CNNs vs. Hybrid Transformers for Brain Tumor Classification on the BRISC Dataset

Muhammad Thahiruddin[1*], Asri Wulandari[2]
[1,2] Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Annuqayah
[1] muhammad.thahiruddin@ua.ac.id, [2] ayoewoelandari338@gmail.com
*Corresponding Author

ABSTRACT

Accurate and timely classification of brain tumors from Magnetic Resonance Imaging (MRI) is critical for effective treatment planning. The advent of deep learning has revolutionized medical image analysis; however, the performance of different model architectures is highly dependent on the quality of benchmark datasets and the specifics of the training methodology. This study presents a rigorous comparative analysis of four prominent deep learning architectures (ResNet18, EfficientNet-B0, MobileNetV3-Small, and the hybrid convolutional-transformer model MobileViTV2) for multi-class brain tumor classification. The models were trained and evaluated on the BRISC dataset, a large-scale and balanced collection of 6,000 T1-weighted contrast-enhanced MRI scans comprising glioma, meningioma, pituitary, and no-tumor classes. Employing a 5-fold cross-validation protocol with a full fine-tuning strategy and robust regularization techniques, this study evaluates the models in terms of both classification accuracy and computational efficiency. The results indicate that MobileViTV2, ResNet18, and EfficientNet-B0 achieve statistically comparable state-of-the-art performance, with mean test accuracies of 98.88%, 98.72%, and 98.72%, respectively. MobileNetV3-Small, while being the most parameter-efficient, demonstrated significantly lower accuracy at 96.94%. A key finding reveals a performance-efficiency paradox, in which the largest model, ResNet18, exhibited the fastest inference latency (2.83 ms), challenging the conventional assumption that fewer parameters directly translate into greater speed. This comprehensive analysis underscores the strengths of hybrid architectures and provides critical insights into the practical trade-offs among model complexity, accuracy, and real-world deployability for clinical decision support systems.

Keywords: Brain Tumor Classification, Convolutional-Transformer, Magnetic Resonance Imaging, BRISC dataset

ABSTRAK

Klasifikasi tumor otak yang akurat dan tepat waktu dari hasil Magnetic Resonance Imaging (MRI) sangat penting untuk perencanaan pengobatan yang efektif. Kemunculan deep learning telah merevolusi analisis citra medis, namun kinerja dari berbagai arsitektur model sangat bergantung pada kualitas dataset acuan (benchmark) dan spesifikasi metodologi pelatihannya. Studi ini menyajikan analisis perbandingan yang teliti terhadap empat arsitektur deep learning terkemuka (ResNet18, EfficientNet-B0, MobileNetV3-Small, dan model hibryd convolusional-transformer MobileViTV2) untuk klasifikasi tumor otak multikelas. Model-model tersebut dilatih dan dievaluasi menggunakan dataset BRISC, sebuah koleksi data skala besar dan seimbang yang terdiri dari 6.000 pindaian MRI T1-weighted dengan peningkat kontras, yang mencakup kelas glioma, meningioma, pituitari, dan tanpa tumor. Dengan menggunakan protokol validasi silang 5-lipat (5-fold cross-validation), strategi fine-tuning penuh, dan teknik regularisasi yang andal, studi ini menilai model berdasarkan akurasi klasifikasi dan efisiensi komputasi. Hasil penelitian menunjukkan bahwa MobileViTV2, ResNet18, dan EfficientNet-B0 mencapai kinerja canggih (state-of-the-art) yang sebanding secara statistik, dengan rata-rata akurasi pengujian masing-masing sebesar 98,88%, 98,72%, dan 98,72%. Sementara itu, MobileNetV3-Small, yang merupakan model paling efisien dari segi parameter, menunjukkan akurasi yang jauh lebih rendah, yaitu 96,94%. Sebuah temuan kunci mengungkapkan adanya paradoks kinerja-efisiensi, di mana model terbesar, ResNet18, justru menunjukkan latensi inferensi tercepat (2,83 ms). Hal ini menantang asumsi konvensional bahwa jumlah parameter yang lebih sedikit berbanding lurus dengan kecepatan yang lebih tinggi. Analisis komprehensif ini menggarisbawahi kekuatan arsitektur hibrida dan memberikan wawasan penting mengenai trade-off praktis antara kompleksitas model, akurasi, dan kemampuan penerapan di dunia nyata untuk sistem pendukung keputusan klinis.

Keywords: *Klasifikasi Tumor Otak, Convolutional-Transformer, Magnetic Resonance Imaging, BRISC dataset*

1. INTRODUCTION

Brain tumors represent a significant global health challenge and are a primary cause of cancer-related mortality [1]. For both primary and metastatic brain tumors, an accurate and early diagnosis is paramount, as it directly influences the selection of therapeutic strategies, surgical planning, and ultimately, patient prognosis and

survival rates [2], [3]. In this clinical context, Magnetic Resonance Imaging (MRI) has been established as the gold-standard non-invasive diagnostic modality [2], [4]. Its exceptional soft-tissue contrast provides detailed anatomical visualization of brain structures, enabling the characterization of a tumor's location, size, and extent [4].

Despite its utility, the manual interpretation of MRI scans by radiologists is a complex, time-consuming task susceptible to human error and significant inter-observer variability [2], [5], [6]. The increasing volume of medical imaging data further exacerbates these challenges, creating a critical need for automated, reliable, and efficient diagnostic support systems [7].

