Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Domain Transfer Learning for Brain Tumor Classification Under Limited MRI Data Regimes: A Fraction-Resolved Benchmark with Explicit Statistical and Methodological Caveats

Nakib Uddin Ahmed

nakibuddin33@gmail.com

Shahjalal University of Science and Technology

Azizur Rahman

Indiana Wesleyan University

Mehjabin Ferdous

St. Francis College

Md. Shahadat Hossain

Metropolitan University

Kadirur Rahman Chowdhury

Birmingham City University

Akhlak Uz Zaman Ashik

Metropolitan University

Research Article

Keywords: Transfer Learning, Brain Tumor Classification, MRI, EfficientNet, Low-Data Regime, Statistical Power, Explainable AI, Grad-CAM++

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10630234/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cross-Domain Transfer Learning for Brain Tumor Classification Under Limited MRI Data Regimes: A Fraction-Resolved Benchmark with Explicit Statistical and Methodological Caveats

Nakib Uddin Ahmed[1*], Azizur Rahman[2], Mehjabin Ferdous[3], Md. Shahadat Hossain[4], Kadirur Rahman Chowdhury[5], Akhlak Uz Zaman Ashik[6]

[1*]Department of Information Technology, Shahjalal University of Science and Technology, University Avenue, Sylhet, 3114, Bangladesh.
[2]Department of Computer Technology (CAPS), Indiana Wesleyan University, Marion, IN, 46953, USA.
[3]St. Francis College, 179 Livingston Street, Brooklyn, NY, 11201, USA.
[4]Department of Computer Science and Engineering, Metropolitan University, Bateshwar, Sylhet, 3104, Bangladesh.
[5]Birmingham City Business School, Birmingham City University, Birmingham, B4 7BD, United Kingdom.
[6]Department of Computer Science and Engineering, Metropolitan University, Bateshwar, Sylhet, 3104, Bangladesh.

*Corresponding author(s). E-mail(s): nakibuddin33@gmail.com;
Contributing authors: azizurusa22@gmail.com;
mehjabinferdousmu@gmail.com; shahadat.hossain.bio@gmail.com;
kadirurrahmanchowdhury@gmail.com; akhlak@metrouni.edu.bd;

Abstract

Brain tumors cause an estimated 251,000 deaths worldwide each year [1], yet the labeled MRI archives that deep learning classifiers typically require remain out of reach for many hospitals. We compared ImageNet-pretrained against randomly initialized EfficientNet-B0 for four-class brain tumor MRI classification across six training-data fractions (5–100% of a 5,600-image, class-balanced pool from the Nickparvar Kaggle collection), on a fixed 1,600-image test set with three seeds

per condition. Transfer learning produced a statistically confirmed improvement only under the most extreme data scarcity we tested (5%, 238 images: +11.4 points, 95% CI [11.0, 11.8] pp, $\boldsymbol{p < .001}$); at every larger fraction the fine-tuned model scored higher on average, but the difference was not statistically distinguishable from chance under our three-seed design. We also report two minor descriptive quantities, a Transfer Efficiency Index and a Minimum Viable Data Threshold, but are explicit that neither has a theoretical derivation or independent validation. Per-class and Grad-CAM++ analyses identify glioma as the most persistently misclassified class, with attention drifting toward "no tumor" under uncertainty, favoring a screen-then-refer deployment over an autonomous one. Because the dataset lacks confirmed patient-level identifiers, we cannot rule out same-patient slices in both partitions, the study's most consequential limitation. We close with a concrete roadmap: more seeds, additional backbones, patient-disjoint validation, and a learning-curve alternative to TEI, all needed before any number here should inform a real deployment decision.

Keywords: Transfer Learning, Brain Tumor Classification, MRI, EfficientNet, Low-Data Regime, Statistical Power, Explainable AI, Grad-CAM++

1 Introduction

Brain tumors are among the deadliest forms of cancer, with survival depending heavily on early and accurate diagnosis. The World Cancer Research Fund estimates approximately 308,000 new diagnoses of brain and central nervous system tumors annually, with over 251,000 deaths [1]. Prognosis and treatment planning depend heavily on tumor subtype, whether glioma, meningioma, or pituitary tumor, which makes accurate, timely classification clinically consequential.

Magnetic resonance imaging (MRI) is the standard modality for brain tumor evaluation, and convolutional neural networks (CNNs) are now widely used to classify MRI findings [2]. This approach has real promise, but the radiological expertise and densely labeled imaging archives that deep learning historically requires are both scarce outside well-resourced academic hospitals. The constraint is most acute in rural clinics, low- and middle-income countries, and healthcare systems serving underserved populations, where the labeling budgets assumed by most published brain tumor classifiers, typically several thousand annotated images, are simply unrealistic.

Transfer learning, initializing a network with weights learned on a large natural-image corpus such as ImageNet and then fine-tuning on the target medical task, is the standard response to this scarcity. Tajbakhsh et al. [3] found that fine-tuned, ImageNet-pretrained CNNs matched or exceeded randomly initialized networks across several medical imaging tasks, particularly with small samples. Raghu et al. [4], by contrast, found that on two large-scale medical imaging tasks the benefit of ImageNet transfer was smaller than commonly assumed once sufficient labeled data was available. These two findings are not necessarily in tension; they may simply describe different

points along a single curve relating transfer benefit to data budget. Testing that possibility requires measuring the benefit continuously as the data budget shrinks, which is the empirical question this paper addresses for brain tumor MRI classification.

This paper reports a completed, 36-run empirical sweep rather than a proposed design. Its contributions are:

- A four-class brain tumor MRI benchmark comparing ImageNet-pretrained and from-scratch EfficientNet-B0 across six training-data fractions (5–100%) and three random seeds, with paired statistical testing that distinguishes confirmed effects from directional trends rather than treating them alike, including the finding that only the smallest fraction reaches conventional significance, a result we treat as informative rather than as a shortcoming to minimize.
- A Grad-CAM++-based explainability analysis across data fractions and training regimes identifying a specific, clinically relevant failure mode (glioma cases drifting toward a "no tumor" prediction under uncertainty), together with an explicit accounting of the patient-level data leakage risk inherent to the dataset family used here, a risk that prior work on this dataset, including an earlier version of this study, did not address.
- Two descriptive summary quantities alongside the benchmark: a Transfer Efficiency Index (TEI) that normalizes the fine-tuned/scratch accuracy gap by $\log_{10}(N)$, and a Minimum Viable Data Threshold (MVDT) marking where fine-tuned macro-F1 first crosses an illustrative reference value. Both are offered as compact shorthand for a pattern already visible in the raw results, not as contributions in their own right.

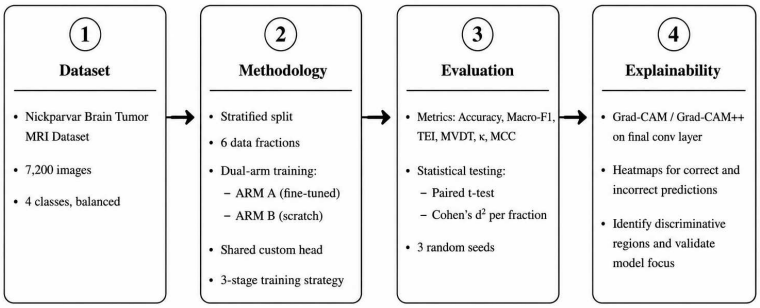


Fig. 1 Overview of the study pipeline: dataset preparation, the dual-arm fine-tuned/scratch training methodology, evaluation, and explainability stages described in this paper.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset and the patient-level independence caveat; Section 4 details the architecture, training protocol, and the proposed metrics with their limitations; Section 5 reports results with statistical qualification maintained throughout; Section 6 discusses clinical implications in provisional terms; Section 7 states limitations directly; and Section 8 concludes.

