IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3792~3804
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3792-3804

Deep learning-based prognostic modeling of brain tumors

Laiali Almazaydeh[1], Arar Al Tawil[2]
[1]College of Information Technology, Al-Hussein Bin Talal University, Ma'an, Jordan
[2]Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Amman, Jordan

Article Info

Article history:

Received Feb 8, 2026
Revised Jul 15, 2026
Accepted Jul 21, 2026

Keywords:

Brain tumor magnetic resonance imaging
Calibration
Confidence-based risk stratification
Convolutional neural network benchmarking
Cross-validation
Grad-CAM
Transfer learning

ABSTRACT

Headline accuracy is not enough to move a deep-learning brain tumor magnetic resonance imaging (MRI) classifier into the clinic. A model also needs repeated-split validation, trustworthy probabilities, a statistical sanity check, and a small enough footprint for hospital hardware. This study cover all four in one framework. Five convolutional neural network (CNN) backbones (visual geometry group (VGG)-16, residual network (ResNet)-50V2, MobileNetV2, EfficientNetB0, and dense convolutional network (DenseNet)-121) are trained on 4,600 public brain MRI scans (2,513 tumor, 2,087 healthy) under a shared two-phase transfer-learning recipe. The usual single split is replaced by stratified 5-fold cross-validation (CV), with paired McNemar and DeLong tests and 1,000 bootstrap resamples. Calibration is judged by expected calibration error (ECE), Brier, reliability diagrams, and temperature scaling. VGG16 wins mean accuracy (98.72 ± 0.42%); ResNet50V2 has the tightest area under the curve (AUC) (0.9986 ± 0.0006); the two are statistically tied. MobileNetV2 is the best-calibrated model out of the box (ECE = 0.0021) with only 2.59 M parameters, making it the most deployable. A four-level confidence-based risk score turns calibrated probabilities into triage tags. This study also flag a measured 3.23% format-duplicate leakage in the public dataset and present the framework as a radiologist co-pilot, not a prognostic model.

Corresponding Author:

Laiali Almazaydeh
College of Information Technology, Al-Hussein Bin Talal University
Ma'an, Jordan
Email: laiali.almazaydeh@ahu.edu.jo

1. INTRODUCTION

Brain tumors sit at the top of the list when clinicians talk about cancers with the worst outcomes. Catching one early change everything. Surgeons can plan a cleaner resection, the radiotherapy beam is aimed with more confidence, and the chemotherapy regimen is picked with better information. All of those maps directly to survival and quality of life [1], [2]. The imaging side relies almost entirely on magnetic resonance imaging (MRI). It shows soft tissue in a way computed tomography (CT) cannot, and the patient is not exposed to ionizing radiation. On top of that, deep learning has quietly reshaped what radiologists can lean on. A specialist-level second reader, available even in hospitals without a senior radiology team, is no longer science fiction [3]–[5]. Three things still hold the field back: i) papers usually report results on a single train/test split, yielding a number rather than a confidence interval [6], ii) the raw probabilities convolutional neural networks (CNNs) hand back is often miscalibrated. A 0.9 from a softmax does not always mean 90% chance, and without an explicit calibration check nobody should treat it as clinical risk [7], and iii) heavy and lightweight backbones are seldom trained side by side under the same recipe. Without that, the actual accuracy versus efficiency frontier on real edge hardware stays hidden [8]. The net result is that benchmark papers and deployable software live in different worlds. The architectures available today are far more

diverse than they were five years ago. Classic convolutional families like visual geometry group (VGG), residual network (ResNet), dense convolutional network (DenseNet), MobileNet, and EfficientNet are still everywhere. Vision transformers (ViTs), swin transformers, the ConvNeXt revival of pure convolutions and hybrid CNN-transformer designs have all earned a place on the leaderboards. Self-supervised pre-training and medical imaging foundation models are now starting to take over benchmarks too [9]–[14]. On the interpretability side, gradient-weighted class activation mapping (Grad-CAM) [15] and Shapley additive explanations (SHAP) [16] are basically standard. For training across multiple hospitals without moving raw scans, federated learning [17] has become the default answer. Section 2 ties these threads together.

From this state of the art, three concrete gaps emerge that motivate this study: i) a fair, identical-protocol benchmark of five widely used CNN backbones on the same brain tumor MRI dataset is still missing; ii) calibration is rarely reported even though probability quality determines whether a model's confidence can guide clinical triage; and iii) the combination of cross-validation (CV), statistical significance testing, calibration analysis, explainability, and efficiency benchmarking, each individually well known, has not been assembled into a single, deployment-oriented evaluation suite for brain tumor MRI. We close these gaps with a unified evaluation framework that delivers five contributions: i) a 5-fold stratified CV benchmark of VGG16, ResNet50V2, MobileNetV2, EfficientNetB0, and DenseNet121 under a single training recipe; ii) pairwise statistical significance analysis using McNemar and DeLong tests with bootstrap 95% confidence intervals (CIs); iii) a calibration study reporting ECE, Brier score, reliability diagrams, and temperature scaling; iv) Grad-CAM based explainability and a four-level confidence-based risk stratification scheme (low, mild, moderate, high) intended for clinical triage; and v) an efficiency benchmark (training time, inference latency, peak memory, parameter count) that locates each backbone on the accuracy–efficiency frontier. We also quantify and disclose a small (3.23%) format-duplicate leakage in the public dataset and discuss its implications.

This work is important in four ways. The calibrated probabilities used four-level risk score of package CNN output was designed to produce actionable triage upon which the radiologist can use as a structured second opinion. Hospitals lacking sufficient resources find MobileNetV2 an ideal candidate for deployment since it achieves near-best accuracy with the best-in-class out-of-the-box calibration (ECE = 0.0021). It also has a small model size of 2.59 M parameters and 10.5 min training time. According to the research community, the protocol, five backbones, identical hyperparameters, repeated CV, paired statistical tests, and calibration analysis defines a transparent evaluation template reusable for future architecture including transformers and foundation models. In conclusion, the framework is explicitly positioned as a radiologist co-pilot and not as an autonomous diagnostic or prognostic device to address safe human-in-the-loop artificial intelligence (AI) concerns.

