Afshar, P., Plataniotis, K. N., & Mohammadi, A. (2019). Capsule Networks for Brain Tumor Classification Based on MRI Images and Coarse Tumor Boundaries. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1368–1372. <https://doi.org/10.1109/ICASSP.2019.8683759>
- Barz, B., & Denzler, J. (2020). Do We Train on Test Data? Purging CIFAR of Near-Duplicates. *Journal of Imaging*, 6(6), 41. <https://doi.org/10.3390/jimaging6060041>
- Bhuvaji, S. (2020). *Brain tumor classification (MRI)[Data set]*. Kaggle.
- Bouthillier, X. (2021). *Accounting for variance in machine learning benchmarks*. 747–769.
- Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., & Feng, Q. (2015). Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition. *PLOS ONE*, 10(10), e0140381. <https://doi.org/10.1371/journal.pone.0140381>
- Deepak, S., & Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*, 111, 103345. <https://doi.org/10.1016/j.compbiomed.2019.103345>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7), 1895–1923. <https://doi.org/10.1162/089976698300017197>
- Fateh, A., Rezvani, Y., Moayedi, S., Rezvani, S., Fateh, F., Fateh, M., & Abolghasemi, V. (2026). BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification. *Scientific Data*, 13(1), 361. <https://doi.org/10.1038/s41597-026-06753-y>
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673. <https://doi.org/10.1038/s42256-020-00257-z>
- Gu, J., Tresp, V., & Hu, H. (2021). Capsule Network is Not More Robust than Convolutional Network. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14304–14312. <https://doi.org/10.1109/CVPR46437.2021.01408>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2).
- Jabbar, A., Naseem, S., Mahmood, T., Saba, T., Alamri, F. S., & Rehman, A. (2023). Brain Tumor Detection and Multi-Grade Segmentation Through Hybrid Caps-VGGNet Model. *IEEE Access*, 11. <https://doi.org/10.1109/ACCESS.2023.3289224>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>

- Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
- Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8), 1231–1251. <https://doi.org/10.1093/neuonc/noab106>
- Lu, D., Zhang, Y., & Chaudhary, D. (2026). MTA-Swin: A Multi-Token Attention Swin Transformer for Brain Tumor Classification with Leakage-Free MRI Benchmarking. *Journal of Medical Systems*, 50(1), 100. <https://doi.org/10.1007/s10916-026-02418-w>
- Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet V2: Practical guidelines for efficient cnn architecture design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11218 LNCS. https://doi.org/10.1007/978-3-030-01264-9_8
- Nickparvar, M. (2021). *Brain tumor MRI dataset*. Kaggle.
- Ostrom, Q. T., Cioffi, G., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2021). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014-2018. *Neuro-Oncology*, 23. <https://doi.org/10.1093/neuonc/noab200>
- Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 32.
- Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems, 2017-December*.
- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. <https://doi.org/10.1145/3381831>
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.
- Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., & Lu, J. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75, 34–46. <https://doi.org/10.1016/j.compmedimag.2019.05.001>
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019, 2019-June*.
- Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digital Medicine*, 5(1), 48. <https://doi.org/10.1038/s41746-022-00592-y>
- Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., SCHEDA, R., Giannelli, M., Tessa, C., Citi, L., & Diciotti, S. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, 11(1), 22544. <https://doi.org/10.1038/s41598-021-01681-w>
- Yunus, R. E. (2024). Radiology Loading and Coverage Hours in Indonesia. *Korean Journal of Radiology*, 25(7), 597. <https://doi.org/10.3348/kjr.2024.0267>