2 Related Work

2.1 Deep Learning for Brain Tumor Classification from MRI

Much published work on brain tumor MRI classification traces back to the contrast-enhanced MRI (CE-MRI) dataset introduced by Cheng et al. [5], who classified meningioma, glioma, and pituitary tumors using tumor-region-of-interest features, reaching 91.28% accuracy. Sultan, Salem, and Al-Atabany [6] later applied a purpose-built CNN to the same data, reporting higher accuracy and helping establish end-to-end feature learning as the dominant paradigm. Subsequent studies have used progressively larger CNN backbones, including VGG [7], ResNet [8], Inception-style networks [9], DenseNet [10], and EfficientNet [11], typically as a transfer-learning backbone rather than trained from scratch.

Most directly relevant here, Iqbal, Jaffar, and Jahangir [12] fine-tuned an EfficientNet-B0 backbone and reported approximately 99% accuracy on a four-class brain tumor MRI task, using the ∼7,000-image Kaggle aggregation curated by Nickparvar [13]. An earlier, preliminary version of this work instead used the smaller, 3,264-image dataset released by Bhuvaji et al. [14]. The present study uses a class-balanced 7,200-image release from the same Nickparvar family Iqbal et al. relied on, letting our full-data result be compared against theirs while additionally reporting how performance changes as the labeling budget shrinks below 100%, a gap almost the entire published literature leaves unaddressed. We return to the Nickparvar aggregation's provenance, specifically its lack of confirmed patient-level identifiers, in Section 3.4 and Section 7.

Across the broader literature, nearly all reported results assume access to a full labeled training set; very few studies report how performance degrades as the labeling budget is deliberately reduced, which is the empirical gap this paper targets.

2.2 Transfer Learning in Medical Imaging

The premise that features learned on ImageNet transfer usefully to other visual tasks goes back to Yosinski et al. [24], who showed that early convolutional layers learn general-purpose edge and texture detectors while later layers become increasingly task-specific, and to the broader transfer-learning framework laid out by Pan and Yang [25]. Whether that premise holds across the much larger domain gap between natural photographs and medical images is a separate empirical question, one Cheplygina, de Bruijne, and Pluim [26] survey at length for medical image analysis specifically.

Tajbakhsh et al. [3] provide an early empirical answer, reporting that across four medical imaging tasks, fine-tuning an ImageNet-pretrained CNN matches or exceeds an identically structured CNN trained from scratch, most clearly where labeled training data is scarce. Raghu et al. [4], in their Transfusion study, paint a less uniformly positive picture: on two large-scale tasks (diabetic retinopathy and Chest X-ray14), transfer learning provided only modest benefits over training from scratch once a large, fixed amount of data was available, with most of the benefit attributable to model capacity and to the earliest convolutional layers rather than to the higher-level features one might hope would transfer from natural images to medical ones.

Our sweep tests both claims on the same task at multiple data budgets rather than at one. Consistent with Tajbakhsh et al., a fine-tuning advantage is present in the point estimates at every fraction we tested. Consistent with Raghu et al., that advantage is statistically confirmed only at the smallest fraction, and its normalized magnitude falls toward a near-zero band once several hundred images per class are available. We read our results as broadly compatible with both prior findings rather than as adjudicating between them, since neither prior study measured the same task across a continuum of data budgets.

2.3 Low-Data Regimes in Medical Imaging

A broader literature addresses learning from limited labeled data through complementary strategies: few-shot learning [15], semi-supervised learning that incorporates unlabeled examples [16], and image-space data augmentation [17], the latter explicitly identifying medical imaging as a domain where big-data assumptions break down. Within medical imaging specifically, Maicas et al. [18] proposed a curriculum- and meta-learning scheme modeled on radiologist training, applied to breast DCE-MRI screening. Setio et al. [19] addressed limited 3-D CT training data for pulmonary nodule false-positive reduction using a multi-view 2-D convolutional architecture, an architecture-level response to data scarcity rather than a purely data-level one. To our knowledge, none of this work jointly targets four-class brain tumor MRI classification with a continuous, fraction-resolved, statistically qualified measure of how a pretraining benefit changes as the labeled-data budget shrinks.

2.4 Research Gap

The literature establishes that brain tumor MRI classification with deep CNNs performs strongly at full data scale on better-curated dataset variants, that ImageNet transfer generally helps in medical imaging but its benefit may depend on the data budget in ways not yet characterized continuously for this task, and that several general-purpose low-data strategies have not been benchmarked against a fraction-resolved transfer-learning baseline for brain tumor MRI specifically. We address this gap with a fraction-resolved sweep and paired statistical testing, while being more explicit than is typical in this literature about where our own statistical power runs out and where our proposed metrics remain descriptive rather than validated.

3 Dataset and Preprocessing

3.1 Dataset Description

All experiments use the Nickparvar "Brain Tumor MRI Dataset" [13], comprising 7,200 T1-weighted contrast-enhanced MRI slices across four classes, pre-partitioned by the dataset curator into a 5,600-image training pool and a 1,600-image test set. This release is class-balanced with exactly 1,400 training and 400 test images per class (Table 1), which removes class-frequency imbalance as a confound in the per-class comparisons reported in Section 5.

Table 1 Class distribution of the Nickparvar [13] dataset used in this study.

Class	Training Pool	Test Set	Total	% of Total
Glioma	1,400	400	1,800	25.0%
Meningioma	1,400	400	1,800	25.0%
No Tumor	1,400	400	1,800	25.0%
Pituitary	1,400	400	1,800	25.0%
Total	**5,600**	**1,600**	**7,200**	**100%**

Class-distribution charts (Supplementary Figure S1) confirm the exact 25/25/25/25 split across both partitions. Representative slices per class (Figure 2) show varied acquisition planes and, visibly, multiple imaging sources, consistent with how images accumulate in a real hospital archive rather than a single-protocol cohort, and a reminder of the provenance question taken up in Section 3.4. Per-class pixel-intensity histograms (Supplementary Figure S2) show broadly similar unimodal distributions, with the no-tumor class slightly narrower and lower-intensity, consistent with the absence of a contrast-enhancing lesion, though too small a difference to explain the per-class gaps in Section 5.3 on its own.

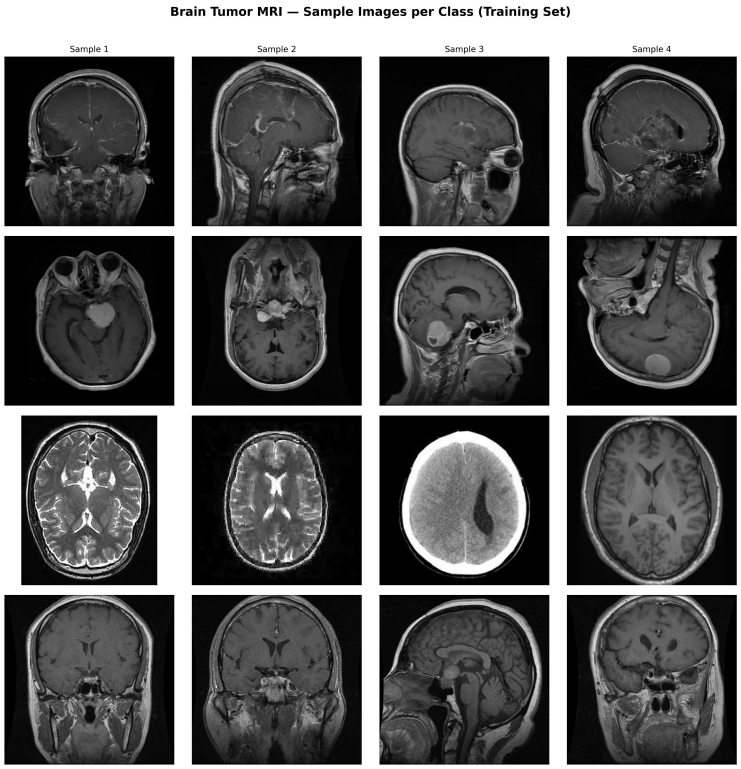


Fig. 2 Representative MRI slices, four samples per class, illustrating the varied acquisition planes and imaging sources present in the training pool.

