TinyHybrid: Ultra-Efficient CNN-Transformer Architecture for Edge-Deployable Brain Tumor Classification

Krystian Krukowski[1,†], Pawel Powroznik[1,*,†] and Maria Skublewska-Paszkowska[1]

[1]Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland

Abstract

Brain tumor classification from magnetic resonance imaging (MRI) is critical for clinical diagnosis, yet state-of-the-art deep learning models require substantial computational resources. We present TinyHybrid, an ultra-efficient CNN-Transformer hybrid architecture that achieves a mean 3-fold CV accuracy of 95.66% with only 89K parameters and 0.047 GFLOPs - a 976× reduction in parameters compared to Swin Transformer. Our design combines depthwise separable convolutions for local feature extraction with a lightweight transformer encoder for global context modeling.

Critically, TinyHybrid achieves 2.5ms CPU inference (24× faster than Swin Transformer's 59.4ms) and requires only 0.37MB storage, enabling real-time processing on edge devices without GPU acceleration. We also present a full-scale Hybrid model achieving a mean 3-fold CV accuracy of 96.73% with 2.25M parameters and 13.3ms CPU inference. Comprehensive ablation studies validate each architectural component, revealing that the transformer encoder is critical (∼11% accuracy impact), while attention mechanisms can be removed without accuracy loss. Aligned with the principles of Green AI, our models enable deployment in resource-constrained clinical settings, significantly reducing the carbon footprint of medical AI inference.

Keywords

Brain tumor classification, Swin Transformer, XAI, CNN

1. Introduction

Brain tumor classification from MRI scans plays a crucial role in clinical diagnosis and treatment planning. Deep learning approaches have achieved remarkable success in this domain, with Vision Transformers [1] and hybrid architectures [2] reaching near-perfect accuracy. However, state-of-the-art models often require substantial computational power and significant storage, presenting significant barriers for deployment in clinical settings with limited infrastructure.

This computational burden has three critical implications: (1) the deployment requires expensive GPU hardware unavailable in many healthcare facilities, (2) edge deployment on mobile or embedded devices is infeasible, and (3) the environmental impact contradicts the principles of Green AI [3]. Recent work has highlighted the substantial carbon footprint of training and deploying large neural networks [4], motivating the need for efficient alternatives that maintain competitive accuracy.

Our contributions are as follows: i) **TinyHybrid**: An ultra-efficient architecture with 89K parameters achieving 95.66% mean CV accuracy - 976× fewer parameters than Swin Transformer. ii) **Hybrid**: A balanced model with 2.25M parameters achieving 96.73% mean CV accuracy. iii) **Real-time CPU inference**: 2.5ms per image on CPU (24× faster than Swin's 59.4ms) - enabling edge deployment without GPU. iv) **Edge-deployable size**: 0.37MB model file (895× smaller than Swin's 331MB) - fits on microcontrollers. v) **Comprehensive ablation studies** validating channel attention [5], spatial attention [6], transformer encoders, and attention pooling. vi) **Multi-level interpretability analysis** using GradCAM [7] and attention visualization.

SuRE'26: Workshop on Sustainability and Resource-Efficiency of Artificial Intelligence, August 17, 2026, Bremen, Germany

*Corresponding author.

†These authors contributed equally.

s97649@pollub.edu.pl (K. Krukowski); p.powroznik@pollub.pl (P. Powroznik); maria.paszkowska@pollub.pl (M. Skublewska-Paszkowska)

https://cvml.cs.pollub.pl/pawel-powroznik-phd/ (P. Powroznik);
https://cvml.cs.pollub.pl/maria-skublewska-paszkowska-phd-eng/ (M. Skublewska-Paszkowska)

0000-0002-5705-4785 (P. Powroznik); 0000-0002-0760-7126 (M. Skublewska-Paszkowska)

2. Related Work

Deep Learning for Medical Imaging. Convolutional neural networks dominate medical image analysis [8, 9]. ResNet [10] and DenseNet [11] achieve over 95% accuracy in brain tumor classification. Vision Transformers [1, 12], including the Swin Transformer [13], further improve accuracy through self-attention but require substantial computational resources (over 86M parameters and 15 GFLOPs per inference).

