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.

Compact CNN--Transformer Refinement for Patient-Disjoint Brain Tumor MRI Classification with an Optimized CNN Comparator

Dian Nova Kusuma Hardani

diannova.kh@ump.ac.id

Universitas Muhammadiyah Purwokerto

Maulida Ayu Fitriani

Universitas Muhammadiyah Purwokerto

Latiful Hayat

Universitas Muhammadiyah Purwokerto

Feri Wibowo

Universitas Muhammadiyah Purwokerto

Research Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11512443/v1

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

Additional Declarations: No competing interests reported.

Compact CNN–Transformer Refinement for Patient-Disjoint Brain Tumor MRI Classification with an Optimized CNN Comparator

Dian Nova Kusuma Hardani ORCID: 0000-0001-5201-4912[1*], Maulida Ayu Fitriani ORCID: 0009-0008-0444-2793[2], Latiful Hayat ORCID: 0009-0001-7746-9718[1], Feri Wibowo ORCID: 0000-0001-5225-158X[2]

[1*]Department of Electrical Engineering, Universitas Muhammadiyah Purwokerto, Jl. KH. Ahmad Dahlan PO BOX 202, Purwokerto, 53182, Indonesia.
[2]Department of Informatics Engineering, Universitas Muhammadiyah Purwokerto, Jl. KH. Ahmad Dahlan PO BOX 202, Purwokerto, 53182, Indonesia.

*Corresponding author(s). E-mail(s): diannova.kh@ump.ac.id;
Contributing authors: maulidaayuf@ump.ac.id; latifulhayat@ump.ac.id; feriwibowo@ump.ac.id;

Abstract

Hybrid CNN–Transformer classifiers can appear advantageous when compared only with fixed or untuned CNN baselines, making it difficult to separate the effect of global-context refinement from hyperparameter optimization and imbalance handling. This study evaluates a compact EfficientNet-B0–Transformer against a matched-budget optimized CNN-only comparator using 3,064 contrast-enhanced T1-weighted MRI slices from 233 patients. All model development used patient-disjoint training, validation, and held-out test partitions. The CNN-only and hybrid pipelines used focal loss and adaptive minority-class augmentation and were independently optimized with 40 Optuna tree-structured Parzen estimator trials; the test set remained inaccessible during selection. On the held-out set, optimized CNN and hybrid models achieved Macro F1 scores of 0.9096 and 0.9341, respectively, with patient Macro F1 of 0.8956 and 0.9327. Across three independent training seeds, hybrid image-level Macro F1 was $\mathbf{0.9249 \pm 0.0187}$ versus $\mathbf{0.9018 \pm 0.0090}$ for the optimized CNN. However, patient-wise five-fold robustness results were close ($\mathbf{0.9355 \pm 0.0383}$ versus $\mathbf{0.9342 \pm 0.0455}$), and the paired patient-level McNemar test was not significant ($\boldsymbol{p = 0.5}$). The hybrid required only 9.5% more parameters, with modest latency and memory overhead. The results support compact self-attention as an incremental refinement, while showing that its benefit should be judged against equally optimized CNN baselines and multiple robustness checks.

Keywords: brain tumor classification, magnetic resonance imaging, CNN, Transformer, class imbalance, patient-disjoint evaluation

1 Introduction

Multiclass brain tumor recognition from magnetic resonance imaging (MRI) is an image-classification problem in which both representation quality and evaluation design affect the observed result. Convolutional neural networks (CNNs) provide strong local and hierarchical image features, while attention mechanisms can model interactions between spatially separated features. Attention-based brain tumor classifiers have consequently been reported in *Signal, Image and Video Processing* and related image-processing literature [1, 2]. Recent work has expanded this direction through global-Transformer refinement, attention-guided multi-scale CNNs, EfficientNet-family attention models, and Swin-based architectures [3–6]. More recent hybrid and unified designs combine convolutional feature extraction with Transformer-based global modeling or reconstruction-classification coupling [7, 8]. Nevertheless, an architectural gain is difficult to establish when a hybrid model is tuned more extensively than its CNN comparator.

A second difficulty arises from the structure of commonly used slice-based MRI datasets. The Cheng/Figshare dataset contains 3,064 contrast-enhanced T1-weighted slices from 233 patients [9, 10]. Multiple correlated slices may originate from the same patient; random image-wise partitioning can therefore expose related anatomy to both development and test sets and inflate performance estimates [11]. Recent leakage-focused benchmarking has independently emphasized that duplicated or correlated MRI samples can yield optimistic estimates if dataset construction and partitioning are not controlled [12]. The dataset is also imbalanced at image level, with 708 meningioma, 1,426 glioma, and 930 pituitary slices. Contemporary reviews identify class imbalance, repeated use of small public datasets, incomplete per-class reporting, and limited external validation as persistent issues in MRI-based brain tumor classification [13, 14]. Patient-level separation and class-aware learning are therefore important experimental controls rather than optional reporting choices.

EfficientNet provides an efficient convolutional backbone through compound scaling [15], whereas Transformer self-attention provides a mechanism for global feature interaction [16, 17]. EfficientNet-based attention models and Transformer-enhanced systems have recently reported strong brain-MRI classification performance [3, 5, 7]. Hybrid CNN–Transformer models have also been used directly for brain tumor MRI classification [8, 18, 19], while a recent direct CNN-versus-Swin comparison highlights both the potential of Transformer models and the importance of external and protocol-aware validation [20]. However, the incremental value of the Transformer can still be confounded by optimization, augmentation, and class-imbalance treatment. Augmentation effects are also modality dependent, as shown in recent SIVP medical-image studies [21].