3.2 Preprocessing and Augmentation

Slices are resized to 224×224 and normalized with ImageNet statistics (mean [0.485, 0.456, 0.406], SD [0.229, 0.224, 0.225]); single-channel intensities are replicated across three channels. No cropping, skull-stripping, or bias-field correction is applied, consistent with an end-to-end deployment scenario rather than a research-grade pipeline. The 100% fraction uses resize-then-crop, horizontal/vertical flips, rotation ($\pm 20°$), color jitter, affine translation/scaling, and random erasing; smaller fractions (5–50%) use a lighter policy (flip, $\pm 15°$ rotation, modest color jitter) to avoid distorting an already-limited signal. No augmentation is applied to validation or test images.

3.3 Data-Fraction Sampling Protocol

The 5,600-image training pool is split 85/15 (class-stratified) into a 4,760-image working training set and an 840-image validation set. From the working set we draw six class-stratified subsamples corresponding to 5, 10, 20, 30, 50, and 100% of it; validation and the fixed 1,600-image test set are held constant across all conditions, isolating training-data quantity as the variable of interest.

Table 2 Realized training-set size per fraction (validation: 840 images; test: 1,600 images, held fixed across all conditions). The right-hand column offers illustrative framing only and should not be read as evidence that our sample sizes match any specific real-world archive's case mix or acquisition protocol.

Fraction	N (training images)	Illustrative Analogue
5%	238	A clinic in its first months of digital archiving
10%	476	A newly digitized hospital archive
20%	952	An early-stage multi-site collaboration
30%	1,428	A growing regional imaging archive
50%	2,380	A mid-size hospital with an established archive
100%	4,760	The full stratified training pool

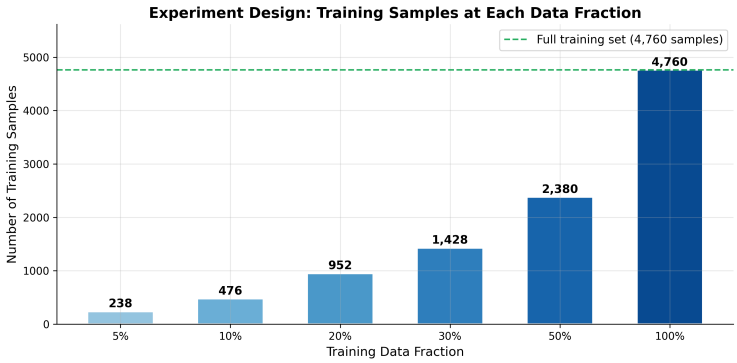


Fig. 3 Supplementary Figure S3. Training samples realized at each fraction, with the full 4,760-image working pool marked for reference.

3.4 A Caveat on Patient-Level Independence

A limitation we did not previously address, and now treat as central rather than incidental, is whether images from the same patient or imaging study can appear in both the training pool and the fixed test set. The Nickparvar [13] aggregation combines three earlier source collections and does not, to our knowledge, publish confirmed patient- or study-level identifiers alongside each slice. Slice-level splitting of this kind is a documented source of bias in brain MRI classification: Rumala [27] shows that subject-wise versus slice-wise splitting of longitudinal brain MRI can change which model appears to perform best, because slice-wise splits let a network learn to recognize the subject rather than the pathology, and Yagis et al. [28], working with 2-D CNNs on T1-weighted brain MRI, found that slice-level cross-validation inflated reported test accuracy by 29 to 55 percentage points relative to a subject-level split, depending on the dataset and disease studied. Neither study used the Nickparvar collection, so we cannot transfer their exact inflation figures to our numbers, but they establish that the mechanism we are flagging is real, well documented, and can be large rather than a minor technicality.

We could not verify, from the metadata publicly distributed with the Nickparvar dataset, whether such leakage is present in our specific 5,600/1,600 partition. If it is, within-patient correlation across near-duplicate slices would tend to inflate every accuracy and macro-F1 figure we report, since a held-out test slice would then be less genuinely held out than a patient-disjoint design would provide. This risk applies uniformly across all six fractions and both regimes, since the same fixed test set is used throughout, so it is unlikely to change the qualitative comparison between fine-tuned and scratch models in Section 5.1; but it directly affects how the absolute performance levels in Table 4, and specifically the values underlying the Minimum Viable Data Threshold in Section 5.5, should be read.

We return to this in Section 7, where we recommend that anyone attempting to replicate the specific numerical thresholds reported here first confirm, or construct, a patient-disjoint split, for example using an alternative dataset release with study-level identifiers, or a prospective single-site cohort where patient identity is known by construction.

4 Methodology

Figure 4 summarizes the pipeline described in this section: dataset partitioning, the dual-arm (pretrained vs. scratch) training setup, the three-stage unfreezing schedule, and the downstream evaluation and explainability steps.

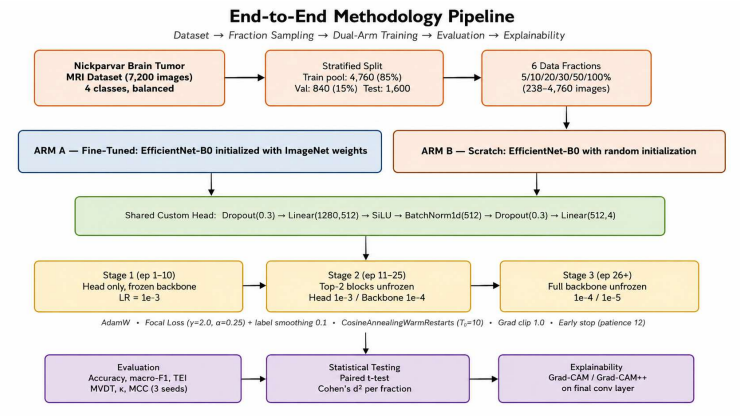


Fig. 4 End-to-end methodology pipeline: dataset partitioning, fraction sampling, the dual-arm (pretrained vs. scratch) EfficientNet-B0 setup with a shared custom head, the three-stage gradual-unfreezing schedule with its optimizer and loss settings, and the downstream evaluation and explainability steps described in Sections 4.1–4.5.

4.1 Model Architecture

We use EfficientNet-B0 [11] (5.3M parameters at 224×224 input resolution) as the backbone for both arms, chosen for its accuracy-per-parameter efficiency, which we consider relevant to modest clinical computing budgets, though we did not benchmark this claim directly (see Section 7). The ImageNet classification head is replaced with: Dropout(0.3) [29] → Linear(1280, 512) → SiLU → BatchNorm1d(512) [30] → Dropout(0.3) → Linear(512, 4), for 4,666,496 total parameters, of which 658,948 belong to the head. The fine-tuned arm initializes from timm's weights pretrained on the ImageNet corpus [31]; the scratch arm uses random initialization. Both arms share an identical head, optimizer, learning-rate schedule, and data pipeline, so that initialization is the only variable that differs between them.

4.2 Training Protocol

Training proceeds in three gradual-unfreezing phases:

- **Phase 1 (epochs 1–10):** backbone frozen; only the head is trained, at learning rate 1×10^{-3}.
- **Phase 2 (epochs 11–25):** the top two MBConv blocks are unfrozen at 1×10^{-4}, while the head continues at 1×10^{-3}.
- **Phase 3 (epoch 26 onward):** the full backbone is unfrozen, with newly unfrozen layers trained at 1×10^{-4} and previously unfrozen layers at 1×10^{-5}.

We use the AdamW optimizer (weight decay 1×10^{-4}) [32], gradient clipping at max norm 1.0, and a CosineAnnealingWarmRestarts scheduler ($T_0 = 10$) [33], restarted at each phase transition. Early stopping uses a patience of 12 epochs on validation loss, and the same protocol is applied identically to the scratch arm.

4.3 Loss Function

We use multi-class focal loss [34] with label smoothing [35] in place of plain cross-entropy:

$$\mathrm{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \cdot \log(p_t), \tag{1}$$

$\gamma = 2.0$, $\alpha = 0.25$, label smoothing $= 0.1$.

Focal loss down-weights easy, confidently correct examples and concentrates gradient signal on hard examples; it was originally proposed for dense object detection [34], where foreground/background imbalance is extreme, but the same down-weighting logic applies to any classification task with unevenly difficult classes. We use it here as a direct, mechanism-level lever for the uneven per-class difficulty documented in Section 5.3: it automatically up-weights harder glioma examples relative to flat cross-entropy, though we did not run a cross-entropy ablation to isolate its independent contribution from the other protocol choices (see Section 7).