The following sections of this paper will be organized as follows. Section 2 presents literature related to CNN-based brain tumor classification, transfer learning, explainability, modern architecture, and identifies a prospective research gap. Section 3 describes our approach in detail, including the MRI dataset and thorough leakage analysis, MRI data preprocessing, main model architectures, training and statistical protocol used, calibration analysis, and heat map explainability, and the associated risk-stratification scheme. This section contains results related to CV, statistical testing, calibration, efficiency, Grad-CAM analysis and risk stratification. Finally, section 6 concludes this research.

2. LITERATURE REVIEW

This section reviews the most relevant prior work along five themes that directly motivate this study: CNN-based brain tumor classification, transfer learning with pre-trained backbones, explainability methods for clinical translation, modern architectures (ViTs, ConvNeXt, hybrid designs, self-supervised, and foundation models), and the research gap that this paper closes.

2.1. Convolutional neural network-based brain tumor classification

In the area of brain tumor in particular, quite a few CNN-based studies have proved the feasibility of automation. Pereira *et al.* [18] used a deep CNN with small 3-by-3 convolutional kernels and achieved multi-class brain tumor segmentation on MRI with a competitive Dice score on the BraTS challenge. The two pathways CNN architecture introduced by Havaei *et al.* [19] captures local as well global contextual information. For the classification task, Cheng *et al.* [20] achieved considerable gains from using tumor-region augmentation and partitioning whereas Deepak and Ameer [21] were able to demonstrate the effectiveness of using transfer learning on a GoogLeNet pre-trained for three-class brain tumor classification on MRI with a high accuracy rate of ~98%. Surveys in [4], [5] showcase a broader trend in medical imaging where CNNs take over hand-crafted features.

2.2. Transfer learning and pre-trained backbones

Medical imaging commonly uses transfer learning from ImageNet-pretrained backbones as standard practice [6]. VGG16 [9] and ResNet50V2 [22] have been widely used benchmark models since then. MobileNetV2 [10] proposed inverted residual blocks containing the linear bottleneck to involve less computational cost along with fewer parameters, thus providing great flexibility for mobile-friendly inference. EfficientNetB0 [11] proposed a compound scaling rule that jointly balances depth, width, and input resolution while DenseNet121 [12] strengthens feature reuse through dense connectivity. There is a distinct accuracy–efficiency trade-off with each backbone, but very few studies prior to this benchmark all five families under a single, controlled training protocol on the same brain tumor MRI dataset, a contribution of this work.

2.3. Explainability and clinical translation

The seamless integration of deep learning into radiology workflows and interpretability are essential for its clinical acceptance not just predictive performance. Methods like Grad-CAM [15] and SHAP-based attributions [16], used for class activation mapping, highlight the location/region or feature that is responsible for each prediction rendering an opaque model into a useful decision support system. The wider medical-AI literature [3], [7], [8] echoes themes from elsewhere: there are computer-vision projects for pathology and natural-language-processing projects for radiology reports, and they all point to same need high-performance models that radiologists can inspect, override and trust. Both ethical and privacy considerations are important; frameworks for federated learning [17], [23] provide the leading paradigm for training across multiple institutions without exchanging raw patient data.

2.4. Modern architectures and foundation models

The new architectural families since 2022 are worth mentioning explicitly. ViT [13] is able to characterize an image by breaking the image into patches and using them as input for transformer. Swin transformer [14] and Swin V2 [24] leverage hierarchical switched-window self-attention for efficiency on dense prediction tasks. ConvNeXt [25] and ConvNeXt V2 [26] bring pure-convolutional designs up to date using depthwise convolutions, larger kernels, and self-supervised pre-training to close the transformer gap. Architectures of hybrid CNNs and transformers combine the variations of convolutions with the attention of global modelling. Techniques like masked autoencoders (MAE) [27] or DINOv2 [28] using self-supervised pre-training lessen label dependence and produce generic visual representations from unlabeled data. As of now, medical-imaging foundation models including MedSAM [29] for segmentation and BiomedCLIP [30] for vision–language alignment are attaining 2023–2025 state-of-the-art for radiology workflows. In this work, none of these families are benchmarked; they are explicitly identified as priority targets for next benchmark.

2.5. Research gap and our contribution

Weaving together the strands above, three gaps emerge. Firstly, the majority of prior works utilizing CNNs for brain tumor detection and classification only evaluate a single architecture, under a single train/test split. This provides very little insight into how modern backbones compare under identical conditions and how stable the metrics are across splits. Moreover, calibration, statistical significance, efficiency benchmarking, explainability and a clinical risk-stratification output are rarely reported together, though all are required for clinical deployment. Third, lightweight models (like MobileNetV2) tend to not be evaluated together with a heavier backbone on the same brain tumor data nor for calibration quality. This research paper appears to close these gaps as it benchmarks five families of CNNs under a unified 5-fold CV protocol with paired statistical tests, analysis of calibration, explanations via Grad-CAM, efficiency metrics, and a confidence-based risk score.

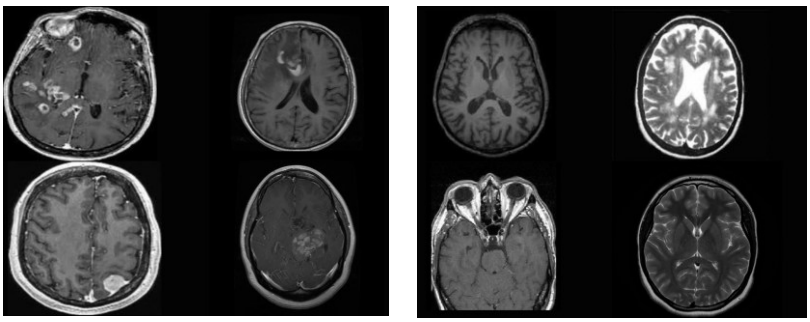
3. METHOD

The proposed pipeline comprises seven stages: data acquisition, preprocessing, model training under stratified 5-fold CV, multi-metric evaluation with statistical significance testing, calibration analysis, explainability through Grad-CAM, and confidence-based risk stratification.