Efficient Architectures. MobileNets [14, 15] introduced depthwise separable convolutions for efficient computation, while EfficientNet [16] proposed compound scaling. MobileViT [2] combined convolutions with transformers, and RegNet [17] and ConvNeXt [18] further improved network efficiency. However, developing sub-100K parameter models with competitive accuracy remains challenging. Our method addresses this gap, achieving competitive performance with 55$\times$ fewer parameters than MobileViT while enabling real-time CPU inference.

Green AI. Schwartz et al. [3] introduced the Green AI framework, advocating for efficiency as a primary evaluation criterion alongside accuracy. Strubell et al. [4] quantified the environmental cost of large models, estimating that training a single transformer can emit as much CO_2 as five cars over their lifetimes. Patterson et al. [19] further emphasized the importance of model efficiency for reducing the carbon footprint of AI systems. This growing awareness has motivated research into efficient model architectures that minimize computational requirements while maintaining accuracy. Our TinyHybrid aligns with these principles, enabling sustainable medical AI deployment in resource-constrained clinical environments worldwide.

Brain Tumor Classification. Brain tumor classification has mainly relied on transfer learning from ImageNet-pretrained models. Abiwinanda et al. [20] applied CNNs, while Deepak et al. [21] improved performance using GoogleNet and AlexNet. In contrast, our hybrid architecture is tailored to medical imaging, combining local feature extraction with global context modeling while maintaining extreme efficiency.

3. Materials and Mathods

In this section, we describe the design of our hybrid CNN-Transformer architecture. We use the Brain Tumor MRI Dataset [22] comprising 7,023 images across four classes (Glioma, Meningioma, No Tumor, Pituitary). The dataset represents a realistic clinical distribution with slight class imbalance: Glioma and Meningioma are the most common tumor types, while Pituitary tumors are less prevalent. All images are T1-weighted MRI scans acquired from multiple clinical sources, providing diversity in imaging conditions and scanner characteristics. The integrated dataset aggregates images from three primary sources: Figshare, Sartaj, and Br35H. The Figshare subset consists of T1-weighted contrast-enhanced (CE-MRI) scans acquired at Nanfang Hospital (Guangzhou) and General Hospital (Tianjin Medical University) in China between 2005 and 2010. These images feature an in-plane resolution of 512×512 with pixel dimensions of 0.49×0.49 mm^2, a slice thickness of 6 mm, and were obtained following Gd-DTPA injection. For the Sartaj and Br35H subsets, specific acquisition protocols and detailed patient demographics were not disclosed by the original authors, though they contribute to the overall heterogeneity of the combined benchmark. The detailed statistics for training and testing splits are provided in Table 1.

3.1. Architecture Overview

Our hybrid architecture combines CNN local feature extraction with transformer global context modeling through four components: (1) CNN Tokenizer, (2) Attention Mechanisms, (3) Transformer Encoder, and (4) Classification Head. The TinyHybrid architecture is illustrated in Figure 1, and the detailed specifications are in Table 2.

CNN Tokenizer The tokenizer uses depthwise separable convolutions [14] organized into residual blocks following the ResNet design [10], replacing standard convolutions to maximize efficiency:

Table 1
Dataset Statistics

Class	Training	Testing
Total	5,712	1,311
Glioma	1,321	300
Mening.	1,339	306
None	1,595	405
Pitui.	1,457	300

Table 2
Architecture specifications

Component	Hybrid	TinyHybrid
CNN Channels	8→16→32	8→16→24
Feature Map	28×28	14×14
Tokens	784	196
Hidden Dimension	256	64
Transformer Layers	4	2
Attention Heads	4	2
MLP Dimension	512	128
Dropout	0.2	0.2
Total Parameters	**2.25M**	**89K**
GFLOPs	**0.88**	**0.047**