4.4 Two Descriptive Summary Statistics

We use two derived quantities as compact shorthand for the sweep. Neither is a validated metric or theoretically derived; both are convenient summaries of a pattern already visible in Table 4, and we flag their status once here rather than repeating it throughout.

Transfer Efficiency Index (TEI). For training-set size N with fine-tuned and scratch accuracy $\mathrm{Acc}_{\mathrm{pre}}$ and $\mathrm{Acc}_{\mathrm{scr}}$:

$$\mathrm{TEI}(N) = \frac{\mathrm{Acc}_{\mathrm{pre}} - \mathrm{Acc}_{\mathrm{scr}}}{\log_{10}(N)} \tag{2}$$

The $\log_{10}(N)$ denominator was a convenient normalization, not a derived one; Table 3 recomputes the same accuracy gaps under three alternative normalizations ($\ln N$, $\sqrt{N}$, $1/N$) and shows the qualitative pattern (largest gap-per-image at 5%, smallest at 30%) holds regardless of the choice.

Table 3 Sensitivity of the accuracy-gap normalization to the choice of denominator (recomputed from Table 4, no new experiments).

Fraction	N	Raw gap	TEI, $\log_{10}(N)$	TEI, $\ln(N)$	TEI, $\sqrt{N}$
5%	238	0.1142	0.0480	0.0209	0.00740
10%	476	0.0258	0.0096	0.0042	0.00118
20%	952	0.0275	0.0092	0.0040	0.00089
30%	1,428	0.0040	0.0013	0.0005	0.00010
50%	2,380	0.0150	0.0044	0.0019	0.00031
100%	4,760	0.0077	0.0021	0.0009	0.00011

Note: TEI, $1/N$ ($\times 1000$) values, in the same fraction order: 0.4797, 0.0543, 0.0289, 0.0028, 0.0063, 0.0016.

Minimum Viable Data Threshold (MVDT). MVDT is the smallest fraction at which fine-tuned macro-F1 first reaches a chosen reference value, set here at

0.85. This threshold is a round number we chose for illustration, not a value drawn from a clinical or regulatory source; the CLAIM checklist [36] and the FDA/Health Canada/MHRA Good Machine Learning Practice principles [37] ask for clear reporting of subgroup performance but specify no numerical cutoff. MVDT should be read as the fraction at which our sweep crossed a pre-chosen number, not as a labeling-budget recommendation.

4.5 Statistical and Explainability Analysis

We treat the three seeds per fraction as paired observations and report a paired t-test and paired Cohen's d for the accuracy gap at each fraction, computing exact p-values given the small n. We compute Cohen's kappa [20] and Matthews correlation coefficient [21] from representative confusion matrices as agreement-quality checks. We use both Grad-CAM [23] and Grad-CAM++ [22] on the backbone's final convolutional layer to visualize prediction-driving regions and audit misclassifications. Three seeds is a small basis for inference. Colas, Sigaud, and Oudeyer [38], building on the broader reproducibility concerns raised by Henderson et al. [39] for deep reinforcement learning, show that the number of seeds needed to reliably detect a given effect size follows directly from standard power-analysis formulas [40, 41], and that small-seed comparisons are well powered only for large effects, exactly the pattern we see in Table 6, where the one large effect (5%) is easily detected and the smaller ones are not. We discuss the consequences of this directly in Section 5.4 rather than only noting it here.

4.6 Experimental Design

The complete sweep comprises 36 runs: 6 fractions $\times$ 2 regimes $\times$ 3 seeds, all evaluated on the same fixed 1,600-image test set, implemented in PyTorch 2.7.1 with CUDA and timm 1.0.27. No experiments were added or removed for this revision.

5 Experimental Results

All performances are test-set results ($n = 1,600$); values are mean $\pm$ SD across three seeds unless stated otherwise. Consistent with Section 3.4, these are observed performance on this specific partition, not a claim of generalization to a confirmed patient-disjoint cohort.

5.1 Classification Performance Across Data Fractions

The fine-tuned models produced a higher mean test accuracy than the scratch models at every fraction tested (Table 4), with the gap largest at 5% (11.4 points) and smallest at 100% (0.8 points). As Section 5.4 details, only the 5% comparison reaches conventional significance under our three-seed paired design; the remaining fractions are directional trends, not confirmed effects.

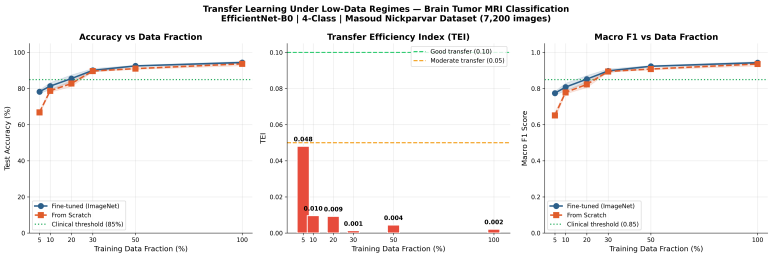


Fig. 5 Test accuracy, TEI, and macro-F1 as a function of training-data fraction for the fine-tuned and scratch models (1 SD bands shown; the 0.85 line is the illustrative macro-F1 reference discussed in Section 4.4, not a validated clinical threshold).

Table 4 Results across all six fractions (mean ± SD, 3 seeds). Only the 5% row is backed by a statistically significant paired comparison (see Table 6); the remaining rows should be read as consistent-direction point estimates.

Fraction	N	Acc (FT) %	Acc (Scratch) %	F1 (FT)	F1 (Scratch)	TEI
5%	238	78.27 ± 0.95	66.85 ± 0.96	0.7745	0.6516	0.0480
10%	476	81.35 ± 1.93	78.77 ± 0.79	0.8082	0.7790	0.0096
20%	952	85.56 ± 1.99	82.81 ± 1.64	0.8525	0.8233	0.0092
30%	1,428	90.10 ± 0.99	89.71 ± 0.40	0.8982	0.8949	0.0013
50%	2,380	92.56 ± 0.41	91.06 ± 0.44	0.9237	0.9083	0.0044
100%	4,760	94.46 ± 0.56	93.69 ± 1.04	0.9436	0.9360	0.0021

5.2 TEI Across Fractions

TEI falls from 0.048 at 5% to 0.0096 at 10%, then declines further and fluctuates near zero through 30–100% (Table 4; sensitivity in Table 3). This is consistent with a scratch network having little opportunity to learn generic low-level filters when N is smallest, and catching up once focal loss and augmentation give it a few hundred images per class to work with, though this remains a pattern rather than a validated causal mechanism.

5.3 Per-Class Performance and the Glioma Problem

Recall for glioma is the lowest of the four classes at every fraction analyzed (65.75% at 10%, 77.00% at 50%, 80.00% at 100%), despite glioma having the highest precision of any class at 100% data (99.38%). In other words, the model is rarely wrong when it does predict glioma, but still fails to detect roughly one in five true glioma cases at full data.

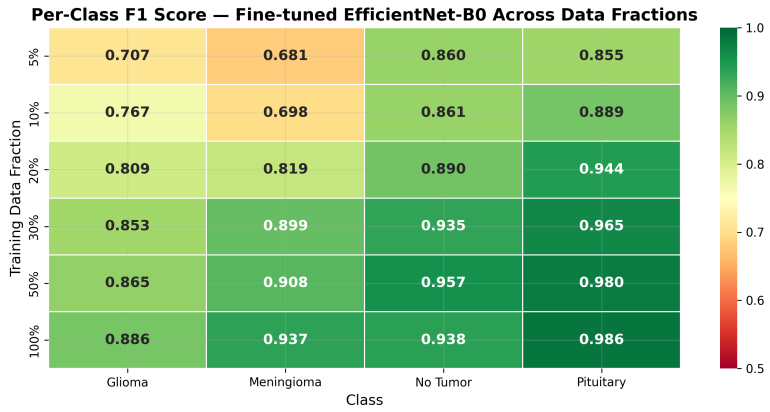


Fig. 6 Per-class F1 heatmap across all six fractions. Glioma and meningioma are visibly the hardest classes at low fractions, converging toward pituitary/no-tumor performance by 30–50% data; glioma remains the single lowest-F1 class throughout.