This work therefore reframes the comparison around an independently optimized CNN comparator. The contributions are: (1) patient-disjoint development and held-out evaluation; (2) matched-budget Bayesian optimization of an imbalance-aware CNN-only comparator and its corresponding compact CNN–Transformer counterpart; (3) three-seed final-training analysis without best-seed selection; (4) patient-level bootstrap confidence intervals and exact paired testing; (5) patient-wise five-fold robustness analysis with frozen hyperparameters; and (6) computational profiling of parameter count, model size, inference latency, throughput, and peak GPU memory. The specific experimental configurations and model identifiers are defined in Section 2.3 and summarized in Table 2. The aim is to determine whether compact global-context refinement provides incremental value after comparable CNN optimization, rather than to claim universal superiority of attention-based architectures.

2 Materials and methods

2.1 Dataset, patient-disjoint partitioning, and preprocessing

The public Cheng/Figshare dataset [9, 10] comprises 3,064 contrast-enhanced T1-weighted MRI slices from 233 patients with meningioma, glioma, or pituitary tumors. Table 1 summarizes the image-level class distribution. Using random seed 42, patients were assigned to mutually exclusive training, validation, and held-out test subsets in nominal proportions of 70%, 10%, and 20%, corresponding to 163, 23, and 47 patients. Every slice from a patient remained in a single subset,

and the held-out test subset remained inaccessible during fitting, hyperparameter optimization, early-stopping selection, and model selection.

Table 1 Image counts and class proportions in the Cheng/Figshare brain tumor dataset [9, 10].

Class	Images	Proportion
Meningioma	708	23.11%
Glioma	1,426	46.54%
Pituitary	930	30.35%
Total	3,064	100%

Images were converted to floating point, min–max normalized to $[0, 1]$, resized to 224×224, replicated to three channels, and standardized with ImageNet channel statistics. Training augmentation was applied only after patient-level partitioning. The common transform used rotation up to 5°, translation 0.02 along each axis, and scale 0.98–1.02. Meningioma, the least represented class at image level, received class-specific adaptive augmentation when enabled; its rotation, translation, and horizontal-flip probability were included in Bayesian optimization. Validation and test images were not augmented.

2.2 CNN and CNN–Transformer models

EfficientNet-B0 pretrained on ImageNet was used as the convolutional feature extractor with its original classifier removed [15]. For the hybrid model, the resulting $7 \times 7 \times 1280$ feature map was projected by a 1×1 convolution to embedding dimension d, flattened into 49 spatial tokens, and combined with a learnable positional embedding. TransformerEncoderLayer blocks used multi-head self-attention, a $4d$ feed-forward layer, GELU activation, dropout, and pre-layer normalization. Scaled dot-product attention was

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right) V \quad (1)$$

Mean token pooling, layer normalization, and a `Linear(`d`,128)`–GELU–dropout–`Linear(128,3)` classifier produced the logits. Softmax was used only during inference/evaluation. Fig. 1 shows the implemented hybrid architecture.

2.3 Controlled configurations and matched-budget optimization

The experimental design comprised six fixed configurations and two matched-budget optimized variants, as summarized in Table 2. The fixed configurations were constructed to separate the contributions of focal loss, adaptive minority-class augmentation, and Transformer-based global-context refinement. The EfficientNet-B0 cross-entropy baseline, focal-loss CNN, imbalance-aware CNN, hybrid cross-entropy model, hybrid focal-loss model, and complete hybrid configuration are denoted E1 through E6, respectively. This controlled design enables individual training and architectural components to be examined before optimization is introduced.

Among the CNN-only configurations, E3 represents the complete imbalance-aware CNN branch because it combines EfficientNet-B0 with focal loss and adaptive minority-class augmentation. E6 represents the corresponding complete CNN–Transformer branch with the same imbalance-handling components and an additional Transformer refinement stage. To avoid attributing performance differences to unequal optimization effort, E3 and E6 were subsequently optimized independently using the same Bayesian-optimization budget. The resulting variants are denoted E3-Opt and E6-Opt, respectively. Thus, the E3-Opt versus E6-Opt comparison evaluates the incremental contribution of global-context refinement after both model families have received comparable hyperparameter optimization.

$$\text{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \quad (2)$$

where p_t denotes the predicted probability assigned to the correct class and γ controls the degree of down-weighting applied to well-classified samples [22]. Under the common fixed settings, all models were trained using AdamW with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, batch size 32, and cosine-annealing learning-rate scheduling. Training was limited to 100 epochs with a minimum of 30 epochs, early-stopping patience of 20 epochs, and minimum improvement threshold of 0.001. For focal-loss configurations, $\gamma = 2.0$ was used.