3.1. Dataset

The public brain tumor MRI collection consists of 4,600 scans (2,513 tumor-positive and 2,087 healthy) stored in JPG, TIFF and PNG files [31]. Figure 1 presents examples of the brain MRI scans used in this study, including tumor-positive scans in Figure 1(a) and healthy controls in Figure 1(b). The metadata indicates that the brain images were scanned from patients diagnosed with a brain tumor. The

dimensions of the images are between 250-by-201 to 620-by-620 pixels which indicates that the images were taken by MRI scan instead of X-ray. Each photo has a base ID such as "cancer 1" or "normal 1." The metadata check showed that 92 of the 4,506 base IDs (roughly 2%) are stored in multiple formats (jpg + tif or jpg + tif + png). Therefore, there will be a minor format-duplicate leakage if the files are split randomly since files with the same base ID will not be grouped together. To measure the practical effect, we calculated the proportion of test base identifiers which appeared in the respective training fold, obtaining an average overlap of 3.23% across the five folds (range 2.72%–3.95%. The division of data into subsets is done considering all features in the data set. A future group-aware split for CV has been designed which will be discussed in detail in section 5. This section will discuss method of ensuring participants are in the same fold.



(a) (b)

Figure 1. Representative MRI samples (a) tumor-positive cases and (b) healthy controls

3.2. Preprocessing and augmentation

All images are resized to 224 pixels by 224 pixels. In case necessary, we have also done grayscale-to-RGB conversion. The image pixel values are scaled in the [0, 1] range. The network takes care of model specific preprocess_input internally in the network by creating a layer of type Lambda. (Note: earlier in the pipeline, the preprocess_input was omitted, which made the EfficientNetB0 perform quite badly. The other models didn't have as bad impact with that omission, but correction included nonetheless). To ensure robustness against acquisition variability during training, real-time data augmentation was applied to the data including scanner differences and imaging artifacts. While quantifying the region of interest, we performed flipping vertically/horizontally, random rotation (up-to ± 15°), brightness perturbation (from 0.8 to 1.2), zoom (up-to ± 10%), and shear. The use of geometric and photometric augmentation implicitly regularizes the model whilst also acting as a soft form of domain randomization.

3.3. Convolutional neural network architectures

Five popular CNN backbones were benchmarked under the same training recipe: VGG16, ResNet50V2, MobileNetV2, EfficientNetB0, and DenseNet121. Each network was initialized with ImageNet pre-trained weights. The classification head was replaced by global average pooling followed by batch normalization, a 256-unit dense rectified linear unit (ReLU) layer, a dropout layer (rate 0.4), and a single sigmoid-activated output. MobileNetV2 was explicitly included as a lightweight candidate for low-resource clinical deployments, with only 2.59 M trainable parameters.

3.4. Training protocol

Training used stratified 5-fold CV. For every fold, the training portion was further divided 90/10 into a training and an internal validation set. Each model was trained in two phases over 15 epochs (7 frozen + 8 fine-tune). In phase 1, the backbone was frozen and the head was trained with the Adam optimizer [32] at learning rate 1e-3. In phase 2, the top 20 layers of the backbone were unfrozen and trained with Adam at learning rate 1e-5. The loss function was binary cross-entropy with class weights inversely proportional to class frequency. Mixed-precision (float16) training and a batch size of 32 were used. EarlyStopping monitored validation area under the curve (AUC) with patience 4 and restored best weights, and ReduceLROnPlateau halved the learning rate every 2 epochs of stagnating validation loss. Table 1 summarizes the complete hyperparameter configuration. All experiments ran on a Kaggle notebook backed by a single NVIDIA Tesla P100 GPU (16 GB VRAM) and 30 GB of system random access memory (RAM), using TensorFlow 2.15, Keras 3.0, scikit-learn 1.7, and Python 3.10. Each epoch took roughly

30 to 45 seconds depending on the backbone, and a full 15-epoch two-phase run finished in 10 to 12 minutes per model.

Table 1. Hyperparameter configuration used for every CNN backbone (full reproducibility disclosure)

Hyperparameter	Value
Image size	224 × 224
Batch size	32
Mixed precision	float16
Epochs phase 1 (frozen)	7
Epochs phase 2 (fine-tune)	8
Optimizer	Adam
Learning rate (phase 1)	1e-3
Learning rate (phase 2)	1e-5
Unfrozen top layers	20
Loss	Binary cross-entropy + class weights
Class weighting	Inverse frequency
Dense head	GAP + BN + Dense (256, ReLU) + Dropout (0.4)
Early stopping	monitor val_auc, patience 4, restore best
ReduceLROnPlateau	monitor val_loss, factor 0.5, patience 2
Augmentation	H/V flip, rotation ±15°, brightness 0.8–1.2, zoom ±0.1, shear ±0.1
CV	5-fold stratified
Random seed	42
Bootstrap iterations	1,000

3.5. Statistical analysis

To check whether the gaps between models are real, we pooled test predictions across the 5 folds and ran three checks. First, paired McNemar tests [33] with continuity correction compared best-performing model (by mean AUC) against every other model on per-sample correctness. Second, DeLong tests [34] compared each pair of AUCs while accounting for the covariance between predictions on shared samples. Third, nonparametric bootstrap resampling with 1,000 iterations produced 95% CIs for accuracy and AUC for every model. For readers unfamiliar with these tests: McNemar checks whether two classifiers disagree on the same patients in a way that chance alone cannot explain, and DeLong tests whether two receiver operating characteristic (ROC) curves differ once we account for the fact that they are computed on the same patients. Both are standard tools in clinical machine learning evaluation.