Table 5 Per-class precision, recall, and F1 at three fractions (support: 400 images per class).

Fraction	Class	Precision	Recall	F1
10%	Glioma	0.9196	0.6575	0.7668
10%	Meningioma	0.7147	0.6825	0.6982
10%	No Tumor	0.7650	0.9850	0.8612
10%	Pituitary	0.8705	0.9075	0.8886
50%	Glioma	0.9872	0.7700	0.8652
50%	Meningioma	0.8562	0.9675	0.9085
50%	No Tumor	0.9194	0.9975	0.9568
50%	Pituitary	0.9776	0.9825	0.9800
100%	Glioma	0.9938	0.8000	0.8864
100%	Meningioma	0.9085	0.9675	0.9370
100%	No Tumor	0.8830	1.0000	0.9379
100%	Pituitary	0.9875	0.9850	0.9862

At 100% data, of 400 true glioma cases, 34 are predicted meningioma and 44 no-tumor (19.5% of glioma cases overall). Because the dataset is class-balanced, this is not a class-frequency artifact; we consider glioma's radiologically heterogeneous appearance the more likely contributor, discussed further in Section 6.

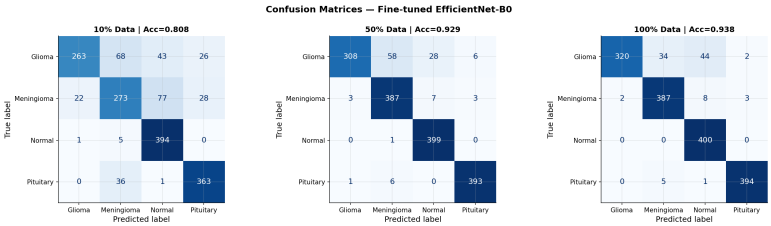


Fig. 7 Confusion matrices, fine-tuned model, at 10/50/100% data (one representative seed per fraction). No-tumor reaches perfect recall (400/400) at 100% data; glioma is the only class with meaningful off-diagonal error mass at full data.

5.4 Statistical Significance and Confidence Intervals: What the Data Actually Supports

The fine-tuned advantage reaches conventional significance only at 5% ($p < .001$, paired Cohen's $d_z = 76.0$); no other fraction reaches $p < .05$ (Table 6). Framed as a single sentence, our result is: transfer learning produced a statistically confirmed improvement only under the most extreme data scarcity we tested (238 images); at every larger fraction the observed advantage is a plausible but statistically unconfirmed point estimate. We use that sentence, not "transfer learning consistently outperformed," as the paper's actual headline finding.

Table 6 adds 95% confidence intervals for the paired accuracy difference, computed from the reported mean and t-statistic (SE = mean/t; CI = mean ± 4.303·SE, the $df = 2$ critical value). This is an analytic interval derived from our existing summary statistics, not a bootstrap over raw per-seed values, since we did not retain individual seed-level accuracies in a form that supports resampling, which we flag as a reproducibility gap in Section 7. The interval still conveys the same information as $p > .05$, in units readers can act on directly.

Table 6 Paired fine-tuned vs. scratch accuracy, $df = 2$, with 95% confidence intervals (pp = percentage points). Only the 5% interval excludes zero; every other interval is wide enough to be consistent with no real difference at all.

Fraction	**Mean Δ (FT−Scratch)**	**95% CI (pp)**	$t(2)$	p	**Cohen's d_z**
5%	+11.41 pp	[11.04, 11.78]	131.69	$< .001^{***}$	76.03
10%	+2.58 pp	[−5.24, 10.40]	1.42	.290	0.82
20%	+2.75 pp	[−4.60, 10.10]	1.61	.248	0.93
30%	+0.40 pp	[−1.47, 2.27]	0.92	.454	0.53
50%	+1.50 pp	[−0.59, 3.59]	3.09	.091	1.79
100%	+0.77 pp	[−1.86, 3.40]	1.26	.334	0.73

Why is d_z so large specifically at 5%? Both arms draw from the same seeded 238-image subsample, so most of the variability from which images were drawn cancels out in the paired difference even though it still shows up in each arm's marginal SD (Table 4: FT = 0.95, scratch = 0.96), consistent with Colas et al.'s [38] point that paired designs can inflate apparent effect size relative to independent-groups designs. An independent-groups replication would confirm this directly (Section 7).

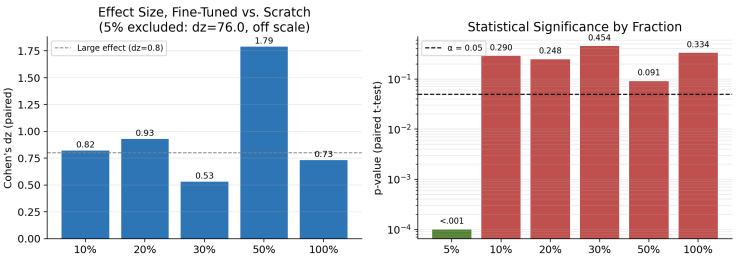


Fig. 8 Paired-seed effect size (Cohen's d_z) and significance (p-value, log scale) across fractions, computed from Table 6. The 5% fraction ($d_z = 76.0$) is displayed on its own scale to keep the remaining bars legible; its p-value is plotted at $< .001$ in the right panel. See the surrounding text for why this value should not be read at face value as an unusually strong biological effect.

Table 7 Chance-corrected agreement metrics from the confusion matrices in Figure 7. Kappa and MCC track accuracy and macro-F1 closely, as expected on a class-balanced dataset; reported mainly as a robustness check.

Fraction	Accuracy	Macro-F1	Cohen's κ	MCC
10%	0.8081	0.8037	0.7442	0.7494
50%	0.9294	0.9276	0.9058	0.9086
100%	0.9381	0.9369	0.9175	0.9198

Given that only one of six fractions clears conventional significance with three seeds, we consider the honest summary of Section 5.1–Section 5.4 to be: a fine-tuning advantage is clearly present and well-supported at the smallest data budget tested; at every larger budget we tested, the point estimate continues to favor fine-tuning, but our design cannot distinguish that from chance, and we do not treat it as confirmed.

5.5 Minimum Viable Data Threshold

The macro-F1 ≥ 0.85 reference from Section 4.4 is first reached at 20% of the training pool (952 images, F1 = 0.8525) and remains met at every larger fraction. As noted above, 952 images is the point at which our sweep crossed a pre-chosen reference value, not a validated labeling-budget recommendation. A lower reference (macro-F1 > 0.60) is already exceeded at 5% (F1 = 0.7745), illustrating the shape of the curve.

5.6 Training Dynamics

Validation loss and accuracy track training loss and accuracy closely at every fraction, with visible improvements at the Stage 2 (epoch 11) and Stage 3 (epoch 26) transitions and no persistent training-validation gap. We had, in an earlier and smaller pilot on a different, imbalanced dataset variant, observed more validation-test divergence; the improvement here is plausibly attributable to the class-balanced dataset and the conservative staged learning-rate schedule, though we did not run a controlled comparison isolating which of the two factors matters more.

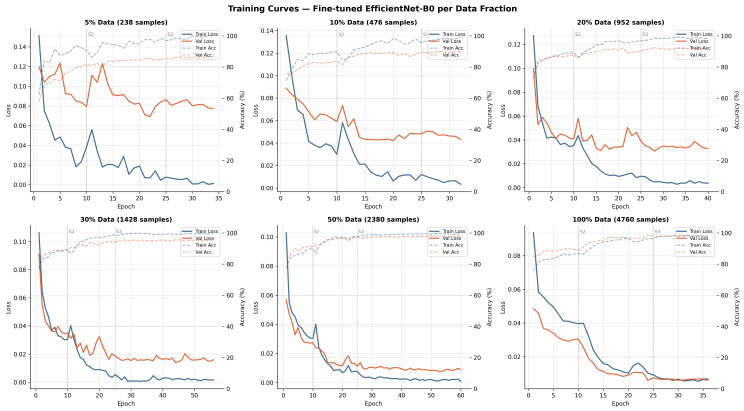


Fig. 9 Training and validation loss/accuracy by data fraction; Stage transitions marked at epochs 10 and 25. Lower data fractions show noisier validation curves and earlier stopping; higher fractions converge more smoothly.