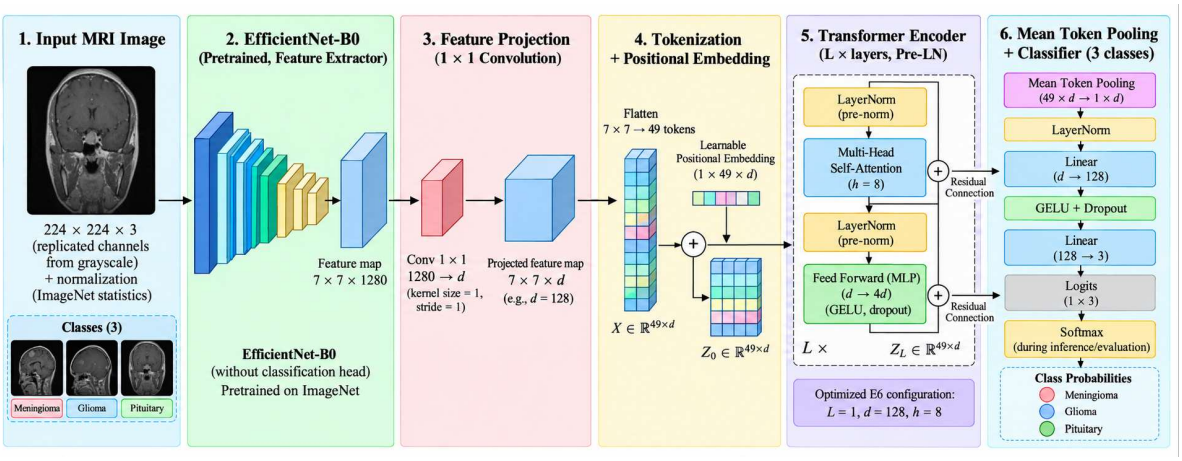


Fig. 1 EfficientNet-B0–Transformer architecture used for the optimized hybrid model. The selected configuration used one pre-layer-normalized Transformer encoder layer, embedding dimension $d = 128$, and eight attention heads.

Table 2 Experimental configurations E1–E6 and matched-budget optimized variants.

ID	Backbone	Transformer	Loss	Adaptive aug.	Optimization / role
E1	EfficientNet-B0	No	Cross-entropy	No	Fixed CNN baseline
E2	EfficientNet-B0	No	Focal	No	Focal-loss CNN
E3	EfficientNet-B0	No	Focal	Yes	Imbalance-aware CNN
E4	EfficientNet-B0	Yes	Cross-entropy	No	Hybrid baseline
E5	EfficientNet-B0	Yes	Focal	No	Hybrid + focal loss
E6	EfficientNet-B0	Yes	Focal	Yes	Full hybrid configuration
E3-Opt	EfficientNet-B0	No	Focal	Yes	40-trial optimized CNN comparator
E6-Opt	EfficientNet-B0	Yes	Focal	Yes	40-trial optimized hybrid model

Matched-budget optimization used the Optuna TPE sampler [23] with 40 trials per model and validation Macro F1 as the objective. The shared search space comprised learning rate $(5\times10^{-5}$–$8\times10^{-4})$, weight decay $(10^{-6}$–$5\times10^{-4})$, batch size $\{16, 32, 64\}$, focal-loss γ (0.5–3.0), minority rotation (5°–12°), translation (0.01–0.04), and horizontal-flip probability (0–0.30). E6-Opt additionally searched $d \in \{128, 256, 384\}$, attention heads $\{4, 8\}$, one to three Transformer layers, and dropout 0.05–0.30. Each trial used at most 40 epochs, a minimum of 15 epochs, patience 12, and a MedianPruner with a 10-epoch warm-up. The held-out test set remained inaccessible throughout optimization and model selection.

Post-HPO sensitivity analysis compared patience values of 10, 20, 30, and no early stopping under a 100-epoch horizon; patience 10 was retained for both models. E3-Opt selected learning rate 7.6145×10^{-4}, weight decay 1.1515×10^{-6}, batch size 16, $\gamma = 1.7160$, rotation 10.87°, translation 0.0389, and horizontal-flip probability 0.0455. E6-Opt selected learning rate 3.3553×10^{-4}, weight decay 1.8092×10^{-6}, batch size 16, $\gamma = 1.6671$, $d = 128$, eight heads, one Transformer layer, dropout 0.2550, rotation 10.27°, translation 0.0176, and horizontal-flip probability 0.0011.

All experiments were implemented using PyTorch 2.11.0+cu128 and Optuna 5.0.0 and executed on an NVIDIA Tesla T4 GPU with 14.56 GB of memory. Automatic mixed precision was enabled during training. Randomness was controlled using deterministic seeds for Python, NumPy, PyTorch, CUDA, and data-loader workers, with deterministic cuDNN behavior enabled and benchmarking disabled.

2.4 Evaluation, stability, statistical analysis, and efficiency

Image-level evaluation used accuracy, macro precision, macro recall, Macro F1, weighted F1, macro ROC-AUC, and macro PR-AUC. Multiclass AUC metrics used one-vs-rest computation

with macro averaging. Patient-level probabilities were obtained by averaging slice probabilities,

$$\bar{p}_c^{(p)} = \frac{1}{N_p} \sum_{i=1}^{N_p} p_{ic} \tag{3}$$

followed by argmax classification.

After hyperparameters were fixed, E3-Opt and E6-Opt were each retrained with seeds 42, 123, and 2026. All three runs were reported; no seed was selected according to test performance. For the seed-42 held-out comparison, 10,000 patient-level bootstrap resamples were used to estimate 95% percentile confidence intervals and paired differences, and an exact two-sided McNemar test compared patient predictions.

A secondary patient-wise stratified five-fold analysis used all 233 patients. `StratifiedGroupKFold` preserved patient identity, while 12.5% of the fold-specific training-validation pool was reserved for validation. E3-Opt and E6-Opt hyperparameters were frozen before this analysis and were not re-optimized inside folds; the five-fold analysis therefore measures robustness rather than nested cross-validation. Computational profiling used the same Tesla T4 for parameter count, model size, batch-1 inference latency, batch-32 throughput, and peak GPU memory. Multiply–accumulate counts were not reported because a validated Transformer-aware MAC estimate was not available from the profiling backend used in the rerun.