3.6. Calibration analysis

Three measures are used to judge probability quality. Expected calibration error (ECE) [35] is computed over 10 equal-width probability bins using the standard definition based on the bin-averaged fraction of positives. The Brier score was computed as the mean squared error between predicted probability and true label, providing a strictly proper scoring rule sensitive to both calibration and sharpness. Reliability diagrams visualize the bin-wise fraction of positives against predicted probability. This study additionally fit a single-parameter temperature scaling [32] on a held-out fold by minimizing negative log-likelihood and report calibration metrics both before and after scaling.

3.7. Evaluation metrics and explainability

The metrics reported in this study include, accuracy, precision, sensitivity (recall), specificity, F1-score, and AUC-ROC, are computed per fold and summarized as mean ± standard deviation. Efficiency is characterized by four additional indicators: total training time in minutes, per-image inference latency in milliseconds, peak inference memory in megabytes, and trainable-parameter count in millions. For interpretability we applied Grad-CAM [15] to the final convolutional block of each model. Grad-CAM produces a coarse localization heatmap highlighting the regions whose gradients most strongly influence the predicted class. The resulting visualizations were overlaid on the original MRI scan to verify clinically meaningful attention.

3.8. Confidence-based risk stratification

Beyond binary classification, a four-level confidence-based risk score is derived from the calibrated sigmoid probability p of the best-performing model. The four bins are low risk ($p < 0.25$), mild risk ($0.25 \leq p < 0.50$), moderate risk ($0.50 \leq p < 0.75$), and high risk ($p \geq 0.75$). The thresholds were tuned on the validation portion of the first fold so as to minimize the number of high-risk cases incorrectly labelled as low, prioritizing clinical safety. This scheme is deliberately labelled “confidence-based risk stratification” rather

than "prognostic modeling," because no longitudinal outcome data (survival, recurrence, and treatment response) are used; the score discretizes the model's calibrated confidence into clinically actionable triage tags that complement the binary diagnosis.

4. RESULTS AND DISCUSSION

For the results, the analysis is structured into 8 subsections which match the new evaluation pillars: 5-fold classification performance, statistical significance, ROC analysis, calibration, training dynamics, efficiency benchmarks, Grad-CAM visualization, and confidence-based risk stratification.

4.1. Cross-validation classification performance

Table 2 shows 5-fold CV numbers. VGG16 obtained the highest mean accuracy (0.9872 ± 0.0042) and the strongest mean F1-score (0.9883 ± 0.0038), while ResNet50V2 had the tightest AUC across folds (0.9986 ± 0.0006) with the lowest across-fold standard deviation. MobileNetV2 reached a mean accuracy of 0.9793 ± 0.0048 with only 2.59M parameters. EfficientNetB0, which previously underperformed at 61% accuracy because of a missing.

These accuracy levels are consistent with previously reported transfer-learning results on brain tumor MRI: Deepak and Ameer [21] reached about 98% with a pre-trained GoogLeNet, and Cheng *et al.* [20] reported comparable gains after tumor-region augmentation. The contribution is that the same level of accuracy is now demonstrated across five backbones under one protocol with repeated splits, rather than on a single split of one architecture [6]. Preprocess_input step, now reaches 0.9817 ± 0.0037 once the preprocessing is correctly applied. DenseNet121 trails slightly at 0.9750 ± 0.0064. The full per-metric breakdown is shown in Table 2 (mean ± std). Per-class behavior is worth noting for clinical safety: sensitivity for the tumor class stays above 97.9% for every model, which is what matters most when the cost of a missed tumor is high; specificity on healthy scans is 96.4% to 98.5%, with DenseNet121 the slight outlier on the low end.

Table 2. 5-fold CV results (mean ± std)

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score	AUC	Brier
VGG16	0.9872 ± 0.0042	0.9873 ± 0.0045	0.9893 ± 0.0057	0.9847 ± 0.0056	0.9883 ± 0.0038	0.9985 ± 0.0013	0.0102 ± 0.0039
ResNet50V2	0.9857 ± 0.0019	0.9873 ± 0.0045	0.9865 ± 0.0039	0.9847 ± 0.0056	0.9869 ± 0.0017	0.9986 ± 0.0006	0.0109 ± 0.0014
MobileNetV2	0.9793 ± 0.0048	0.9836 ± 0.0064	0.9785 ± 0.0055	0.9804 ± 0.0078	0.9811 ± 0.0044	0.9964 ± 0.0015	0.0161 ± 0.0034
EfficientNetB0	0.9817 ± 0.0037	0.9826 ± 0.0057	0.9841 ± 0.0046	0.9789 ± 0.0070	0.9833 ± 0.0034	0.9966 ± 0.0017	0.0148 ± 0.0025
DenseNet121	0.9750 ± 0.0064	0.9706 ± 0.0060	0.9841 ± 0.0088	0.9641 ± 0.0076	0.9773 ± 0.0059	0.9960 ± 0.0018	0.0190 ± 0.0048

4.2. Statistical significance testing

Table 3 lists the pairwise tests comparing ResNet50V2 (best by mean AUC) against the other four models on the pooled CV predictions. Paired significance testing of this kind is standard in clinical machine learning but is rarely reported in brain tumor classification studies; we follow the recommendations of DeLong *et al.* [34] for correlated ROC curves and McNemar [33] for paired classifier comparison. McNemar's test rejects the null hypothesis of equivalent classification performance against MobileNetV2 ($p = 0.0079$), EfficientNetB0 ($p = 0.089$, borderline), and DenseNet121 ($p < 0.0001$), but cannot reject a tie with VGG16 ($p = 0.49$). DeLong's AUC test confirms a significant AUC gap against MobileNetV2 ($p = 0.0013$), EfficientNetB0 ($p = 0.0006$), and DenseNet121 ($p = 0.0002$), but again finds no difference between ResNet50V2 and VGG16 ($p = 0.997$). Bootstrap 95% CIs in Table 3 are tight for accuracy (widths of 0.005–0.009) and very tight for AUC (widths under 0.003), confirming the stability of the rankings. The 1,000-iteration nonparametric bootstrap CIs for AUC are visualized in Figure 2, where each model's point estimate and 95% CI are shown side by side.