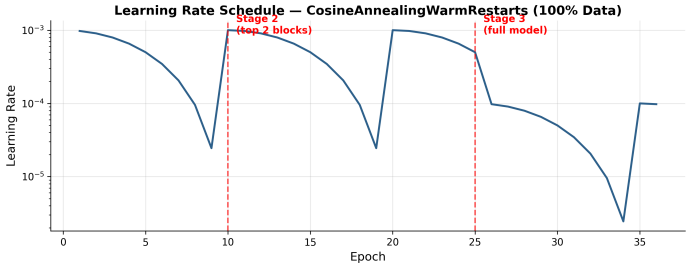


Fig. 10 Learning-rate schedule (100% data fraction), showing the CosineAnnealingWarmRestarts sawtooth pattern with resets at each gradual-unfreezing stage boundary.

5.7 Explainability: Grad-CAM++ Analysis

At 10% data, Grad-CAM++ attention is diffuse across all four classes and frequently spreads into skull and background regions. By 50–100% data, attention for meningioma, no-tumor, and pituitary concentrates more tightly on anatomically plausible regions; glioma attention improves over the same range but remains comparatively less well-localized even at full data, consistent with its lower recall.

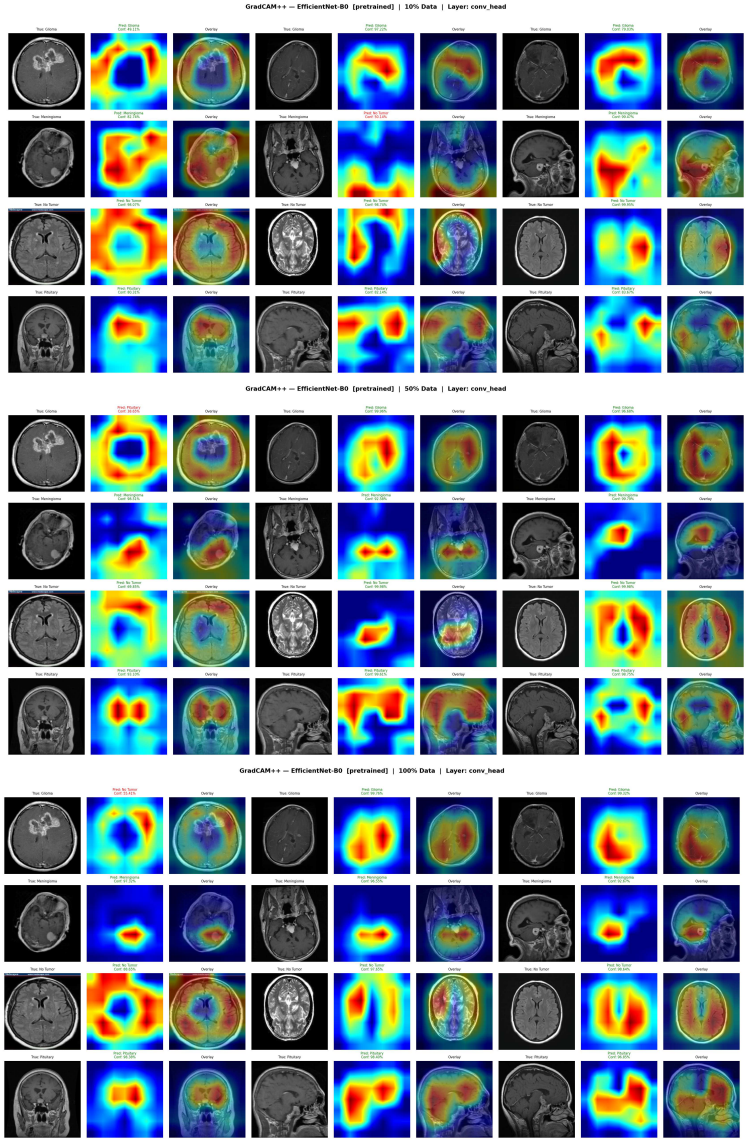


Fig. 11 Grad-CAM++ attention grids at 10/50/100% data (top to bottom), three samples per class. Attention sharpens and localizes more tightly as training data increases, most visibly between 10% and 50%.

Misclassified glioma cases show attention displaced away from the visible lesion toward regions resembling the model's typical no-tumor or meningioma attention pattern. When uncertain about a genuine glioma, the model appears to drift toward a "no tumor" reading rather than remaining anchored on an ambiguous lesion. We consider this the single most clinically relevant pattern in our explainability analysis, since a

false “no tumor” prediction is the error direction a screening deployment can least afford (Section 6).

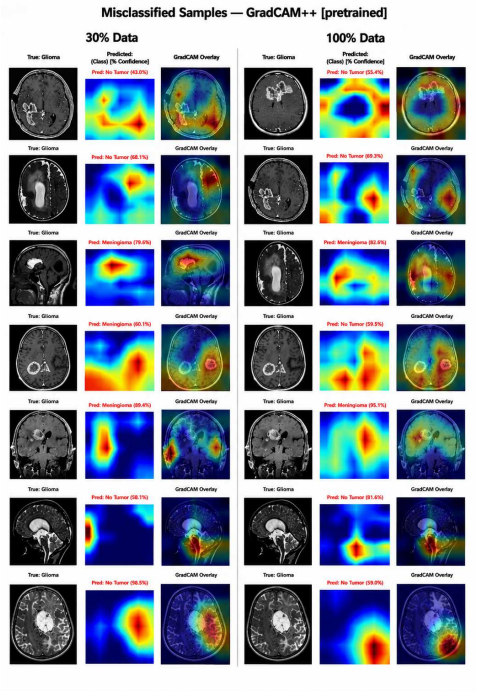


Fig. 12 Misclassified glioma cases at 30% and 100% data. Attention is visibly displaced away from the lesion in most cases shown.

Tracking one representative case per class from 10% to 100% data, three of four classes are classified correctly and increasingly confidently throughout; the glioma example shown is misclassified until 100% data, with attention only weakly anchored on the lesion even in the correct full-data prediction.

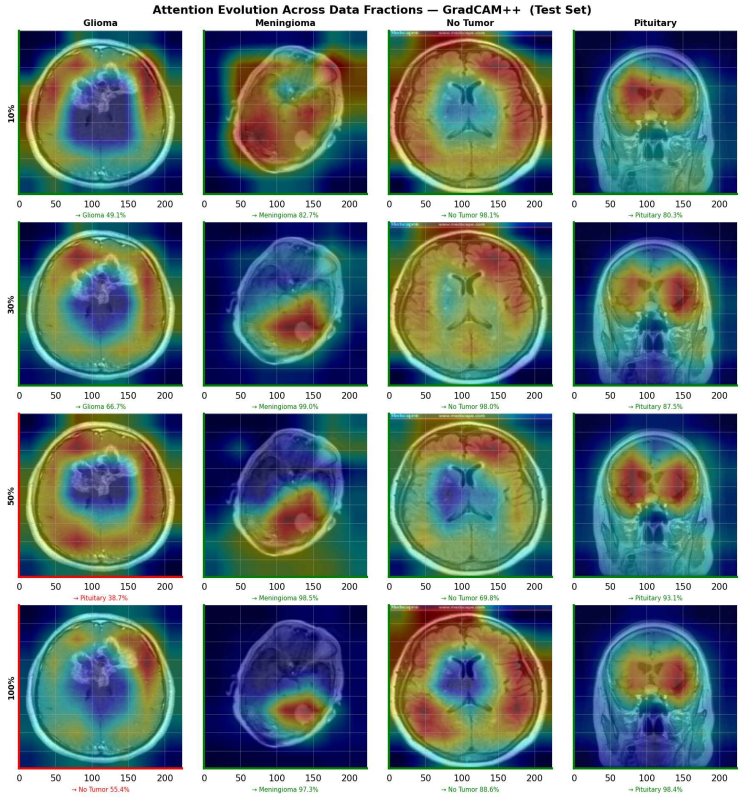


Fig. 13 Attention evolution per class as training data increases from 10% to 100%. Border color indicates correct (green) or incorrect (red) prediction.