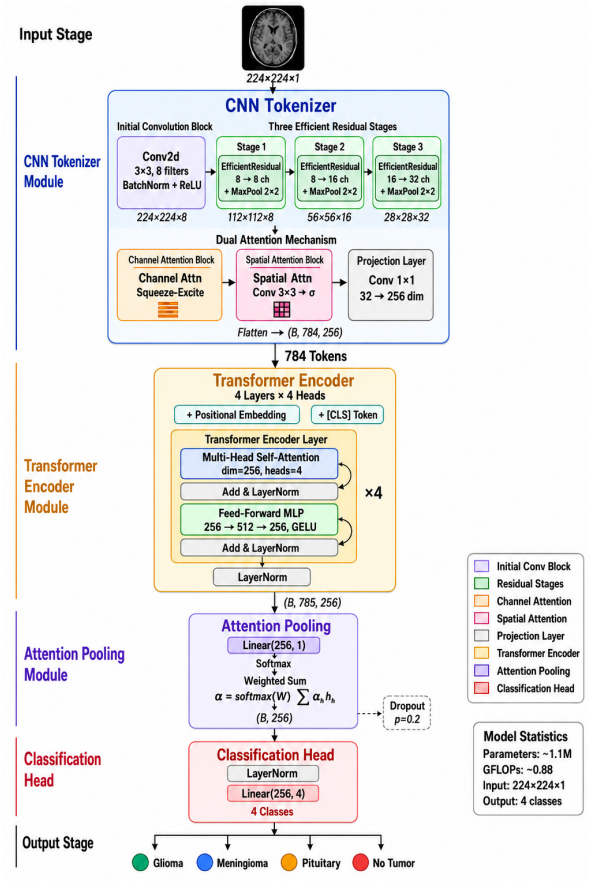


Figure 1: TinyHybrid architecture

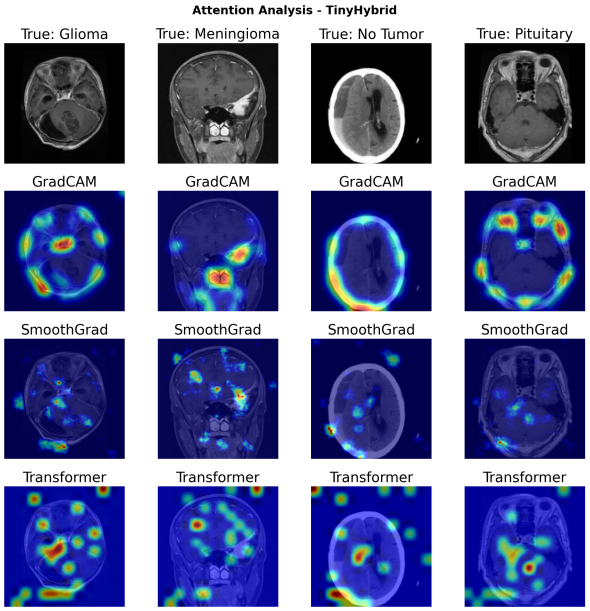


Figure 2: Attention Analysis - TinyHybrid: Four visualization rows showing Original images, CNN feature activations, SmoothGrad saliency, and Transformer attention for each tumor class

RB(X) = ReLU(DSC$_2$(DSC$_1$$(X)$) + X) where DSC$_1$ applies a full depthwise separable convolution (depthwise 3×3 followed by pointwise 1×1), and DSC$_2$ applies a second DSC without an activation between the two convolution pairs, before the residual addition. This reduces computational cost from $O(K^2 \cdot C_{in} \cdot C_{out})$ to $O(K^2 \cdot C_{in} + C_{in} \cdot C_{out})$.

Attention Mechanisms Channel Attention (SE-style) [5] adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels: CA(X) = $X \cdot \sigma$(FC(GAP(X))).