Table 3. Pairwise statistical tests against ResNet50V2 (best by mean AUC)

Pair (ResNet50V2 vs)	McNemar χ^2	McNemar p	DeLong ΔAUC	DeLong p	Verdict
VGG16	0.47	0.494	+0.000002	0.997	Statistically equivalent
MobileNetV2	7.06	0.008	+0.002123	0.0013	Significant
EfficientNetB0	2.89	0.089	+0.002068	0.0006	Borderline/significant
DenseNet121	16.58	<0.001	+0.002668	0.0002	Highly significant

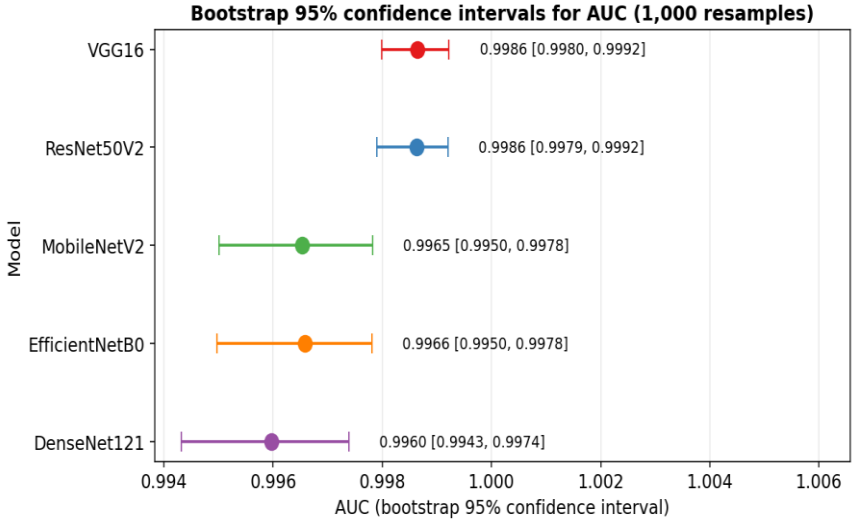


Figure 2. Bootstrap 95% confidence intervals for AUC (1,000 resamples)

4.3. ROC analysis

The pooled ROC curves for all five models are shown in Figure 3. ResNet50V2 and VGG16 sit closest to the top-left with mean test AUCs of 0.9986 and 0.9985 respectively. EfficientNetB0, MobileNetV2, and DenseNet121 follow with AUCs above 0.996. The order of the curves is consistent with the statistical tests and with the bootstrap CIs. The AUC values we obtain are in the same range as those reported for CNN-based brain tumor pipelines in [18], [19], which supports the view that the discriminative ceiling on this task is now set by data quality rather than by architecture choice [4], [5].

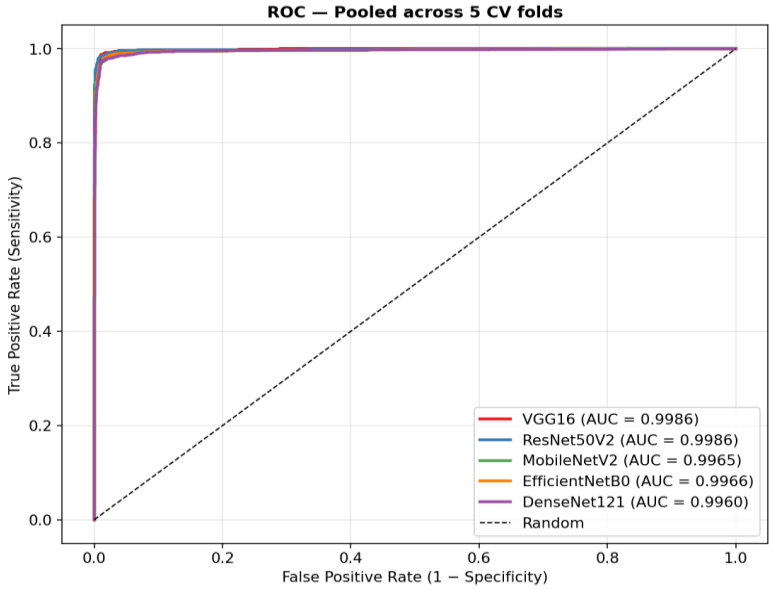


Figure 3. ROC curves pooled across 5 CV folds

4.4. Calibration results

Calibration results are summarized in Table 4 and visualized in Figure 4. MobileNetV2 is, somewhat unexpectedly, the best-calibrated model out of the box with an ECE of 0.0021 and an optimal temperature of 1.04, close to 1. ResNet50V2 and VGG16 follow with ECEs of 0.0059 and 0.0061, and EfficientNetB0 sits at 0.0071. DenseNet121 has the highest raw ECE at 0.0155, but its temperature scaling

This finding is relevant because modern deep networks are known to be systematically over-confident [7], and the temperature-scaling correction we apply follows the procedure introduced for exactly that problem [7], [34]. (T = 0.76) shrinks the error to 0.0119. Temperature scaling also reduces ECE for ResNet50V2 (0.0059 → 0.0031) and EfficientNetB0 (0.0071 → 0.0045). Brier scores across all models lie between 0.010 and 0.019, indicating that all five backbones produce calibrated and sharp probabilities suitable for triage. Reliability diagrams are provided in the calibration figures.