The field of medical image analysis has been transformed by the application of deep learning, particularly Convolutional Neural Networks (CNNs) [4], [7], [8]. CNNs have demonstrated a remarkable ability to automatically learn hierarchical feature representations directly from pixel data, obviating the need for manual feature engineering and setting new performance benchmarks in diagnostic tasks [2], [9]. The evolution of CNN architectures has progressed from foundational models like VGG and ResNet to highly optimized variants such as EfficientNet and MobileNet, which focus on improving the trade-off between accuracy and computational efficiency [1], [10], [11].

More recently, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs [12]. By employing a self-attention mechanism, ViTs can model long-range dependencies and capture global context across an entire image, overcoming the inherent limitation of the local receptive fields in standard convolutional operations [13], [14]. This capability is particularly relevant for medical imaging, where diagnostic information may be encoded in the spatial relationships between distant regions. The current state-of-the-art is trending towards hybrid architectures that combine the strengths of both paradigms: leveraging CNNs for robust local feature extraction and ViTs for global contextual understanding [15], [16]. This hybrid approach is exemplified by models like MobileViTV2, which integrates these concepts into an efficient framework [17].

The rapid advancement of these sophisticated deep learning models is fundamentally dependent on the availability of large-scale, high-quality, and meticulously annotated datasets. While several public datasets have facilitated progress, they are not without limitations. For instance, the widely used BraTS benchmark focuses primarily on gliomas and often utilizes pre-processed, standardized data, which may not reflect the variability of real-world clinical imaging and can limit the generalizability of trained models. Other resources, such as the Figshare dataset, have been noted to suffer from class imbalance and a lack of diversity in patient demographics and imaging conditions, which can restrict model robustness [18].

To address these shortcomings, this study utilizes the recently introduced BRISC dataset. This modern benchmark was specifically curated to provide a more robust and clinically relevant platform for developing and evaluating neuro-oncological AI models. Its key advantages include a large scale (6,000 T1-weighted contrast-enhanced MRI scans), a balanced distribution across four clinically significant classes (glioma, meningioma, pituitary, and no tumor), and high-quality annotations validated by certified radiologists. The inclusion of a "no tumor" class and multiple common tumor types makes it an ideal testbed for developing general-purpose classification models [18].

The primary contribution of this paper is a rigorous and systematic empirical comparison of four distinct deep learning architectures for brain tumor classification on the BRISC dataset. The selected models represent a spectrum of design philosophies:

1. ResNet18: A classic, robust CNN architecture.
2. EfficientNet-B0: A state-of-the-art CNN optimized for accuracy and efficiency.
3. MobileNetV3-Small: A lightweight CNN designed for resource-constrained environments.
4. MobileViTV2: A modern hybrid CNN-Transformer architecture.

A crucial aspect of the methodology is the adoption of a full fine-tuning strategy, wherein all layers of the pre-trained models were unfrozen and trained on the target dataset. This approach was chosen to comprehensively evaluate the adaptability of ImageNet-derived features to the specific domain of brain MRI analysis. The scope of this study extends beyond mere classification accuracy to include a holistic evaluation of computational efficiency, encompassing parameters, FLOPs, and real-world inference latency. Furthermore, the statistical significance of performance differences is rigorously assessed to provide robust conclusions about the relative merits of each architecture.

2. MATERIAL AND METHODS

2.1. DATASET DESCRIPTION

This study utilizes the BRISC 2025 dataset, a high-quality collection of expert-annotated Magnetic Resonance Imaging (MRI) images of brain tumors. This dataset was specifically designed to address several limitations found in previously available datasets, such as class imbalance and annotation inconsistencies, making it highly suitable for training reliable classification models [18].

The BRISC 2025 dataset comprises a total of 6,000 T1-weighted contrast-enhanced MRI images, curated from a combination of three sources: Figshare, SARTAJ, and Br35H. The images underwent a meticulous review and re-annotation process by certified radiologists to ensure high quality and diagnostic accuracy. For the

classification task, all images are categorized into four distinct classes: Glioma, Meningioma, Pituitary and No Tumor. Each image in the dataset is presented in one of three anatomical planes: axial, coronal, or sagittal. This diversity in planes is intended to ensure that the developed model is robust and not dependent on the image acquisition orientation [18].

The dataset is carefully partitioned into two separate sets: 5,000 images for training and 1,000 images for testing. This division ensures an objective evaluation of the model's performance. The sample distribution for each class across the different MRI planes is summarized in the table below. This table demonstrates that the class distribution is kept relatively balanced in both the training and testing sets, which is crucial for preventing bias during the model training process [18].

Table 1. Class Distribution Based on MRI Planes in the BRISC 2025 Dataset

Class/Plane	Train			Test		
	Axial	Coronal	Sagittal	Axial	Coronal	Sagittal
Glioma	347	428	372	85	81	88
Meningioma	423	426	480	134	89	83
Pituitary	428	496	533	116	98	86
No Tumor	352	310	405	52	48	40
Total Per Plane	1550	1660	1790	387	316	297
Total		**5000**			**1000**	

Although the BRISC 2025 dataset also provides pixel-wise segmentation masks for each image, this study focuses solely on the image classification task. Therefore, the data used consists of the raw MRI images and their corresponding class labels, without utilizing the segmentation mask information [18].