Spatial Attention (CBAM-style) [6] complements this by applying a spatial attention map to emphasize informative regions while suppressing background noise: SA$(X) = X \cdot \sigma(\text{Conv}_{3\times3}(X))$.

We employ **Transformer Encoder**, a lightweight transformer with learnable positional embeddings and a CLS token for classification, following the ViT formulation [1]: Attention(Q, K, V) = softmax $\left(\frac{QK^T}{\sqrt{d_k}}\right) V$, where Q, K, V represent the query, key, and value matrices projected from the input, and d_k is the dimension of the key vectors.

Table 3
Performance and deployment efficiency comparison. Accuracy is reported as mean ± std across 3-fold CV. Params are given in millions, latency in ms.

Model	Mean Acc ± Std	AUC	Params [M]	GFLOPs	GPU	CPU	Size [MB]	Img/s
TinyHybrid	94.34%±0.69	0.992	**0.089**	**0.047**	1.02	2.5	0.37	400
Tiny: No Attn	95.66%±0.10	0.995	**0.089**	**0.047**	0.90	2.5	0.37	400
Hybrid	96.73%±0.53	0.993	2.25	0.88	1.50	13.3	8.65	75
+Optimized	96.44%±0.81	0.996	2.25	0.88	–	–	–	–
DenseNet-121	97.10%±0.38	0.997	6.96	2.90	5.07	17.5	27.12	57
MobileNetV3	97.22%±0.28	0.998	4.21	0.22	2.13	5.5	16.25	182
ConvNeXt	97.43%±0.13	0.999	87.57	15.35	3.53	51.2	334.18	19
RegNetY-032	95.12%±0.53	0.994	17.93	3.18	4.39	16.9	68.81	59
ResNet-50	98.23%±0.20	0.999	23.52	4.13	1.96	14.9	90.01	67
EfficientNetV2	98.37%±0.08	0.998	20.18	2.71	5.73	20.7	77.84	48
MobileViT	98.60%±0.33	0.998	4.94	1.42	3.03	11.4	19.01	88
ViT	98.48%±0.34	0.999	85.80	16.85	2.83	49.5	327.36	20
Swin Trans.	98.95%±0.25	0.999	86.75	15.17	7.06	59.4	331.04	17

4. Experiments and Results

We conducted all experiments using the following settings: i) **Preprocessing**: 224×224 grayscale images with augmentation including flipping, rotation ±15°, affine transformations, and Gaussian blur; ii) **Optimization**: AdamW with cosine annealing; iii) **Custom models**: 150 epochs, LR=3×10^{-4}, patience=15. The Hybrid and TinyHybrid models were trained from scratch without ImageNet pretraining, requiring more epochs to learn medical-image representations; iv) **Pretrained baselines**: 20 epochs, LR=1×10^{-4}, patience=5. Since ImageNet-pretrained models already contain robust feature extractors learned from 1.2M natural images, only the classification head and minor backbone components were fine-tuned. These models typically converged within 10–15 epochs, while early stopping limited overfitting; v) **Regularization**: label smoothing=0.1 and weight decay=1×10^{-4}; vi) **Validation**: 3-fold cross-validation, with Table 3 reporting mean accuracy ± standard deviation across folds; vii) **Hardware**: AMD Ryzen 7 7800X3D CPU, NVIDIA RTX 4080 SUPER GPU, and 32GB DDR5 6000MHz RAM under Pop!_OS 22.04 LTS.

We compared the proposed models with Swin Transformer [13], ViT [1], MobileViT [2], MobileNetV3 [23], EfficientNetV2 [16], ResNet-50 [10], DenseNet-121 [11], ConvNeXt [18], and RegNetY-032 [17], focusing on the accuracy–efficiency trade-off. As shown in Table 3, TinyHybrid achieved 94.34%±0.69 mean accuracy and 0.992 AUC with only 89K parameters, corresponding to a 976× parameter reduction compared with Swin Transformer.