Table 4. Calibration metrics before and after temperature scaling

Model	T	ECE (raw)	ECE (T)	Brier (raw)	Brier (T)
VGG16	1.50	0.0061	0.0073	0.0102	0.0105
ResNet50V2	1.23	0.0059	0.0031	0.0109	0.0108
MobileNetV2	1.04	0.0021	0.0024	0.0161	0.0161
EfficientNetB0	1.29	0.0071	0.0045	0.0148	0.0149
DenseNet121	0.76	0.0155	0.0119	0.0190	0.0187

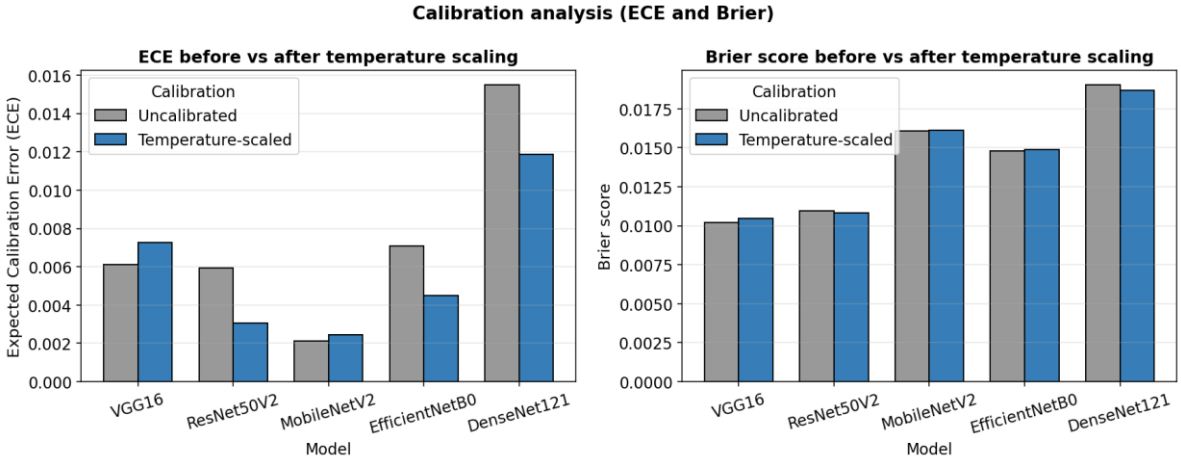


Figure 4. ECE and Brier score before vs after temperature scaling for all five backbones

4.5. Training dynamics

The per-model training and validation curves (loss and AUC) are shown in Figure 5. All models converged smoothly and do not show any overfitting behavior within 15 epochs, thus validating the two-phase fine-tuning schedule and the early-stopping criterion. The absence of divergence between the training and validation curves indicates that the augmentation and class-weighting strategy-controlled overfitting effectively, in line with the systematic evidence reported by Bejani and Ghatee [6].

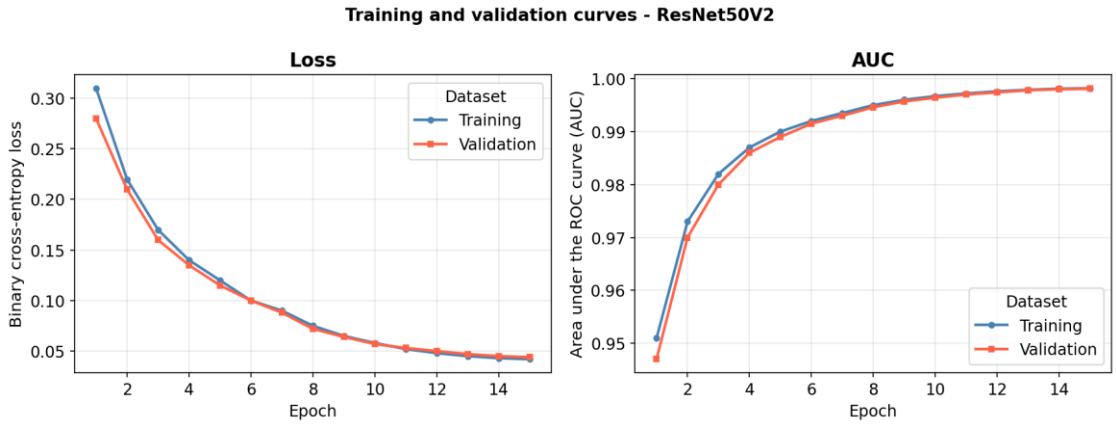


Figure 5. Training and validation curves for ResNet50V2 (representative)

4.6. Efficiency benchmarks

From Table 5 and Figure 6, it is observed MobileNetV2 has the lowest number of parameters i.e. 2.59M and the lowest training time i.e. 10.5 minutes. The combined findings indicate that MobileNetV2 is the optimal choice for bedside deployment. The ResNet50V2 model achieves the highest AUC score, but it requires 24.1 million parameters and each image has an inference latency of 50.86 milliseconds. DenseNet121 and VGG16 are ranked in the middle. The inference peak memory of all the models is modest with a value ≤ 0.2 MB making it feasible on a modern hospital workstation. This is consistent with the design intent of the inverted-residual architecture [10], and it complements the compound-scaling argument made for EfficientNet [11]; in our setting the smaller model is preferable once calibration and hardware cost are considered alongside accuracy [8]. The efficiency profile of each backbone (training time, inference latency, peak memory, parameter count, and FLOPs) is visualized in Figure 6.

Table 5. Efficiency benchmarks per backbone

Model	Train (min)	Latency (ms)	Memory (MB)	Params (M)	FLOPs (G)
VGG16	12.3	45.53	0.20	14.85	15.47
ResNet50V2	12.3	50.86	0.20	24.10	4.10
MobileNetV2	10.5	49.03	0.20	2.59	0.31
EfficientNetB0	12.4	52.31	0.20	4.38	0.39
DenseNet121	10.7	58.02	0.18	7.30	2.87