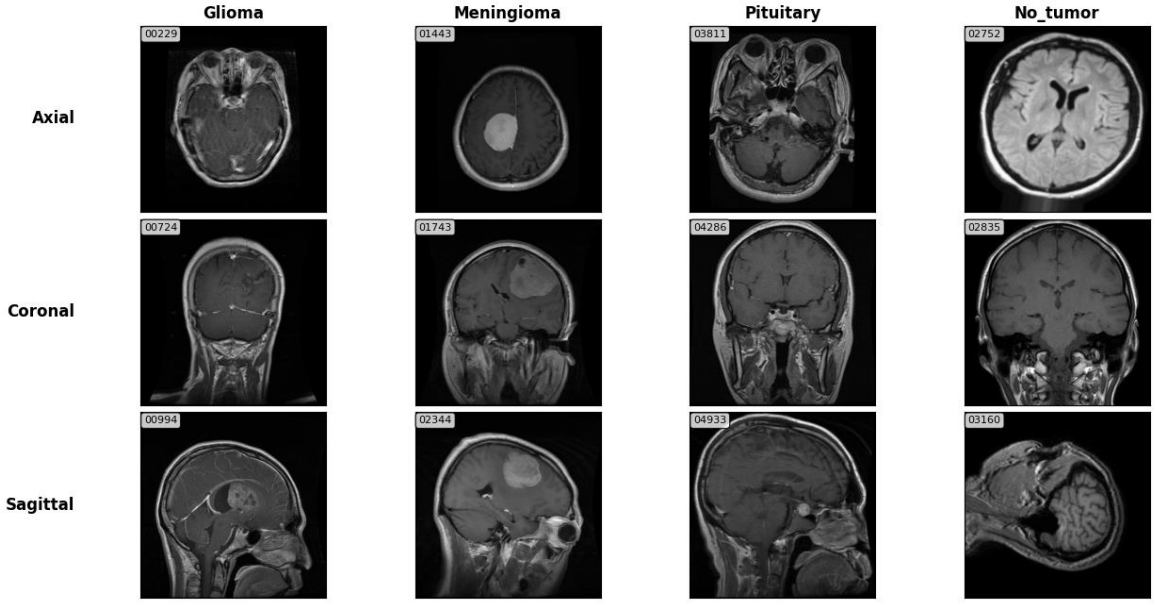

Figure 1. Sample Image for Each Class and Planes

Figure 1 displays representative T1-weighted contrast-enhanced MRI scans for each of the four classes in the BRISC dataset. The images illustrate the distinct morphological characteristics that the models must learn to differentiate. Glioma tumors often present with irregular borders and heterogeneous signal intensity, reflecting their infiltrative nature. Meningiomas typically appear as well-circumscribed, dural-based masses with homogenous contrast enhancement. Pituitary tumors are localized to the sella turcica and present as distinct masses. The "no tumor" class includes scans of healthy brain tissue, which serve as a crucial negative control for training robust classifiers. The visual diversity within and between classes underscores the complexity of the classification task [18].

2.2 MODEL ARCHITECTURES

For this study, four pre-trained deep learning models were selected to represent diverse architectural paradigms. This selection aims to evaluate how different approaches—ranging from conventional CNNs to

efficient hybrid architectures—impact performance on the brain tumor classification task. The details of each model are summarized in the table below.

Table 2. Comparison of the Model Architectures Used

Model Architecture	Core Paradigm	Key Features & Components	Primary Advantage & Use Case
ResNet18 [19]	Deep Residual Network	Utilizes "shortcut connections" to allow gradients to flow more easily to deeper layers.	Fundamental & Robust: Prevents the vanishing gradient problem, enabling the training of very deep networks. It often serves as a strong baseline in various computer vision tasks.
EfficientNet-B0 [20]	Compound Scaling	Introduces a balanced scaling method for network depth, width, and resolution simultaneously.	Optimal Balance: Achieves state-of-the-art accuracy with significantly fewer parameters and computational resources compared to other models.
MobileNetV3-Small [21]	Lightweight Mobile Architecture	Built upon depthwise separable convolutions, inverted residual blocks, and squeeze-and-excite modules.	Efficient & Fast: Specifically designed for resource-constrained devices (e.g., mobile phones). It offers very low computational cost while maintaining competitive performance.
MobileViTV2 [22]	Hybrid CNN-Transformer	Combines convolutions (for local features) with MobileViT blocks that use separable self-attention (for global features).	Local & Global: Capable of efficiently learning both local and global feature relationships within a single lightweight framework, combining the strengths of CNNs and Vision Transformers (ViTs).

2.3. EXPERIMENTAL PROTOCOL

The experiments were conducted in a reproducible and systematic manner to ensure a fair comparison.

a) Framework and Hardware: All models were implemented using the PyTorch deep learning framework [23]. The training and evaluation were performed in a Kaggle environment equipped with two NVIDIA Tesla T4 GPUs, enabling parallel training strategies.
b) Cross-Validation: A 5-fold cross-validation scheme was employed on the 5,000 training images. This process involves splitting the data into five equal folds, training the model five times with each fold serving once as the validation set, and averaging the results. This ensures that the reported performance is robust and less sensitive to the particularities of a single data split.
c) Training Strategy: A full fine-tuning approach was used for all models. The weights of each model, pre-trained on the ImageNet dataset, were unfrozen, allowing all parameters to be updated during training. This strategy tests the model's capacity to adapt its learned features to the specific domain of brain MRI classification.
d) Hyperparameters: The training process for all models was standardized using the hyperparameters specified in Table 3.