For qualitative model inspection, Grad-CAM was applied post hoc to the last convolutional block of EfficientNet-B0 immediately before the 1×1 projection layer [24]. Explainability remains an active concern in recent Transformer-based brain-tumor imaging literature, particularly because visually plausible saliency does not by itself establish localization validity [13]. For target class c, channel weights were obtained by global averaging of class-specific gradients and combined with the corresponding activation maps followed by a ReLU operation. The resulting map was normalized to $[0, 1]$, upsampled to the input resolution using cubic interpolation, and overlaid on the MRI. Because the source feature map was only 7×7, Grad-CAM was treated strictly as coarse class-discriminative saliency and not as lesion segmentation or quantitative localization evidence.

3 Results

3.1 Matched optimization and held-out comparison

The E3-Opt HPO selected trial 33 with a validation Macro F1 of 0.9630. In the subsequent 100-epoch early-stopping sensitivity run, patience values of 10, 20, 30, and no early stopping all recovered the same best validation Macro F1 of 0.95275 at epoch 30. These values are not expected to coincide because the HPO trials used a 40-epoch scheduling horizon, whereas the post-HPO sensitivity runs used a 100-epoch horizon. Patience 10 was retained because it stopped at epoch 40 with the least additional computation. The corresponding E6-Opt sensitivity analysis selected the same patience and recovered its best validation Macro F1 of 0.96181 at epoch 33.

Table 3 shows the key held-out comparison. E3 was the strongest fixed CNN configuration, increasing Macro F1 from 0.9138 for E1 to 0.9207. Importantly, Bayesian optimization did not improve E3 held-out performance: E3-Opt obtained Macro F1 0.9096 despite a higher validation objective. In contrast, E6-Opt achieved 0.9398 accuracy and 0.9341 Macro F1. Relative to the matched-budget E3-Opt comparator, E6-Opt improved held-out accuracy by 2.17 percentage points, image Macro F1 by 2.45 points, patient accuracy by 4.26 points, and patient Macro F1 by 3.70 points. This indicates that validation-driven optimization alone does not explain the E6 result, while also demonstrating that optimization can over-specialize a CNN configuration to a small validation cohort.

At class level, E6-Opt exceeded E3-Opt F1 for all three classes: meningioma 0.9207 versus 0.8938, glioma 0.9611 versus 0.9477, and pituitary 0.9205 versus 0.8874. The largest held-out patient-level gain was nevertheless uncertain because only 47 patients were available. E3-Opt patient accuracy was 0.8936 (95% bootstrap CI 0.8085–0.9787) and E6-Opt achieved 0.9362 (0.8511–1.0000). The paired accuracy difference was +0.0426 (95% CI 0–0.1064), and the patient Macro F1 difference was +0.0370 (95% CI 0–0.0976). Two patient predictions were discordant, both favoring E6-Opt; the exact McNemar test was not significant ($p = 0.5$).

Table 3 Held-out comparison of fixed and optimized configurations (seed 42).

Model	Acc.	Macro P	Macro R	Macro F1	ROC-AUC	PR-AUC	Patient acc.	Patient Macro F1
E1 fixed	0.9231	0.9124	0.9153	0.9138	0.9897	0.9758	0.8936	0.8964
E3 fixed	0.9281	0.9207	0.9207	0.9207	0.9889	0.9713	0.8936	0.8964
E3-Opt	0.9181	0.9168	0.9045	0.9096	0.9878	0.9721	0.8936	0.8956
E6-Opt	**0.9398**	**0.9389**	**0.9318**	**0.9341**	0.9822	0.9733	**0.9362**	**0.9327**

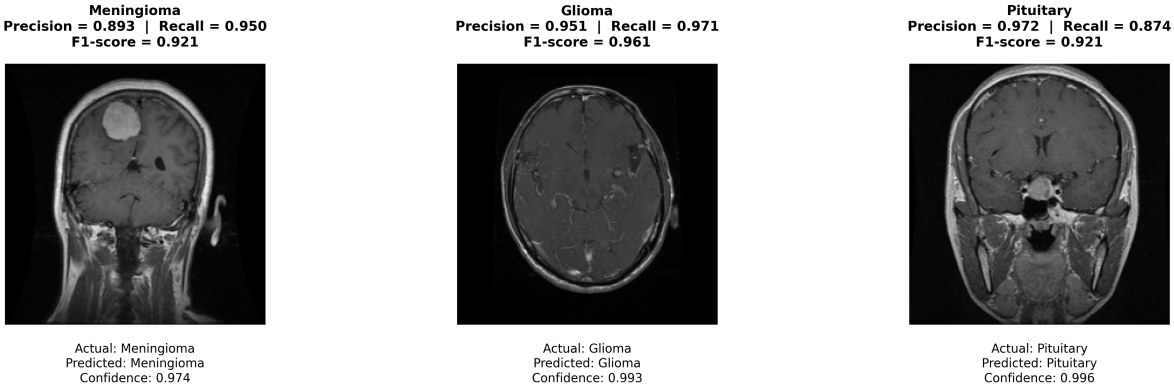


Fig. 2 Representative correctly classified held-out MRI examples from E6-Opt for meningioma, glioma, and pituitary tumor. Class-level precision, recall, and F1-score are reported above each category, while the displayed confidence values are uncalibrated maximum softmax scores provided only for qualitative illustration. MRI images are derived from the Cheng/Figshare brain tumor dataset [9, 10], licensed under CC BY 4.0.