Comparing pretrained and scratch attention on the same 30% subsample, the pretrained model's attention concentrates more on central, anatomically coherent structures, while the scratch model's attention extends more often into skull and background regions, a visual pattern consistent with why ImageNet pretraining helps despite the domain gap between natural photographs and MRI.

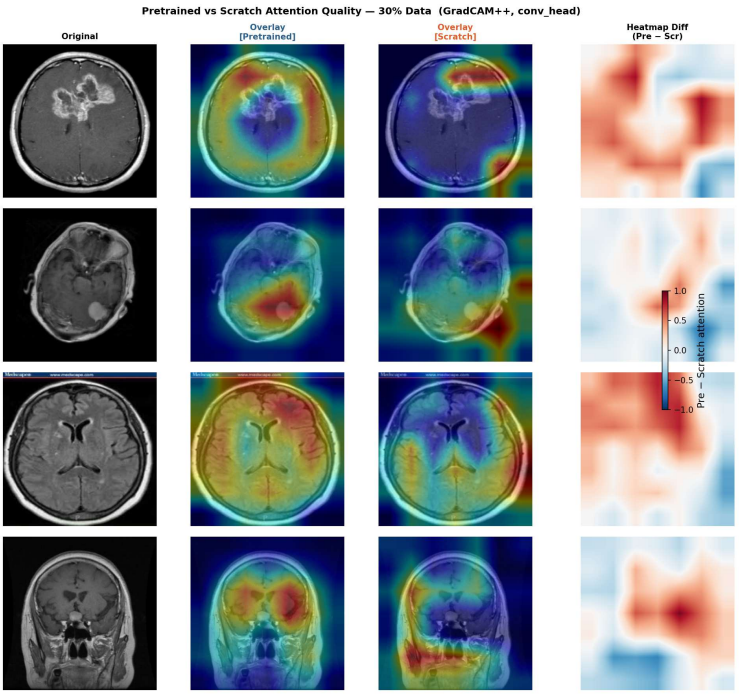


Fig. 14 Attention comparison, fine-tuned vs. scratch, same 30% subsample. Pretrained attention stays more consistently within central brain tissue.

5.8 Comparison with Prior Work

Our full-data result (94.46% accuracy, macro-F1 0.9436) is approximately 4.5 points below Iqbal et al.'s [12] reported ∼99% on the same dataset family. We attribute the gap tentatively to differences in training length, augmentation policy, and the exact data partition rather than to fine-tuning strategy per se, since our protocol does unfreeze the full backbone by epoch 26; but we note, per Section 3.4, that neither figure has been verified against a confirmed patient-disjoint partition, so a precise, apples-to-apples comparison is not currently possible. We view our full-data number as an independently obtained benchmark and, more importantly, as the fixed upper-data reference against which the reduced-fraction results in Table 4 should be read, while keeping in mind that both numbers may be subject to the leakage risk described above.

Table 8 Comparison with prior studies. None of the studies listed, including this one, has confirmed a patient-disjoint train/test split for the Nickparvar-family datasets; absolute accuracy figures across this table should be compared cautiously in light of Section 3.4.

Study	Method	Classes	Accuracy	Dataset
Cheng et al. [5]	ROI features + BoW/GLCM	3	91.28%	CE-MRI (3,064)
Sultan et al. [6]	Custom CNN	3	∼96.1%	CE-MRI (3,064)
Iqbal et al. [12]	EfficientNet-B0 (full)	4	∼99%	Nickparvar (∼7k)
This study (100%)	EfficientNet-B0 (3-stage)	4	94.46% (F1 0.94)	Nickparvar (7,200)
This study (20%, MVDT)	EfficientNet-B0 (3-stage)	4	85.56% (F1 0.85)	Nickparvar (952)

6 Discussion

6.1 Why the Efficiency Advantage Concentrates at the Smallest Fraction

TEI is largest at 5% and compresses rapidly by 30% (Section 5.2), matching the one fraction where the fine-tuning advantage is statistically confirmed (Section 5.4). This is consistent with a bias–variance argument: a scratch network must learn every level of visual abstraction from the training signal alone; with only 238 images there may not be enough signal to learn even low-level edge and texture detectors, so a pretrained model's head start dominates. Once a fraction reaches a few hundred images per class, focal loss and augmentation let the scratch model catch up on low-level features, compressing the normalized transfer benefit even as a modest raw accuracy edge persists. This reading helps reconcile Tajbakhsh et al. [3] with Raghu et al. [4] as different points on the same curve rather than contradictory findings.

6.2 Why EfficientNet-B0 May Suit Limited-Data MRI Classification

EfficientNet-B0's 5.3M parameters, roughly an order of magnitude fewer than ResNet-50 or VGG-16, plausibly reduce overfitting risk when only 238–952 images are available, and its depthwise-separable convolutions give early, ImageNet-transferable layers a larger share of the network than in deeper architectures (Figure 14). Whether this holds for ResNet-50 [8], ConvNeXt [43], or ViT [42] remains untested.

6.3 The Narrowing Gap and Clinical Implications

The raw accuracy gap falls from 11.4 points at 5% to under 1 point at 100%, consistent with a bias–variance account in which a strong prior on useful representations matters most where task data cannot estimate that representation reliably on its own. If this pattern holds under replication, transfer learning would be disproportionately valuable exactly where resource-constrained hospitals sit on the data curve (238–952 images).

We do not recommend fully autonomous deployment at any fraction tested, given glioma's persistently lower recall and its tendency to be misclassified as "no tumor" under uncertainty (Section 5.7), an error direction a screening deployment cannot tolerate. A two-stage screen-then-refer architecture, a high-recall flag reviewed by a radiologist rather than a standalone diagnostic tool, is the more defensible use of a model like this one.

6.4 Extending Prior Comparisons

Cheng et al. [5], Sultan et al. [6], and Iqbal et al. [12] each establish strong full-data performance but do not report what happens as the labeling budget is deliberately constrained, the question a resource-limited hospital needs answered before committing to a data-collection plan. Our contribution is a continuous, six-point fraction sweep with paired, multi-seed statistical testing rather than a single full-data measurement, which is what makes it possible to say plainly that only the smallest fraction clears conventional significance.

7 Limitations

Six limitations bear directly on how the results in this paper should be used.

1. **Unconfirmed patient-level independence.** The Nickparvar aggregation does not, to our knowledge, publish patient- or study-level identifiers, so we could not verify whether same-patient slices appear in both the training pool and the test set. If they do, every accuracy and macro-F1 figure here is likely inflated relative to a genuinely patient-disjoint evaluation. We consider this the most consequential limitation in the paper.
2. **Limited statistical power beyond the smallest fraction.** Table 6 shows $p > .05$ at 10, 20, 30, 50, and 100% data with only three seeds, and the unusually large 5% effect size ($d_z = 76.0$) has a plausible but unconfirmed explanation rooted in our paired, shared-seed design (Section 5.4). More seeds and an independent-groups replication would resolve both points.
3. **TEI and MVDT are unvalidated descriptive quantities, not established metrics.** TEI's $\log_{10}(N)$ normalization was a convenient choice rather than a derived one (Section 4.4), and MVDT's 0.85 macro-F1 threshold is an illustrative reference point, not a clinically or regulatorily validated criterion.
4. **Single dataset, single backbone.** TEI, MVDT, and the glioma failure mode are demonstrated on one architecture, one dataset, and one balanced four-class task; whether they generalize to other backbones, datasets, or class balances is untested, and no external validation cohort was used.
5. **Explainability is qualitative only.** No segmentation masks were available to compute a quantitative attention-to-lesion overlap metric (e.g., IoU); a dataset such as BraTS [44] would allow this and would let Grad-CAM++ be benchmarked against ground truth.
6. **Glioma remains the weakest class** even with a balanced dataset and focal loss. Its diffuse, infiltrative margins are plausibly harder to distinguish from normal

parenchyma on a single 2-D slice than more circumscribed lesions; further gains may require 3-D volumetric context or glioma-specific augmentation rather than additional 2-D images alone.