Table 3. Training Hyperparameters

Parameter	Value	Parameter	Value
Input Image Size	224^2	Scheduler	Cosine Annealing [24]
Batch Size	128 (64 per GPU)	Learning Rate	5×10^{-4}
Number of Epochs	50 + 5 warmup epoch	Weight Decay	1×10^{-4}
Optimizer	AdamW [25]	K-Folds	5

2.4 DATA AUGMENTATION AND REGULARIZATION

To enhance model generalization and prevent overfitting, several regularization techniques were applied during training.

a) Mixup [26]: This data augmentation technique generates new training samples by linearly interpolating between two existing samples and their corresponding labels. A new sample $(\tilde{x}, \tilde{y})$ is formed from two samples (x_i, y_i) and (x_j, y_j) as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \quad (1)$$

where the mixing coefficient λ is drawn from a Beta distribution, $\lambda \sim Beta(\alpha, \alpha)$, with α set to 0.2.

b) Label Smoothing [27]: This regularization technique discourages the model from making overconfident predictions by replacing hard one-hot labels with smoothed labels. For a given class k and a total of K classes, the new label y'_k is calculated from the original one-hot label y_k using a smoothing parameter ϵ:

$$y'_k = y_k(1 - \epsilon) + \frac{\epsilon}{K} \quad (2)$$

In this experiment, ϵ was set to 0.001 and K was 4.

c) Dropout [28]: A dropout layer with a probability of 0.2 was added to the final classifier head of each model. This technique randomly sets a fraction of neuron activations to zero during training, preventing complex co-adaptations between neurons.

2.5. EVALUATION METRICS AND STATISTICAL ANALYSIS

Model performance was assessed using a combination of classification and efficiency metrics.

a) Classification Metrics: Performance was evaluated using standard metrics: Accuracy (overall correct predictions), Precision (positive predictive value), Recall (sensitivity), and the macro-averaged F1-Score (harmonic mean of precision and recall, averaged across all classes) [29].

b) Efficiency Metrics: Computational cost was measured by: Total Parameters (number of trainable weights), Model Size (storage space in MB), FLOPs (floating-point operations, a measure of theoretical complexity), and Inference Time (latency in ms for a single-batch prediction on the specified hardware).

c) Statistical Analysis: To determine if performance differences were statistically meaningful, a one-way Analysis of Variance (ANOVA) and a non-parametric Friedman test were first applied to the distribution of test accuracies from the 5-fold cross-validation. Subsequently, post-hoc pairwise comparisons were conducted using paired t-tests to identify significant differences between specific model pairs. A p-value less than 0.05 was considered statistically significant.

3. RESULTS

3.1. OVERALL CLASSIFICATION PERFORMANCE

The comparative performance of the four architectures, evaluated using 5-fold cross-validation, is summarized in Table 3. The results indicate that three of the four models achieved exceptionally high and closely matched classification accuracy.

Table 3: Overall Model Performance (5-Fold Cross-Validation)

Model	Test Accuracy	Validation Accuracy	Macro Precision	Macro Recall	Macro F1-Score
MobileViTV2	0.9888 ± 0.0025	0.9944 ± 0.0014	0.9879 ± 0.0031	0.9904 ± 0.0023	0.9891 ± 0.0026
ResNet18	0.9872 ± 0.0010	0.9904 ± 0.0019	0.9883 ± 0.0013	0.9886 ± 0.0010	0.9884 ± 0.0011
EfficientNet-B0	0.9872 ± 0.0038	0.9920 ± 0.0025	0.9875 ± 0.0037	0.9880 ± 0.0035	0.9877 ± 0.0036
MobileNetV3-Small	0.9694 ± 0.0053	0.9742 ± 0.0055	0.9672 ± 0.0072	0.9730 ± 0.0054	0.9698 ± 0.0061

MobileViTV2 emerged as the top-performing model with a mean test accuracy of 98.88%. It was followed almost identically by ResNet18 and EfficientNet-B0, both achieving a mean test accuracy of 98.72%. MobileNetV3-Small, while still performing well, registered a noticeably lower mean test accuracy of 96.94%. The low standard deviations across all metrics for the top three models suggest stable and consistent performance across the different data folds.

3.2. COMPUTATIONAL AND EFFICIENCY ANALYSIS

An analysis of the computational requirements and efficiency of each model reveals significant architectural trade-offs, as detailed in Table 4. MobileNetV3-Small is unequivocally the most lightweight model, with only 1.5M parameters and a 5.94 MB footprint. In contrast, ResNet18 is the largest, with over 7 times more parameters and a model size 7 times greater than MobileNetV3-Small.