To complement the quantitative held-out evaluation, Fig. 2 presents one representative correctly classified MRI slice from the E6-Opt model for each tumor class. The examples illustrate variation in tumor appearance and imaging orientation within the held-out cohort. These examples are intended only as qualitative demonstrations of model predictions. The displayed confidence values correspond to uncalibrated maximum softmax scores and should not be interpreted as calibrated probabilities or measures of diagnostic certainty.

3.2 Training-repeat stability and computational cost

Table 4 shows that, across three independent final-training seeds, E6-Opt maintained higher average performance than E3-Opt. Image Macro F1 was 0.9249 ± 0.0187 for E6-Opt and 0.9018 ± 0.0090 for E3-Opt; patient Macro F1 was 0.9149 ± 0.0203 and 0.8725 ± 0.0225, respectively. These repeats characterized stability rather than selecting a favorable seed; all runs were retained.

The hybrid refinement added 381,824 parameters, increasing the model from 4.01 to 4.39 million parameters (+9.5%) and model size from 15.46 to 16.92 MB. Batch-1 latency increased from 8.22 to 9.28 ms/image (+12.9%), while batch-32 throughput decreased from 758.6 to 731.0 images/s (3.6%). Peak GPU memory changed only from 479.3 to 480.8 MB (+0.3%). Thus, the Transformer produced a measurable but modest computational overhead relative to the optimized CNN.

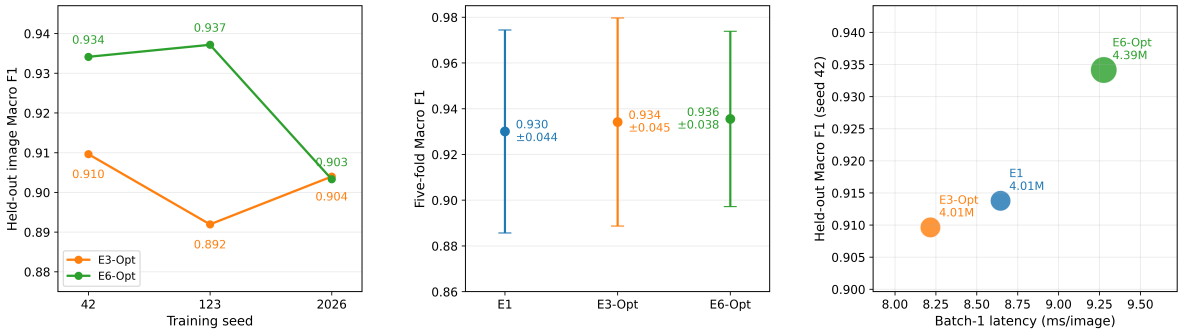
3.3 Patient-wise five-fold robustness

The patient-wise five-fold analysis provides a complementary assessment of robustness across different patient compositions. Table 5 summarizes the classification-oriented metrics. E3-Opt and E6-Opt achieved nearly identical mean accuracy and Macro F1, with E6-Opt showing marginally higher mean values and slightly lower variability. Both optimized models also exceeded the fixed E1 baseline in mean accuracy and Macro F1.

Table 6 provides the corresponding ranking-based metrics. In contrast to the small advantage of E6-Opt in accuracy and Macro F1, E3-Opt achieved the higher mean ROC-AUC and PR-AUC. This difference is not contradictory because the optimization objective was validation Macro

Table 4 Three-seed stability and computational profile of the optimized CNN and hybrid models.

Model	Image acc.	Image Macro F1	Patient acc.	Patient Macro F1	Params (M)	Latency (ms)	Throughput (img/s)
E3-Opt	0.9119 ± 0.0079	0.9018 ± 0.0090	0.8723 ± 0.0213	0.8725 ± 0.0225	4.01	8.22	758.6
E6-Opt	$\mathbf{0.9320 \pm 0.0165}$	$\mathbf{0.9249 \pm 0.0187}$	$\mathbf{0.9149 \pm 0.0213}$	$\mathbf{0.9149 \pm 0.0203}$	4.39	9.28	731.0



(a) Three-seed repeatability (b) Patient-wise five-fold robustness (c) Performance-efficiency trade-off

Fig. 3 Performance, robustness, and efficiency synthesis. (a) Held-out Macro F1 across three training seeds. (b) Patient-wise five-fold Macro F1 (mean ± SD). (c) Seed-42 held-out Macro F1 versus batch-1 latency; labels indicate parameter count.

Table 5 Patient-wise stratified five-fold classification performance (mean ± SD).

Model	Accuracy	Macro F1
E1 fixed	0.9372 ± 0.0385	0.9300 ± 0.0444
E3-Opt	0.9421 ± 0.0381	0.9342 ± 0.0455
E6-Opt	$\mathbf{0.9425 \pm 0.0342}$	$\mathbf{0.9355 \pm 0.0383}$

Table 6 Patient-wise stratified five-fold ranking performance (mean ± SD).

Model	ROC-AUC	PR-AUC
E1 fixed	0.9821 ± 0.0216	0.9628 ± 0.0366
E3-Opt	$\mathbf{0.9883 \pm 0.0134}$	$\mathbf{0.9746 \pm 0.0263}$
E6-Opt	0.9836 ± 0.0219	0.9601 ± 0.0470

F1 rather than an AUC-based criterion, and the models may differ in score ranking even when their final discrete classifications are similar.