The 3-fold cross-validation results showed (Table 6) that removing all attention mechanisms in the **Tiny: No All Attn** variant increased mean accuracy to 95.66%±0.10 (+1.32%) while preserving the same 89K-parameter footprint. This result approaches MobileNetV3 accuracy (97.22%) while using a much smaller number of parameters (0.089M vs. 4.21M) and FLOPs (0.047 vs. 0.22), indicating that strong compression can still provide competitive performance. The full **Hybrid** model achieved 96.73%±0.53, remaining competitive with larger models such as MobileViT (98.60%) and MobileNetV3 (97.22%) while using fewer parameters, whereas **Hybrid: Optimized** reached 96.44%±0.81, slightly below Hybrid due to higher fold variance.

Edge Deployment Analysis A critical contribution of this work is the feasibility of deploying these models on highly constrained edge devices. Table 3 highlights this capability. TinyHybrid's model file size is just 0.37MB, which fits comfortably within the on-chip flash memory of low-cost microcontrollers like the ESP32 (typically 4MB). This eliminates the need for external SDRAM or network dependency. Furthermore, the inference speed is transformative. At 2.5ms per image on a Ryzen 7-7800X3D CPU, TinyHybrid enables a theoretical throughput of 400 FPS, far exceeding real-time requirements (typically 30 FPS). In contrast, Swin Transformer requires nearly 60ms per image, making it unsuitable for real-time applications on CPU-only hardware. These characteristics enable deployment

Table 4
Ablation Results showing the impact of removing specific components (Tiny stands for TinyHybrid).

Ablation	Hybrid	Δ	Tiny	Δ
Full Model	97.10%	—	94.58%	—
- Channel Attn	95.12%	-1.98%	95.88%	+1.30%
- Spatial Attn	95.58%	-1.52%	93.59%	-0.99%
- Transformer	86.27%	-10.83%	83.30%	-11.28%
- Attn Pooling	96.80%	-0.30%	94.13%	-0.45%
- Dropout	96.64%	-0.46%	93.90%	-0.68%
- All Attention	97.10%	+0.00%	96.49%	+1.91%
Optimized	97.48%	+0.38%	95.65%	+1.07%

Table 5
Per-class metrics. G – Glioma, M – Meningioma, N – No Tumor, P – Pituitary.

Model	Class	Sens.	Spec.	Prec.	F1
Tiny: No All Attn	G	0.950	0.990	0.966	0.958
	M	0.912	0.986	0.952	0.932
	N	0.993	0.980	0.957	0.975
	P	0.997	0.996	0.987	0.992
Hybrid: Optimized	G	0.950	0.999	0.997	0.973
	M	0.951	0.984	0.948	0.949
	N	1.000	0.985	0.967	0.983
	P	0.990	0.998	0.993	0.992
Swin Transformer	G	0.987	0.999	0.997	0.992
	M	0.990	0.995	0.984	0.987
	N	1.000	0.997	0.993	0.996
	P	0.993	1.000	1.000	0.997

in portable diagnostic devices, where low latency and offline operation are essential. TinyHybrid's small footprint and fast CPU inference make it well suited to these applications. Its computational efficiency also reduces energy consumption, extending battery life, lowering operational costs, and supporting reliable deployment in resource-constrained healthcare environments.

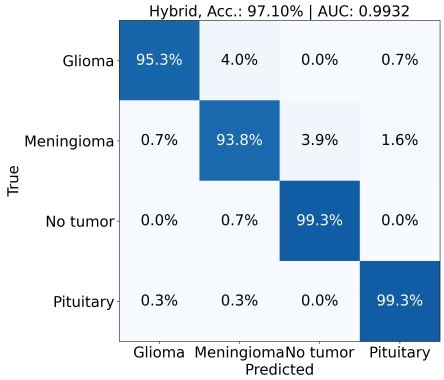
Ablation Studies (Table 4) show that the **Transformer Encoder** is the key component, with its removal reducing accuracy by ∼11% in both Hybrid and TinyHybrid models, highlighting the importance of global context modeling. Attention mechanisms exhibit scale-dependent behavior: they improve the larger Hybrid model but may slightly reduce TinyHybrid performance. This suggests that, at only 89K parameters, allocating capacity to the backbone is more beneficial than attention modules.