Table 4: Model Efficiency and Computational Cost

Model	Total Parameters	Size (MB)	GFLOPs	Inference Time (ms)
MobileNetV3-Small	1,521,956	5.94	0.056	5.52
EfficientNet-B0	4,012,672	15.59	0.385	8.67
MobileViTV2	4,390,893	16.91	1.412	9.26
ResNet18	11,178,564	42.72	1.824	2.83

However, a counter-intuitive relationship between theoretical complexity (FLOPs) and practical speed (Inference Time) was observed. ResNet18, despite having the highest FLOPs, recorded the fastest inference time at just 2.83 ms. Conversely, MobileViTV2, with fewer FLOPs than ResNet18, was the slowest at 9.26 ms. This suggests that factors such as architectural design and hardware-level optimization play a more significant role in determining real-world latency than theoretical operational counts alone.

3.3. ERROR AND MISCLASSIFICATION ANALYSIS

An analysis of the aggregated confusion matrices (Figure 2) from the cross-validation revealed specific patterns of misclassification for each model, highlighting their respective strengths and weaknesses.

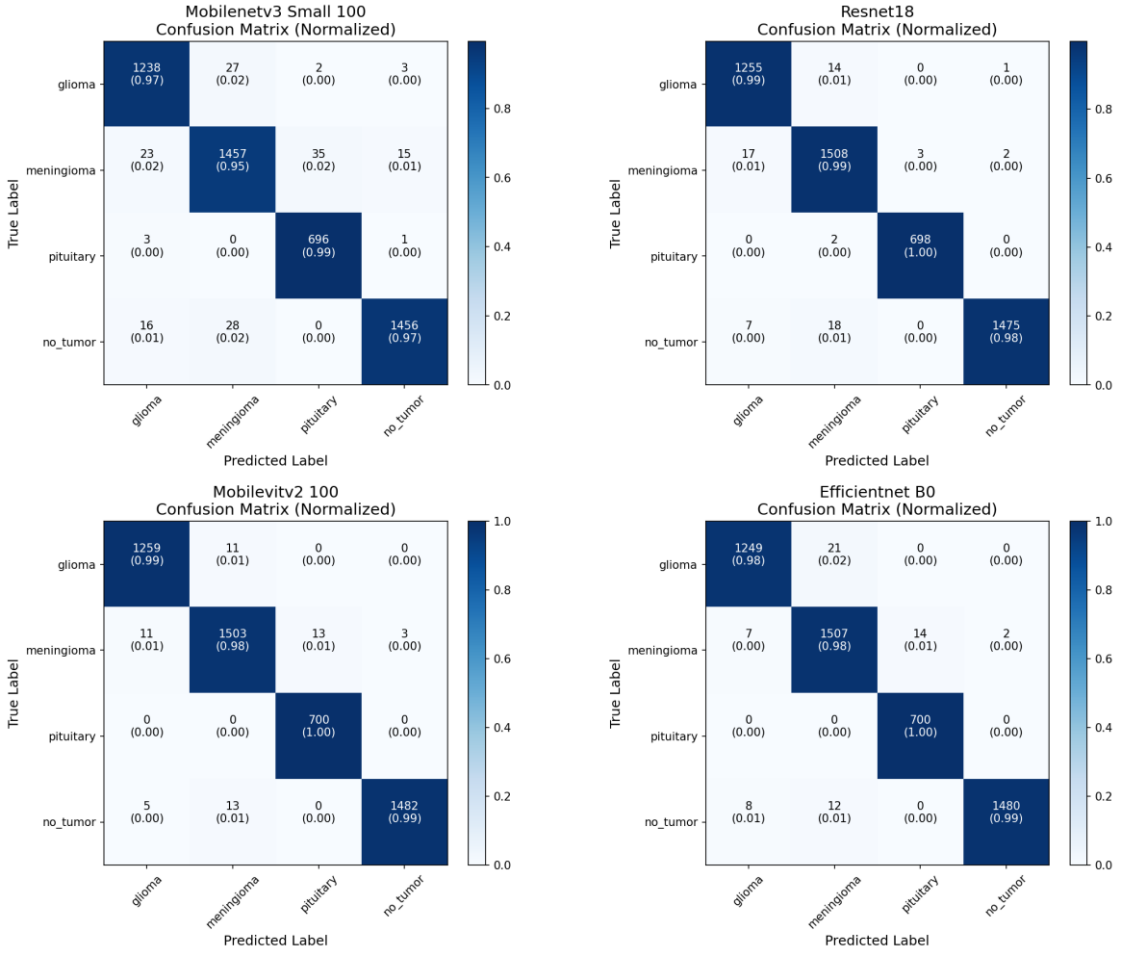

Figure 2. Confusion Matrices

1) MobileNetV3-Small: The most frequent error was the misclassification of meningioma as pituitary tumors, occurring in 35 instances. This suggests difficulty in distinguishing between these two often similarly appearing tumor types.
2) ResNet18: The most common error was misclassifying no-tumor scans as meningioma (18 instances). This may indicate that the model is sensitive to subtle anatomical variations or artifacts in the meninges that mimic early-stage tumors.
3) MobileViTV2: This model also struggled most with confusing meningioma for pituitary tumors, but did so with the lowest frequency among all models (13 instances), indicating greater robustness in differentiating these classes.

4) EfficientNet-B0: Its primary confusion was misclassifying glioma as meningioma (21 instances), pointing to a potential challenge in distinguishing between the infiltrative boundaries of gliomas and the more well-defined appearance of meningiomas.

Overall, the top-performing models, particularly MobileViTV2, made fewer critical errors, suggesting a more nuanced understanding of the distinguishing features between clinically similar tumor types.