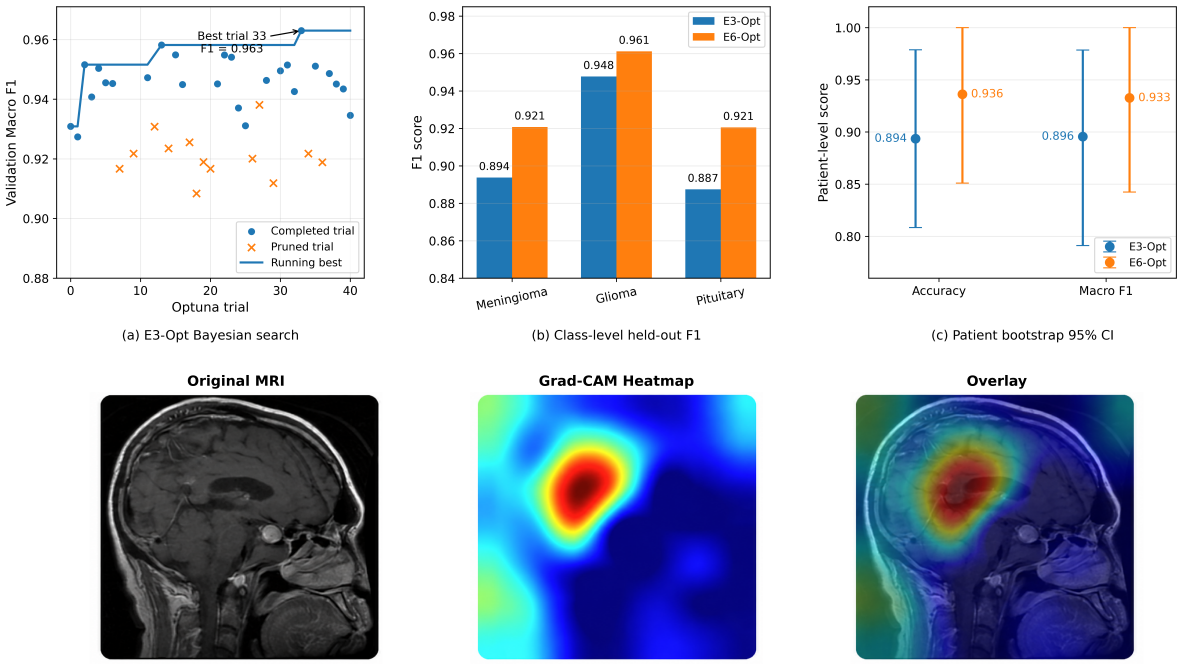
Overall, the five-fold results indicate that the held-out advantage of E6-Opt was not uniformly reproduced across every patient composition. The optimized CNN and hybrid pipelines occupied closely overlapping robustness ranges, although their relative behavior differed between classification-oriented and ranking-based metrics.

Fig. 3 places these findings together with the three-seed analysis and computational profile. The repeated-seed panel shows that E6-Opt maintained a higher average held-out Macro F1, whereas the five-fold panel demonstrates that the two optimized pipelines became much closer when evaluated across different patient partitions. The efficiency panel further shows that the seed-42 gain of E6-Opt was obtained with a moderate increase in parameter count and inference latency.

3.4 Optimization behavior, class-level response, and qualitative inspection

Fig. 4 complements the aggregate comparisons with four additional views. First, the E3-Opt search trajectory shows that the best completed trial occurred at trial 33 with validation Macro F1 0.9630, while several other completed trials clustered close to the optimum and pruned trials occupied a broader lower range. This pattern indicates that the search did not identify a single isolated parameter combination; instead, several configurations achieved similar validation performance, which increases the importance of evaluating the selected model outside the optimization objective.

Second, the class-wise comparison shows that E6-Opt improved F1 for all three tumor classes relative to E3-Opt. The absolute gains were



(a) E3-Opt Bayesian search

(b) Class-level held-out F1

(c) Patient bootstrap 95% CI

(d) Representative Grad-CAM inspection for a correctly classified pituitary MRI

Fig. 4 Complementary analyses. (a) E3-Opt Bayesian-search outcomes and running-best curve; the selected optimum was trial 33. (b) Class-level held-out F1. (c) Patient-level accuracy and Macro F1 with 95% bootstrap confidence intervals from 10,000 resamples. (d) Representative Grad-CAM inspection for a correctly classified pituitary MRI; the saliency is qualitative and is not a segmentation or localization result.

0.0270 for meningioma, 0.0134 for glioma, and 0.0331 for pituitary. The largest improvement therefore occurred for pituitary, whereas meningioma also benefited despite being the least represented image-level class. Third, the patient-level bootstrap intervals remain wide for both models because the held-out cohort contains only 47 patients. The E6-Opt point estimates are higher for both patient accuracy and patient Macro F1, but the overlapping intervals visually reinforce the non-significant paired McNemar result.

Finally, the representative Grad-CAM example for a correctly classified pituitary MRI shows a broad activation pattern that overlaps the lesion-related sellar/suprasellar region and surrounding structures. The saliency map is intentionally presented only as a qualitative inspection of model focus. Its coarse spatial resolution and the absence of tumor masks preclude any claim of segmentation accuracy, localization precision, or clinical interpretability.

4 Discussion

4.1 Architectural contribution and model-selection behavior

The main contribution of this study is not the introduction of another CNN–Transformer combination, but a controlled assessment of whether compact global-context refinement adds value after the CNN comparator has received comparable optimization effort. This distinction is important because architectural effects can otherwise be confounded by unequal hyperparameter tuning, class-imbalance handling, or augmentation strategies. Under the primary patient-disjoint evaluation, E6-Opt achieved better held-out performance than the matched-budget E3-Opt comparator, and the same overall direction was preserved across repeated final-training runs. These findings suggest that a shallow self-attention stage can provide useful refinement of EfficientNet features after focal loss and adaptive augmentation have already been applied.