7.1 What Would Strengthen This Work

These are extensions rather than corrections: the design here is sound within its stated scope, and none of the limitations above threatens the validity of the comparisons we do report. The most useful next step would simply be more of the same experimental design applied more broadly: additional compute and a second, patient-verified dataset, not a change in method.

8 Conclusion

This paper delivers a fraction-resolved benchmark of ImageNet-pretrained versus from-scratch EfficientNet-B0 for brain tumor MRI classification, built on a completed 36-run sweep with paired statistical testing across six data budgets. The central, statistically supported finding is that transfer learning's advantage is confirmed at the smallest data budget tested (5%, 238 images: 95% CI [11.0, 11.8] pp) and directionally consistent, though not independently confirmed, at every larger budget, a result that reframes how much labeled data a resource-constrained hospital actually needs before pretraining stops paying an outsized dividend. Per-class and Grad-CAM++ analysis adds a second concrete contribution: glioma is identified as the persistent weak point, with a specific failure mode, drift toward "no tumor" predictions under uncertainty, that argues directly for a screen-then-refer deployment rather than an autonomous one.

We report two minor descriptive quantities, TEI and MVDT, as shorthand for patterns already visible in the raw results, and are explicit that neither carries independent validation (Section 4.4). We also flag, more directly than prior work on this dataset family has, that unconfirmed patient-level identifiers in the Nickparvar collection leave open a leakage risk that future work should resolve with a patient-disjoint split (Section 7). Within those stated bounds, this benchmark offers a concrete, paired-statistics answer to a question the literature has largely left open: how the value of transfer learning changes as a labeling budget shrinks, and where the evidence for that value is strong versus merely suggestive.

Declarations

Ethics approval: This study used only publicly available, fully anonymized datasets. No human participants were recruited, no patient-identifiable information was accessed, and no interventions involving humans or animals were performed. Therefore, institutional ethics approval and informed consent were not required.

Author contributions: NUA contributed to the conceptualization of the study, methodology, software development, experimental design, formal analysis, data curation, visualization, validation, and preparation of the original manuscript draft. AR

contributed to conceptualization, research supervision, methodology refinement, critical review of the manuscript, and overall scientific guidance. MSH contributed to software implementation, validation, data curation, experimental evaluation, and manuscript review. MF contributed to clinical interpretation, validation of the medical relevance of the study, manuscript review, and editing. KRC contributed to conceptualization, supervision, interpretation of results, manuscript review and editing, and provided overall academic guidance. AUA supervised this work and contributed to methodology validation, manuscript review, and editing. All authors read and approved the final version of the manuscript.

Funding: The authors received no specific financial support from any public, commercial, or not-for-profit funding agency for this research.

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data availability: The datasets used in this study are publicly available and can be obtained from their respective repositories. The implementation code, trained model weights, and supplementary materials will be made publicly available upon acceptance of the manuscript to support transparency and reproducibility.

Author Biographies

Nakib Uddin Ahmed received the MSc degree in Information Technology from Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh. His research interests include artificial intelligence, deep learning, medical image analysis, explainable AI, computer vision, clinical decision support systems and trustworthy machine learning.

Azizur Rahman is affiliated with Indiana Wesleyan University, Marion, Indiana, USA. His research interests include artificial intelligence, medical image analysis, deep learning, explainable AI, clinical decision support systems, and healthcare informatics.

Md. Shahadat Hossain received the BSc degree in Computer Science and Engineering from Metropolitan University, Sylhet, Bangladesh. His research interests include machine learning, deep learning, medical image analysis, data validation, and intelligent healthcare systems.

Mehjabin Ferdous received the Master of Public Health (MPH) degree from St. Francis College, Brooklyn, New York, USA. Her research interests include public health, evidence-based healthcare, clinical validation, digital health, and AI applications in medicine.

Kadirur Rahman Chowdhury is affiliated with Birmingham City University, Birmingham, United Kingdom. His research interests include biomedical engineering, artificial intelligence, medical imaging, clinical decision support systems, and intelligent healthcare technologies.

Akhlak Uz Zaman Ashik is an Assistant Professor in the Department of Computer Science and Engineering at Metropolitan University, Sylhet, Bangladesh. His research interests include artificial intelligence, machine learning, deep learning, and medical image analysis.

References

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. https://doi.org/10.3322/caac.21660

[2] 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

[3] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

[4] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 32.

[5] 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), Article e0140381. https://doi.org/10.1371/journal.pone.0140381

[6] Sultan, H. H., Salem, N. M., & Al-Atabany, W. (2019). Multi-classification of brain tumor images using deep neural network. *IEEE Access*, 7, 69215–69225. https://doi.org/10.1109/ACCESS.2019.2919122

[7] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

[9] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9). IEEE. https://doi.org/10.1109/CVPR.2015.7298594

[10] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708). IEEE. https://doi.

org/10.1109/CVPR.2017.243

[11] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR.

[12] Iqbal, A., Jaffar, M. A., & Jahangir, R. (2024). Enhancing brain tumour multi-classification using Efficient-Net B0-based intelligent diagnosis for Internet of Medical Things (IoMT) applications. *Information*, 15(8), Article 489. https://doi.org/10.3390/info15080489

[13] Nickparvar, M. (2021). Brain tumor MRI dataset [Data set]. *Kaggle*. https://doi.org/10.34740/KAGGLE/DSV/2645886

[14] Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S., & Kanchan, S. (2020). Brain tumor classification (MRI) [Data set]. *Kaggle*. https://doi.org/10.34740/KAGGLE/DSV/1183165

[15] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3), Article 63. https://doi.org/10.1145/3386252

[16] Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109, 373–440. https://doi.org/10.1007/s10994-019-05855-6

[17] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, Article 60. https://doi.org/10.1186/s40537-019-0197-0

[18] Maicas, G., Bradley, A. P., Nascimento, J. C., Reid, I., & Carneiro, G. (2018). Training medical image analysis systems like radiologists. In A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Vol. 11070, pp. 546–554). Springer. https://doi.org/10.1007/978-3-030-00928-1_62

[19] Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S. J., Wille, M. M. W., Naqibullah, M., Sánchez, C. I., & van Ginneken, B. (2016). Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5), 1160–1169. https://doi.org/10.1109/TMI.2016.2536809

[20] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

[21] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2), 442–451.

https://doi.org/10.1016/0005-2795(75)90109-9

[22] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision* (pp. 839–847). IEEE. https://doi.org/10.1109/WACV.2018.00097

[23] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

[24] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 3320–3328.

[25] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[26] Cheplygina, V., de Bruijne, M., & Pluim, J. P. W. (2019). Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54, 280–296. https://doi.org/10.1016/j.media.2019.03.009

[27] Rumala, D. J. (2023). How you split matters: Data leakage and subject characteristics studies in longitudinal brain MRI analysis. In *Machine Learning in Clinical Neuroimaging* (Vol. 14312, pp. 246–257). Springer.

[28] 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, Article 22544. https://doi.org/10.1038/s41598-021-01681-w

[29] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.

[30] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 448–456). PMLR.

[31] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

[32] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*.

[33] Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

[34] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2980–2988). IEEE. https://doi.org/10.1109/ICCV.2017.324

[35] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826). IEEE. https://doi.org/10.1109/CVPR.2016.308

[36] Mongan, J., Moy, L., & Kahn, C. E., Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2), Article e200029. https://doi.org/10.1148/ryai.2020200029

[37] U.S. Food and Drug Administration, Health Canada, & United Kingdom Medicines and Healthcare Products Regulatory Agency. (2021). *Good machine learning practice for medical device development: Guiding principles*. U.S. Food and Drug Administration.

[38] Colas, C., Sigaud, O., & Oudeyer, P.-Y. (2018). How many random seeds? Statistical power analysis in deep reinforcement learning experiments. *arXiv*. https://doi.org/10.48550/arXiv.1806.08295

[39] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 3207–3214. https://doi.org/10.1609/aaai.v32i1.11694

[40] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.

[41] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

[42] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning*

Representations (ICLR 2021).

[43] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11976–11986). IEEE. https://doi.org/10.1109/CVPR52688.2022.01167

[44] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., ... Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694