3.4. STATISTICAL SIGNIFICANCE OF RESULTS

Statistical analysis was performed to validate the observed performance differences.

a) Overall Difference: Both the ANOVA test ($F(3,16) = 24.21, p < 0.001$) and the Friedman test ($\chi^2(3) = 9.37, p = 0.0248$) confirmed that there were statistically significant differences in the mean test accuracies across the four models.

b) Pairwise Differences: Post-hoc paired t-tests revealed that the performance of MobileNetV3-Small was significantly lower than that of ResNet18 ($p = 0.004$), MobileViTV2 ($p = 0.006$), and EfficientNet-B0 ($p = 0.009$). However, no statistically significant differences were found in the pairwise comparisons among the top three models: ResNet18, MobileViTV2, and EfficientNet-B0 (all $p > 0.05$). This indicates that these three models form a top tier of performance on this specific task and dataset.

4. DISCUSSION

4.1. INTERPRETATION OF FINDINGS: HYBRID VIT VS. PURE CNNS

The results demonstrate that modern CNNs and hybrid architectures can achieve state-of-the-art performance in brain tumor classification. While the top three models were statistically indistinguishable in terms of accuracy, the slight numerical advantage of MobileViTV2 is noteworthy. Its hybrid design, which integrates the local feature extraction capabilities of convolutions with the global context modeling of transformers, is theoretically well-suited for medical imaging [15], [17]. Tumors are characterized not only by their internal texture and cellular structure (local features) but also by their shape, mass effect, and relationship to surrounding anatomical structures (global features). The self-attention mechanism in MobileViTV2 allows it to weigh the importance of features across the entire image, potentially enabling a more holistic understanding than purely convolutional approaches [13]. The model's lower number of critical meningioma → pituitary misclassifications further supports the notion that its global context awareness may help resolve ambiguities between tumor types that appear in similar locations.

4.2. THE PERFORMANCE VS. EFFICIENCY TRADE-OFF

This study uncovered a critical and non-linear relationship between theoretical model complexity and practical inference speed. The most striking finding is the performance of ResNet18, which, despite being the largest model by parameter count, size, and FLOPs, was the fastest in single-batch inference. This paradox can be attributed to the deep optimization of standard 3x3 convolutions within GPU-accelerated libraries like cuDNN [30]. In contrast, newer architectures like MobileNetV3 and MobileViTV2 rely on less common operations such as depthwise separable convolutions and attention mechanisms. While these operations reduce the theoretical FLOP count, they can lead to increased memory access costs or less optimized execution paths on current hardware, resulting in higher real-world latency [31]. This finding serves as a crucial reminder that for clinical deployment, theoretical efficiency metrics like FLOPs are not a substitute for empirical benchmarking on the target hardware [32]. ResNet18, despite its size, may represent a highly practical choice where low latency is a primary concern.

4.3. IMPACT OF THE FULL FINE-TUNING STRATEGY

The experiment's use of a full fine-tuning strategy, where all pre-trained layers were unfrozen, proved highly effective. The high accuracies achieved by all models, coupled with the low performance gap between validation and test sets (indicating good generalization), suggest that this approach is well-suited for medical imaging tasks when a moderately large dataset (i.e., several thousand images) is available. The success of this strategy was likely enabled by the strong regularization techniques employed—Mixup, Label Smoothing, and Dropout—which collectively prevented the models from overfitting despite their large number of trainable parameters. This finding is valuable from a practical standpoint, as it simplifies the training pipeline by removing the need for complex, staged unfreezing protocols, making state-of-the-art model training more accessible

4.4. CLINICAL RELEVANCE OF ERROR PATTERNS

The misclassification patterns observed in the results are not random but reflect known diagnostic challenges in neuroradiology. The confusion between meningioma and pituitary tumors, seen in multiple models, is clinically plausible as both can present as well-circumscribed, contrast-enhancing masses at the base of the skull

[33]. Similarly, the confusion of no-tumor scans with meningioma by ResNet18 could stem from the model's high sensitivity to subtle, benign thickening of the meninges, which can mimic early-stage pathology [34]. The fact that the models' failure modes align with human diagnostic ambiguities suggests that they are learning clinically relevant radiological features. This reinforces their potential as diagnostic aids but also highlights the need for future work, such as incorporating multi-modal MRI sequences (e.g., T2-weighted or FLAIR) to provide complementary information that could help resolve these specific ambiguities [35], [36].

4.5. CONTEXTUALIZATION WITH EXISTING LITERATURE

The high accuracies achieved in this study, with top models exceeding 98.7%, are consistent with and advance the findings of recent literature in the field. Several studies utilizing CNN architectures like ResNet and EfficientNet on similar multi-class brain tumor datasets have reported accuracies in the 95-99% range [4], [8], [37], [38]. The superior performance of the hybrid MobileViTV2 aligns with a growing body of research demonstrating the advantages of transformer-based models in medical imaging, which are increasingly outperforming traditional CNNs by effectively capturing global dependencies [13], [14]. This study therefore validates these broader trends on a new, large-scale, and balanced benchmark dataset.