The matched optimization experiment also highlights an important model-selection issue. Although E3-Opt achieved a strong validation objective during Bayesian optimization, this did not translate into better held-out performance than the fixed E3 configuration. This should not be interpreted as evidence that hyperparameter optimization is inherently detrimental. Rather, it illustrates how optimization against a relatively small validation cohort may favor a configuration that is well adapted to the validation partition without producing a corresponding gain on unseen patients. Keeping the held-out test set inaccessible throughout optimization was therefore essential for distinguishing model-selection performance from generalization.

The repeated-seed and five-fold analyses further qualify the architectural interpretation. E6-Opt maintained the higher average performance across repeated training runs, but its advantage was not uniform across all seeds. Likewise, the difference between E3-Opt and E6-Opt became small relative to fold-to-fold variability in the patient-wise five-fold analysis. Taken together, these results support the Transformer block as an incremental global-context refinement rather than evidence of universal architectural superiority. They also demonstrate why hybrid-model claims should be supported by matched optimization, repeated training, and patient-level robustness analysis rather than by a single favorable split.

4.2 Relation to recent brain-MRI classification literature

Recent brain-tumor MRI studies have reported strong performance using global Transformer refinement, attention-guided CNNs, EfficientNet-based attention, Swin Transformer, and hybrid CNN–Transformer designs [3–7]. However, direct ranking by reported accuracy is inappropriate because these studies differ in datasets, partitioning rules, optimization procedures, and evaluation units.

Recent leakage-aware work further emphasizes the importance of evaluation design. MTA-Swin incorporated leakage-free benchmarking, repeated seeds, cross-validation, and Grad-CAM analysis [12], while a recent CNN-versus-Swin study used external testing and highlighted the need for patient-level and prospective validation [20]. The present study addresses a complementary question: whether a compact attention stage adds value after the CNN comparator receives matched optimization effort. This emphasis is consistent with recent reviews identifying public-dataset dependence, class imbalance, explainability, and incomplete validation as persistent limitations [13, 14].

4.3 Class-specific behavior, uncertainty, and practical trade-offs

The held-out improvement of E6-Opt was distributed across all three tumor classes rather than being driven by a single category. Because both optimized pipelines already used focal loss and adaptive minority-class augmentation, the additional gain of E6-Opt is more appropriately interpreted as complementary global-context refinement than as an isolated effect of self-attention.

The optimized Transformer configuration remained compact, adding only 9.5% more parameters and a moderate latency increase while leaving peak GPU-memory demand nearly unchanged. E3-Opt therefore remains attractive where simplicity and throughput are prioritized, whereas E6-Opt provides a stronger held-out operating point when modest additional computation is acceptable.

Patient-level uncertainty nevertheless limits the strength of any superiority claim. The exact McNemar test was not significant and the paired bootstrap intervals included zero. Together with the small five-fold difference relative to fold-to-fold variability, these findings support a cautious interpretation: compact Transformer refinement can improve performance, but the magnitude of benefit remains dependent on patient composition and training realization.

4.4 Validity boundaries and future work

Several limitations remain. The study uses a single public contrast-enhanced T1-weighted dataset, so generalization across scanners, institutions, acquisition protocols, and populations cannot be established. This limitation is especially important

because recent work has shown that external testing can materially strengthen claims of generalization, while current reviews continue to identify dataset concentration and incomplete cross-site validation as major barriers [14, 20]. The model operates on two-dimensional slices, and patient probability averaging does not replace a dedicated volumetric or multimodal patient model. The held-out cohort contains only 47 patients, limiting power for paired statistical inference. Probability calibration was not evaluated, and the five-fold analysis does not quantify uncertainty from a fully nested model-selection procedure. Future work should prioritize external multi-center validation, probability calibration, and compact three-dimensional or multimodal extensions while preserving patient-level grouping, matched optimization of comparators, repeated-training analysis, and explicit efficiency reporting.

5 Conclusion

This study evaluated compact CNN–Transformer refinement against an independently optimized imbalance-aware CNN under patient-disjoint brain tumor MRI classification. E6-Opt outperformed E3-Opt on the held-out split and across three training seeds, while adding 9.5% parameters and modest inference overhead. However, the patient-level paired comparison was not statistically significant and five-fold Macro F1 was nearly identical between optimized CNN and hybrid models. The evidence therefore supports Transformer attention as an incremental global-context refinement rather than a universally superior architecture. More broadly, the results show that claims of hybrid-model benefit should be based on equally optimized comparators, patient-level leakage control, repeated training, robustness analysis, and explicit computational trade-offs.

Acknowledgements

The authors gratefully acknowledge Universitas Muhammadiyah Purwokerto and the Research and Community Service Board for institutional support during this study.

Declarations

- Funding
  This research was supported by Universitas Muhammadiyah Purwokerto through the Research and Community Service Board under contract number A.11-III/8699-S.Pj./LPPM/II/2026.
- Competing interests
  The authors have no relevant financial or non-financial interests to disclose.
- Ethics approval and consent to participate
  Ethics approval was not required because the study used a publicly available de-identified MRI dataset and involved no new recruitment, intervention, or collection of identifiable human-participant data. Consent to participate and consent for publication are not applicable.
- Consent for publication
  Not applicable.
- Data availability
  The Brain Tumor MRI Dataset is publicly available through Figshare at https://doi.org/10.6084/m9.figshare.1512427.v8.
- Materials availability
  Not applicable.
- Code availability
  Research code and trained model checkpoints are available from the corresponding author on reasonable request, subject to institutional intellectual-property procedures.
- Author contribution
  D.N.K.H. has conceptualized the study, prepared the methodology, developed the software, conducted the formal analysis, created the visualizations, provided supervision, and written the original draft, as well as reviewed and edited the manuscript. M.A.F., L.H., and F.W. have conducted investigations, validated the findings, and reviewed and edited the manuscript. All authors have read and approved the final manuscript.