The ablation study identified two optimal configurations. **Hybrid: Optimized**, without spatial attention and dropout, achieved the highest accuracy (97.48%), while **TinyHybrid: No All Attention** reached 96.49%, outperforming the full TinyHybrid by 1.91%. These results suggest that highly compressed models benefit from simpler architectures. For deployment, Hybrid: Optimized is preferable when maximizing accuracy, whereas TinyHybrid without attention provides the best accuracy–efficiency trade-off. Table 4 summarizes all evaluated configurations.

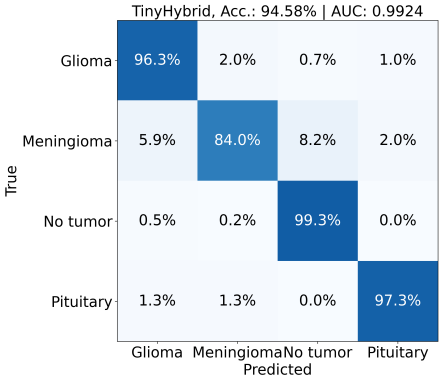
Confusion Matrices The confusion matrices (Figures 3a and 3b) exhibit strong diagonal dominance, indicating high classification accuracy across all classes. The 'No Tumor' and 'Pituitary' categories achieve near-perfect precision, minimizing the risk of misdiagnosis. Despite using 25× fewer parameters, TinyHybrid produces error patterns nearly identical to the Hybrid model, demonstrating that extreme compression preserves discriminative capability. For both models, most errors occur between Glioma and Meningioma due to their similar MRI characteristics, reflecting data complexity rather than model limitations. Both architectures maintain over 95% precision for the 'No Tumor' class and excellent discrimination of the 'Pituitary' class.

Per-Class Metrics and Confidence Intervals (Table 5 and 6) present detailed per-class performance metrics including sensitivity (recall), specificity, precision, and F1-score for key models. All models achieve high specificity (>98%) across all classes, indicating reliable negative predictions. The 'No Tumor' and 'Pituitary' classes achieve near-perfect metrics, while Glioma-Meningioma discrimination remains the primary challenge. Table 6 reports 3-fold cross-validation results with 95% confidence intervals (CI). Our models show low variance, with Tiny: No All Attn achieving 95.66%±0.10 (CI: [95.38, 95.94]), indicating highly consistent performance. Swin Transformer achieves the highest mean accuracy at 98.95%±0.25 but requires 976× more parameters.

To evaluate **Cross-Dataset Generalization**, we tested the models on the independent Bangladesh Brain Cancer MRI Dataset [24], comprising 6,056 MRI images from three tumor classes. As shown in Table 7, ImageNet-pretrained models retained high accuracy (94–98%), whereas our models showed a larger performance drop, likely due to training from scratch on a smaller dataset. Nevertheless, **Tiny:**



(a) Hybrid Confusion Matrix showing strong diagonal dominance.



(b) TinyHybrid Confusion Matrix showing similar performance patterns.

Figure 3: Comparison of confusion matrices obtained for the Hybrid and TinyHybrid models.

Table 6
3-fold cross-validation results with 95% confidence intervals. All models evaluated under identical training conditions.