4.6 LIMITATIONS AND FUTURE WORK

This study has several limitations that provide avenues for future research. First, the analysis was conducted on a single, albeit high-quality, dataset (BRISC). Validating the findings on external, multi-institutional datasets is necessary to confirm the generalizability of the models. Second, only T1-weighted contrast-enhanced MRI scans were used [39], [40]. Future work should explore multi-modal fusion, incorporating T2-weighted, FLAIR, and other sequences, which could provide complementary information to resolve diagnostic ambiguities [41]. Additionally, while full fine-tuning was effective, a comparative study against other training strategies, such as progressive unfreezing [42], could yield further insights into optimal knowledge transfer. Finally, this was a retrospective study; the ultimate validation of any diagnostic model requires prospective clinical trials to assess its real-world impact on clinical workflows and patient outcomes [43], [44].

5. CONCLUSION

This study conducted a comprehensive comparative analysis of four deep learning models for multi-class brain tumor classification on the BRISC dataset. The findings demonstrate that the hybrid vision transformer architecture, MobileViTV2, and the established CNNs, ResNet18 and EfficientNet-B0, achieve statistically equivalent, state-of-the-art performance with test accuracies approaching 99%. While MobileNetV3-Small offers significant advantages in model size and parameter count, its lower accuracy makes it less suitable for this high-stakes diagnostic task.

The analysis revealed a crucial performance-efficiency paradox, where the model with the highest theoretical complexity, ResNet18, yielded the lowest practical inference latency, highlighting the importance of empirical hardware-specific testing over reliance on theoretical metrics like FLOPs for deployment decisions. Furthermore, the success of a full fine-tuning strategy, supported by robust regularization, validates it as a powerful and straightforward approach for adapting pre-trained models to medical imaging tasks with moderately large datasets. The models' error patterns were found to be clinically relevant, reinforcing their potential as diagnostic aids while also pinpointing areas for future improvement. Ultimately, this work validates the BRISC dataset as a robust benchmark and confirms that both advanced CNNs and hybrid transformer architectures are capable of achieving exceptional accuracy in brain tumor classification, paving the way for more reliable automated tools in clinical neuro-oncologyConclusion is statement referring to the purpose linked research with results and Discussion from research.

BIBLIOGRAPHY

[1] M. S. Madhan S, E. N. Rani S.E, S. K. Santhiya K, P. S. Praveena S, and S. P. M. D, “Brain Tumor Classification with Optimized EfficientNet Architecture,” *J Neonatal Surg*, vol. 14, no. 14S, pp. 627–637, Apr. 2025, doi: 10.52783/jns.v14.3990.

[2] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, “Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging,” *Cancers (Basel)*, vol. 15, no. 16, 2023, doi: 10.3390/cancers15164172.

[3] X. Jiang and W. Yu, “Brain tumor classification based on SFFM-ResNet18,” in *Proc.SPIE*, Jul. 2025, p. 136643U. doi: 10.1117/12.3070774.

[4] R. R. Ali *et al.*, “Learning Architecture for Brain Tumor Classification Based on Deep Convolutional Neural Network: Classic and ResNet50,” *Diagnostics*, vol. 15, no. 5, 2025, doi: 10.3390/diagnostics15050624.

[5] M. P. Kumar, D. Hasmitha, B. Usha, B. Jyothsna, and D. Sravya, "Brain Tumor Classification Using MobileNet," in *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 2024, pp. 1–7. doi: 10.1109/ICICACS60521.2024.10499117.

[6] Y. B. A. Sembiring and E. Indra, "MRI Image Classification Analysis of Brain Cancer Using ResNet18 and VGG16 Deep Learning Architectures," *INFOKUM*, vol. 13, no. 05, pp. 1537–1547, Jul. 2025, doi: 10.58471/infokum.v13i05.2955.

[7] M. M. Haj Hashem Khani and F. Maleki Nodehi, "Brain Tumor Detection in MRI Images Using ResNet18 Convolutional Neural Network and Transfer Learning," *Transactions on Machine Intelligence*, vol. 7, no. 4, pp. 269–275, 2024, doi: 10.47176/TMI.2024.269.

[8] S. R. Kempanna *et al.*, "Revolutionizing brain tumor diagnoses: a ResNet18 and focal loss approach to magnetic resonance imaging-based classification in neuro-oncology," *International Journal of Electrical and Computer Engineering (IJECE); Vol 14, No 6: December 2024DO - 10.11591/ijece.v14i6.pp6551-6559*, Dec. 2024, [Online]. Available: https://ijece.iaescore.com/index.php/IJECE/article/view/36023

[9] M. A. Rahman, M. B. Miah, Md. A. Hossain, and A. S. M. S. Hosen, "Enhanced Brain Tumor Classification Using MobileNetV2: A Comprehensive Preprocessing and Fine-Tuning Approach," *BioMedInformatics*, vol. 5, no. 2, 2025, doi: 10.3390/biomedinformatics5020030.

[10] A. K. M. Masum *et al.*, "Comparative Evaluation of Transfer Learning for Classification of Brain Tumor Using MRI," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 697–702. doi: 10.1109/ICMLA58977.2023.00102.

[11] C. Guo, Q. Zhou, J. Jiao, Q. Li, and L. Zhu, "A Modified MobileNetV3 Model Using an Attention Mechanism for Eight-Class Classification of Breast Cancer Pathological Images," *Applied Sciences*, vol. 14, no. 17, 2024, doi: 10.3390/app14177564.