References

[1] Shaik, N.S., Cherukuri, T.K.: Multi-level attention network: application to brain tumor classification. Signal, Image and Video Processing **16**(3), 817–824 (2022) https://doi.org/10.1007/s11760-021-02022-0

[2] Han, Q., Ren, C.: An attention-based hybrid

deep learning technique for brain tumor classification. Signal, Image and Video Processing **20**(7) (2026) https://doi.org/10.1007/s11760-026-05455-7

[3] Dutta, T.K., Nayak, D.R., Pachori, R.B.: GT-Net: global transformer network for multiclass brain tumor classification using MR images. Biomedical Engineering Letters **14**(5), 1069–1077 (2024) https://doi.org/10.1007/s13534-024-00393-0

[4] Dutta, T.K., Nayak, D.R., Zhang, Y.-D.: ARM-Net: Attention-guided residual multi-scale CNN for multiclass brain tumor classification using MR images. Biomedical Signal Processing and Control **87**, 105421 (2024) https://doi.org/10.1016/j.bspc.2023.105421

[5] Pacal, I., Celik, O., Bayram, B., *et al.*: Enhancing EfficientNetv2 with global and efficient channel attention mechanisms for accurate MRI-based brain tumor classification. Cluster Computing **27**(8), 11187–11212 (2024) https://doi.org/10.1007/s10586-024-04532-1

[6] Pacal, I.: A novel swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in mri images. International Journal of Machine Learning and Cybernetics **15**(9), 3579–3597 (2024) https://doi.org/10.1007/s13042-024-02110-w

[7] Huang, Z., Duan, J., Xie, Y., Liu, Y.: UDNet: Unified deep network based on transformer and multi-stage fusion for brain tumor classification from undersampled MRI. Neurocomputing **619**, 129109 (2025) https://doi.org/10.1016/j.neucom.2024.129109

[8] Abdellatef, E., Al-Makhlasawy, R.M., El-Mawla, N.A., Shalaby, W.A.: A novel hybrid transformer-based framework (h-convnext–swin) to classify brain tumors using mri. Scientific Reports **16**(1) (2026) https://doi.org/10.1038/s41598-026-63915-z

[9] Cheng, J., Huang, W., Cao, S., *et al.*: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS ONE **10**(10), 0140381 (2015) https://doi.org/10.1371/journal.pone.0140381

[10] Cheng, J.: Brain Tumor Dataset. figshare (2017). https://doi.org/10.6084/m9.figshare.1512427.v8

[11] Yagis, E., Atnafu, S.W., Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., Diciotti, S.: Effect of data leakage in brain mri classification using 2d convolutional neural networks. Scientific Reports **11**(1) (2021) https://doi.org/10.1038/s41598-021-01681-w

[12] Lu, D., Zhang, Y., Chaudhary, D.: MTA-Swin: A multi-token attention swin transformer for brain tumor classification with leakage-free MRI benchmarking. Journal of Medical Systems **50**(1) (2026) https://doi.org/10.1007/s10916-026-02418-w

[13] Hosny, K.M., Mohammed, M.A.: Explainable ai and vision transformers for detection and classification of brain tumor: a comprehensive survey. Artificial Intelligence Review **58**(9) (2025) https://doi.org/10.1007/s10462-025-11221-x

[14] Noor, N., Turner, C., Holdsworth, S.J., *et al.*: Advanced deep learning architectures in mri-based brain tumor classification: A systematic review focused on meningiomas. Journal of Imaging Informatics in Medicine (2026) https://doi.org/10.1007/s10278-026-02069-w

[15] Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv (2019). https://doi.org/10.48550/ARXIV.1905.11946

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2017). https://doi.org/10.48550/ARXIV.1706.03762

[17] Dosovitskiy, A., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

[18] Aloraini, M., Khan, A., Aladhadh, S., *et al.*: Combining the transformer and convolution for effective brain tumor classification using mri images. Applied Sciences **13**(6), 3680 (2023) https://doi.org/10.3390/app13063680

[19] Goceri, E.: An efficient network with CNN and transformer blocks for glioma grading and brain tumor classification from MRIs. Expert Systems with Applications, 126290 (2025) https://doi.org/10.1016/j.eswa.2024.126290

[20] Zakavi, S.A., Radnia, P., Abdollahi Karizno, M., Tafazolimoghadam, A., Amiri, H., Deravi, N., Fathi, M.: Comparative evaluation of convolutional neural networks and swin transformer for multiclass brain tumor mri classification. Discover Applied Sciences **8**(8) (2026) https://doi.org/10.1007/s42452-026-08767-y

[21] Rainio, O., Klén, R.: Comparison of simple augmentation transformations for a convolutional neural network classifying medical images. Signal, Image and Video Processing **18**(4), 3353–3360 (2024) https://doi.org/10.1007/s11760-024-02998-5

[22] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (2020) https://doi.org/10.1109/tpami.2018.2858826

[23] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: KDD, pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701

[24] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (2019) https://doi.org/10.1007/s11263-019-01228-7

Supplementary Files

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

- ClassificationExamplestwosamples.png
- BayesianOptimizationWorkflow.png