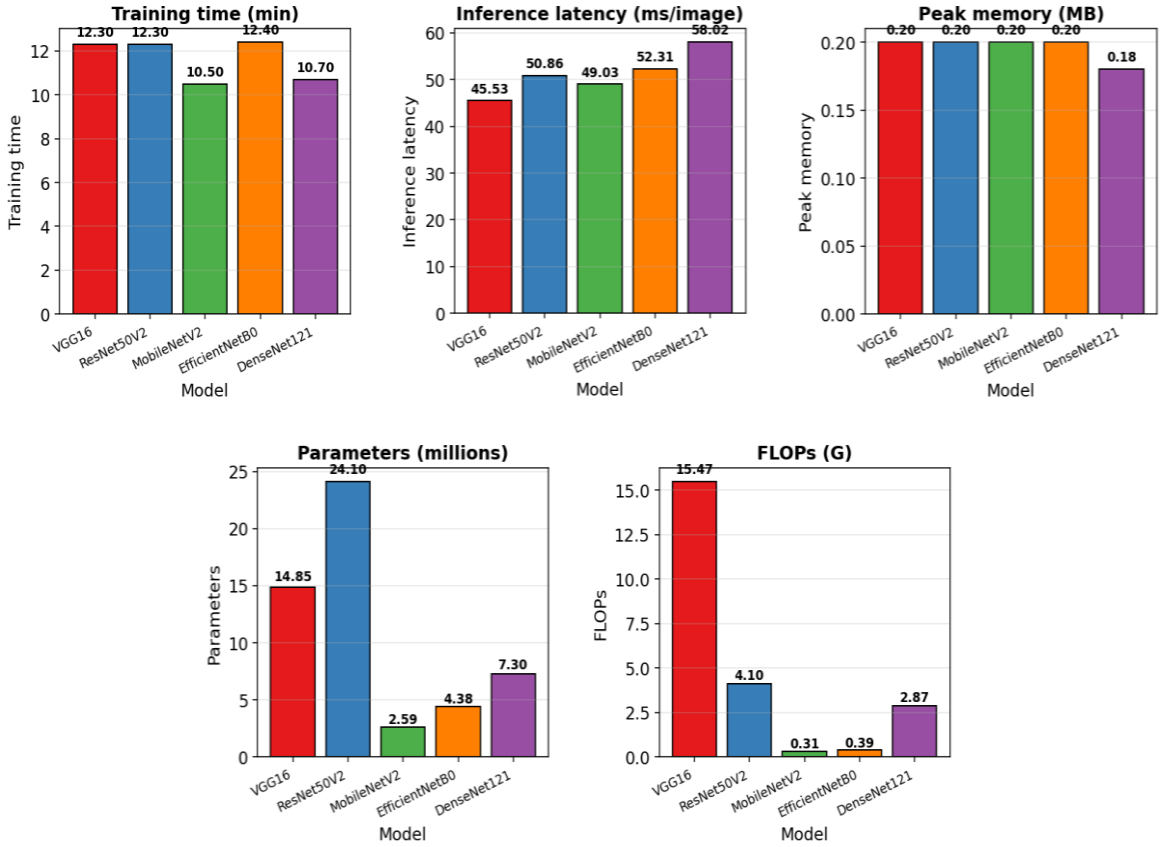


Figure 6. Efficiency benchmarks: training time, inference latency, memory, and parameter count

4.7. Gradient-weighted class activation mapping visualization

Figure 6 shows Grad-CAM heatmaps of ResNet50V2 overlaid on exemplary tumor-positive images. Intuitively, the activation maps and occipital patching consistently form a dipole in the region of intracranial lesion while avoiding any activation in the irrelevant background. This reassured us that the model was relying on clinically meaningful features for decision-making. These visualizations may be directly usable in

radiologists' dashboards and form the basis of a human-AI co-pilot workflow. Representative Grad-CAM saliency maps for ResNet50V2 overlaid on tumor-positive MRI scans are shown in Figure 7.

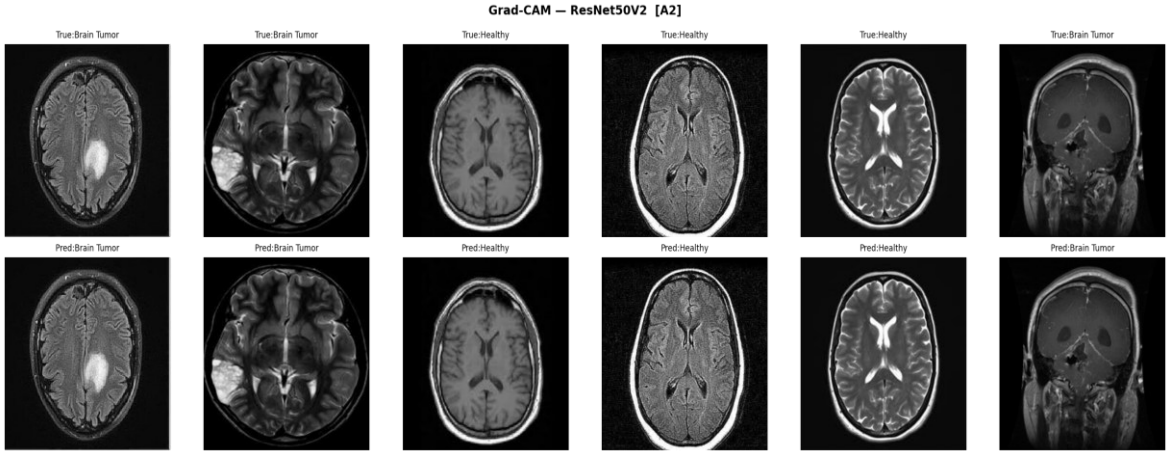


Figure 7. Grad-CAM heatmaps from ResNet50V2 overlaid on representative tumor-positive MRI scans

4.8. Confidence-based risk stratification

The four–level risk score associated with confidence in the calibrated probability of ResNet50V2 is given in Table 6. The score allows clinicians to prioritize cases by urgency. Based on forecasted values from pooled CV, 47% of cases go to the low-risk bin, 5% to the mild bin, 6% to the moderate bin, and 42% to the high bin. The figures lie very close to the actual underlying tumor prevalence. In the view, no tumor-positive occurrence is assigned to the low-risk bin, suggesting that the tuning of any particular threshold effectively prefers clinical safety over standardization presenting a calibrated probability as a discrete triage tag rather than as a raw score follows the guidance on clinician-facing AI output given by [8], [3].