Model	Mean Acc.	F1	95% CI
Tiny: No Attn	95.66%±0.10	0.964	[95.38, 95.94]
TinyHybrid	94.34%±0.69	0.944	[92.63, 96.05]
+Optimized	96.44%±0.81	0.974	[94.44, 98.44]
Hybrid	96.73%±0.53	0.970	[95.43, 98.03]
RegNetY-032	95.12%±0.53	0.950	[93.81, 96.43]
DenseNet-121	97.10%±0.38	0.971	[96.16, 98.04]
MobileNetV3	97.22%±0.28	0.971	[96.53, 97.91]
ConvNeXt	97.43%±0.13	0.974	[97.11, 97.75]
ResNet-50	98.23%±0.20	0.983	[97.74, 98.72]
EfficientNetV2	98.37%±0.08	0.984	[98.17, 98.57]
MobileViT	98.60%±0.33	0.987	[97.78, 99.42]
ViT	98.48%±0.34	0.985	[97.64, 99.32]
Swin Transf.	98.95%±0.25	0.993	[98.33, 99.57]

Table 7
Cross-Dataset Generalization on Bangladesh Brain Cancer MRI dataset.

Model	Accuracy	AUC	Δ Acc
Swin Transformer	98.1%	0.999	-1.2%
MobileViT	96.6%	0.997	-2.1%
ResNet-50	94.9%	0.992	-3.5%
ViT	93.7%	0.989	-4.8%
ConvNeXt	90.4%	0.988	-8.7%
RegNetY-032	89.7%	0.983	-8.9%
DenseNet-121	89.6%	0.985	-8.5%
EfficientNetV2	86.9%	0.981	-11.6%
Tiny: No All Attn	86.0%	0.964	-10.5%
Hybrid: Optimized	82.2%	0.949	-15.3%
MobileNetV3	81.6%	0.952	-15.7%
Hybrid	81.3%	0.942	-15.8%
TinyHybrid	80.1%	0.956	-14.5%

No All Attn achieved 86.0% accuracy, outperforming MobileNetV3 (81.6%) with 47× fewer parameters. All models maintained high AUC values (>0.94), indicating robust ranking performance despite reduced accuracy.

To **interpret** TinyHybrid's predictions, we apply complementary visualization techniques (Figure 2): (1) original MRI, (2) CNN feature maps, (3) SmoothGrad saliency maps, and (4) transformer attention maps. CNN features capture local textures, SmoothGrad highlights regions influencing predictions, and transformer attention emphasizes global tumor-related context while suppressing background. These visualizations confirm that TinyHybrid learns clinically meaningful features despite its compact architecture.

5. Discussion and Conclusions

Novelty and Contribution. TinyHybrid's contribution lies in systematically identifying the optimal parameter regime for edge medical imaging. We combine depthwise separable convolutions, SE-style channel attention, CBAM spatial attention, and a lightweight transformer into a cohesive architecture. Our ablation studies reveal that attention mechanisms become counterproductive at extreme compression (89K parameters), providing a clear guideline: when scaling down hybrid architectures, prioritise transformer encoder capacity over attention refinement modules.

Role of Spatial Attention in TinyHybrid. Our ablation shows accuracy improves when spatial attention is removed from both models. We retain it in the default architecture to maintain design symmetry with Hybrid and to provide an explicit, evidence-based ablation data (Table 4). The module adds fewer than 500 parameters. The issue is that at 89K total parameters the optimisation landscape is already highly constrained, so auxiliary gradient paths compete with the backbone.

Architectural Comparison and Compute Trade-offs. TinyHybrid's 55× parameter reduction vs. MobileViT results from grayscale input, DSC residual blocks, and a compact 64-d transformer. Since the depthwise separable tokenizer costs $O(K^2C_{in} + C_{in}C_{out})$ per location and the transformer costs $O(N^2d)$, the two transformer layers in TinyHybrid (N=196, d=64) contribute ≈60% of total GFLOPs, justifying the aggressive 14×14 downsampling before attention. The higher CPU latency of EfficientNetV2 compared with ResNet-50 (20.7ms vs 14.9ms), despite lower GFLOPs (2.71 vs 4.13), confirms that FLOPs are an incomplete proxy for efficiency, as Fused-MBConv blocks may suffer from cache misses and limited CPU parallelism.