[12] R. R. Sharma, A. Sungheetha, M. Tiwari, I. A. Pindoo, V. Ellappan, and G. G. S. Pradeep, "Comparative Analysis of Vision Transformer and CNN Architectures in Medical Image Classification," in *Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024)*, Atlantis Press, 2025, pp. 1343–1355. doi: 10.2991/978-94-6463-718-2_112.

[13] Y. Zhang, "A Comparative Analysis Between CNNs and ViTs for MRI-based Brain Tumor Classification," *Highlights in Science, Engineering and Technology*, vol. 124, pp. 30–37, Feb. 2025, doi: 10.54097/s64djm51.

[14] Z. Aboobacker, "A Comparative Analysis of CNN and Vision Transformer Architectures for Brain Tumor Detection in MRI Scans ," Jul. 2025, *Zenodo*. doi: 10.5281/zenodo.15973756.

[15] A. M. Kocharekar, S. Datta, Padmanaban, and R. R, "Comparative Analysis of Vision Transformers and CNN-based Models for Enhanced Brain Tumor Diagnosis," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 2024, pp. 1217–1223. doi: 10.1109/ICACRS62842.2024.10841744.

[16] S. Jraba, M. Elleuch, H. Ltifi, and M. Kherallah, "Comparative Analysis of CNNs and Vision Transformer Models for Brain Tumor Detection," in *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, SciTePress, 2025, pp. 1432–1439. doi: 10.5220/0013381900003890.

[17] J. Liu, X. Luo, D. Wang, F. Li, J. Li, and R. Lan, "MobileVitV2-Based Fusion of Vision Transformers and Convolutional Neural Networks for Underwater Image Enhancement," in *2023 13th International Conference on Information Science and Technology (ICIST)*, 2023, pp. 195–204. doi: 10.1109/ICIST59754.2023.10367056.

[18] A. Fateh *et al.*, "BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification with Swin-HAFNet," Jun. 2025. doi: 10.48550/arXiv.2506.14318.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 630–645.

[20] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, Sep. 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[21] A. Howard *et al.*, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[22] S. Mehta and M. Rastegari, "Separable Self-attention for Mobile Vision Transformers," 2022. doi: 10.48550/arXiv.2206.02680.

[23] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[24] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts.," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://arxiv.org/abs/1608.03983v5

[25] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization.," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *International Conference on Learning Representations*, ICLR, 2018. [Online]. Available: https://arxiv.org/abs/1710.09412

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jun. 2014.

[29] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, 2009, doi: https://doi.org/10.1016/j.ipm.2009.03.002.

[30] S. Chetlur *et al.*, "cuDNN: Efficient Primitives for Deep Learning," *arXiv preprint arXiv:1410.0759*, 2014, [Online]. Available: https://arxiv.org/abs/1410.0759

[31] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 122–138.

[32] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017, doi: 10.1109/JPROC.2017.2761740.

[33] Q. T. Ostrom *et al.*, "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012–2016," *Neuro Oncol*, vol. 21, no. Supplement_5, pp. v1–v100, Nov. 2019, doi: 10.1093/neuonc/noz150.

[34] J. G. Smirniotopoulos, F. M. Murphy, E. J. Rushing, J. H. Rees, and J. W. Schroeder, "Patterns of Contrast Enhancement in the Brain and Meninges," *RadioGraphics*, vol. 27, no. 2, pp. 525–551, Mar. 2007, doi: 10.1148/rg.272065155.

[35] S. Bakas *et al.*, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," *arXiv e-prints*, p. arXiv:1811.02629, Nov. 2018, doi: 10.48550/arXiv.1811.02629.

[36] S. H. Patel *et al.*, "T2–FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project," *Clinical Cancer Research*, vol. 23, no. 20, pp. 6078–6085, Oct. 2017, doi: 10.1158/1078-0432.CCR-17-0560.

[37] N. F. Othman and S. W. Kareem, "Enhancing Brain Tumor Classification Accuracy Using Deep Learning with Real and Synthetic MRI Images," *Zanco J Pure Appl Sci*, vol. 37, no. 4, pp. 126–149, 2025, doi: 10.21271/ZJPAS.37.4.11.

[38] A. Iqbal, M. A. Jaffar, and R. Jahangir, "Enhancing Brain Tumour Multi-Classification Using Efficient-Net B0-Based Intelligent Diagnosis for Internet of Medical Things (IoMT) Applications," *Information*, vol. 15, no. 8, 2024, doi: 10.3390/info15080489.

[39] A. Kaushal, R. Altman, and C. Langlotz, "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms," *JAMA*, vol. 324, no. 12, pp. 1212–1213, Sep. 2020, doi: 10.1001/jama.2020.12067.

[40] M. Nagendran *et al.*, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, p. m689, Mar. 2020, doi: 10.1136/bmj.m689.

[41] K. S. Choi, S. H. Choi, and B. Jeong, "Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network," *Neuro Oncol*, vol. 21, no. 9, pp. 1197–1209, Sep. 2019, doi: 10.1093/neuonc/noz095.

[42] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339. doi: 10.18653/v1/P18-1031.

[43] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *Lancet Digit Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019, doi: 10.1016/S2589-7500(19)30123-2.

[44] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat Med*, vol. 28, no. 1, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.