Green-AI and Environmental Impact. TinyHybrid supports sustainable medical AI by combining 323× fewer FLOPs than Swin Transformer with 24× faster CPU inference, in line with Green-AI principles [3]. Although training still consumes energy, the compact architecture reduces fine-tuning cost and substantially lowers the inference carbon footprint, which is critical for large-scale clinical deployment. Moreover, its 2.5ms CPU inference enables real-time, offline diagnosis in resource-constrained settings without GPU infrastructure, cloud connectivity, or additional computing costs.

Edge Deployment Feasibility. TinyHybrid's 0.37MB file size enables deployment on devices with severe memory constraints: mobile phones, Raspberry Pi (1GB RAM), and embedded medical devices. The 895× compression versus Swin Transformer (331MB) means our model can coexist with other applications on edge devices. Real-time inference (400 images/second on CPU) enables deployment in mobile clinics and remote diagnostic settings where internet connectivity is unreliable.

Limitations. TinyHybrid trades about 3.6% accuracy for extreme efficiency. Glioma–Meningioma confusion likely reflects radiological similarity, while the model's compact size limits the detail of gradient-based saliency maps compared with larger architectures. Future work will explore domain adaptation, multimodal fusion with clinical metadata, and accuracy improvements through ImageNet pretraining or knowledge distillation from stronger teachers such as Swin Transformer or ResNet-50.

AI systems for brain tumor MRI classification should be evaluated not only by accuracy, but also by clinical reliability, transparency, accessibility, and patient safety. TinyHybrid supports responsible clinical AI by reducing computational and infrastructure barriers, enabling efficient local inference in resource-limited settings such as rural hospitals, mobile diagnostic units, and smaller clinical centers. Grad-CAM visualizations are used to support qualitative inspection of model decisions, but they should be treated only as auxiliary explanations, not as clinical evidence or a substitute for expert radiological assessment.

The proposed model is intended as a decision-support tool rather than an autonomous diagnostic system. Its predictions should be reviewed by qualified clinicians and interpreted together with the full radiological and clinical context. Cross-dataset evaluation also shows that performance may decrease under distribution shifts caused by differences in scanners, acquisition protocols, institutions, or patient populations, which remains an important clinical and ethical limitation.

Ethical Statement

There are no ethical issues.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.

[2] S. Mehta, M. Rastegari, Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, in: International Conference on Learning Representations (ICLR), 2022.

[3] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green ai, Communications of the ACM 63 (2020) 54–63.

[4] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.

[5] J. Hu, L. Shen, G. Sun, et al., Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[6] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

[9] P. Powroznik, Deep learning approaches for retinal diseases recognition : From data augmentation towards vision transformers, Politechnika Lubelska, 2026. doi:10.35784/9788379476671.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[12] P. Powroznik, M. Skublewska-Paszkowska, K. Nowomiejska, B. Gajda-Deryło, M. Brinkmann, M. Concilio, M. D. Toro, R. Rejdak, Residual self-attention vision transformer for detecting acquired vitelliform lesions and age-related macular drusen, Scientific Reports 15 (2025) 17107.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, in: International Conference on Machine Learning (ICML), 2021.

[17] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[19] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training, arXiv preprint arXiv:2104.10350 (2021).

[20] N. Abiwinanda, M. Hanif, S. Ezra, A. Russel, Brain tumor classification using convolutional neural

network, in: World Congress on Medical Physics and Biomedical Engineering, 2019.
[21] S. Deepak, P. Ameer, Brain tumor classification using deep cnn features via transfer learning, Computers in Biology and Medicine 111 (2019).
[22] M. Nickparvar, Brain tumor mri dataset, Kaggle, 2021. URL: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.
[23] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[24] M. M. Rahman, Brain cancer - mri dataset, 2024. doi:10.17632/mk56jw9rns.1.