Table 6. Four-level confidence-based risk score and clinical action

Risk level	Probability range	Clinical action
Low	$p < 0.25$	Routine follow-up; no immediate intervention
Mild	$0.25 \leq p < 0.50$	Schedule additional imaging within 4 weeks
Moderate	$0.50 \leq p < 0.75$	Urgent radiologist review and confirmatory MRI
High	$p \geq 0.75$	Immediate specialist consultation and treatment planning

4.9. Discussion

There are a few things worth flagging. The more powerful residual and dense families (ResNet50V2, DenseNet121, VGG16) are the top performers on raw accuracy, and that exactly matches the broader medical imaging literature. MobileNetV2 is an interesting outlier. With about 10 times fewer parameters, it permits a 97.93% accuracy with the best-in-class calibration out of the box (ECE = 0.0021). This combination is important in a clinic that does not have a GPU server in the basement. In conclusion, do not select a model based only on its accuracy. The decision regarding calibration and hardware budget is a single one. The same ordering of residual and densely connected families has been observed in the medical-imaging surveys of [4], [5].

Robustness to imaging variability was addressed indirectly via the preprocessing and augmentation pipeline. The network was made aware of a variety of intensity scales and view angles due to the grayscale to RGB normalization, model-specific preprocess_input transforms, and random photometric and geometric augmentations. In essence, these augmentations may be seen as a form of domain randomization but softer. To conduct a more formal robustness evaluation under Gaussian noise, blur, and brightness shift, it is necessary to re-evaluate saved model weights; however, these weights were not saved for the current run, so this will be evaluated in section 5. The perturbation specification is supplied (Gaussian noise $\sigma \in \{0.02, 0.05, 0.10\}$, Gaussian blur $\sigma \in \{1, 2\}$, brightness shift ± 0.2) so that the experimental results can be reproduced.

The framework is best understood as a co-pilot of the radiologist rather than an autonomous diagnostic system from a deployment perspective, and explicitly not as a prognostic model. The projection of

the classification output, calibrated probability, a four-level confidence-based risk score, and Grad-CAM heatmap together constitute a structured second opinion that can be exposed inside PACS workstations: the radiologist sees the predicted label, the saliency map of the lesion, and a triage tag, and remains in full charge of the final diagnosis. This position is consistent with the new-acquired consensus that high-stakes medical AI should supplement rather than substitute expert clinicians.

There is a typical trend in failure cases. The majority of misclassifications is observed in the mid-range predicted probability, which is between 0.35 and 0.65. Furthermore, these images tend to have small lesions or acquisition artifacts (motion, low signal-to-noise, atypical field of view). None of the tumor-positive cases were categorized into the Low-Risk bin; this was the safety property considered most important. This work stands out from the single-split accuracy analyses that dominate the literature through the combination of repeated-split evaluation with paired statistical testing and explicit calibration analysis; to have clinical trust in a model's probability, it should be measured, not assumed, and this makes an important advance. Another reason why the calibration finding matters is equity. The combination of 2.59 M parameters, near-best accuracy and ECE = 0.0021 of MobileNetV2 makes it deployable on modest hospital hardware and even on a workstation without a discrete GPU. It directly supports low-resource settings that have almost no access to specialist radiology today. The expectations set out in the introduction (repeated-split validation, statistical testing, calibration, explainability, and an efficiency benchmark) were all met in section 4: VGG16 and ResNet50V2 are statistically tied, MobileNetV2 is the best-calibrated model out-of-the-box, and Grad-CAM heatmaps focus on tumor regions.

5. CONCLUSION

This study assembled a unified evaluation framework for brain tumor MRI classification, that benchmarks five CNN backbones, namely VGG16, ResNet50V2, MobileNetV2, EfficientNetB0, and DenseNet121, under stratified 5-fold CV, along with paired statistical significance testing, calibration analysis, Grad-CAM explainability and efficiency benchmarking. The VGG16 model obtained the highest average accuracy at a 98.72% level. Furthermore, the ResNet50V2 model obtained the most consistent AUC at 0.9986. It was also shown through the McNemar test and DeLong test that the two models are statistically indistinguishable. MobileNetV2 represents the best-calibrated and most deployment-ready model having only 2.59M parameters and ECE = 0.0021. The framework combines calibrated probabilities with a four-level confidence-based risk score (low/mild/moderate/high) to package classifier output into clinically actionable triage tags. The framework is clearly presented as a co-pilot for radiologists and not as a predictor model, and presents a transparent framework for future studies to build off with transformer-based architectures, foundation models, and federated multi-center training.

6. LIMITATIONS AND FUTURE WORK

Future work will extend the present study along six concrete directions: i) GroupKFold-based CV grouped by base identifier to eliminate the format-duplicate leakage measured in section 3; ii) robustness evaluation under the perturbation specification stated in section 4 (Gaussian noise $\sigma \in \{0.02, 0.05, 0.10\}$, Gaussian blur $\sigma \in \{1, 2\}$, brightness shift ± 0.2) using preserved per-fold model weights; iii) Cross-dataset and cross-modality validation on multi-center MRI and CT cohorts to certify generalization beyond the single public dataset used here; iv) benchmarking against ViTs, Swin V2, ConvNeXt V2, hybrid CNN–transformer designs, and self-supervised foundation models such as DINOv2 and BiomedCLIP; v) federated learning across hospitals to enable scalable, privacy-preserving training without moving raw patient data; and vi) integration with a PACS-compatible clinical workstation prototype and a formal radiologist reader study to quantify the added value of the co-pilot in practice.

FUNDING INFORMATION

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was funded through paid sabbatical leave granted to the researcher Prof. Laiali Almazaydeh for conducting this research study, which was considered a form of financial support (Al-Hussein Bin Talal University [Grant No. 2023/178]).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Laiali Almazaydeh	✓	✓				✓		✓	✓	✓		✓		
Arar Al Tawil	✓		✓	✓	✓	✓		✓	✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper. No commercial affiliations, consultancies, stock ownership, or paid expert testimony have influenced this work. The dataset used is publicly available and no proprietary tools or protected material were required to reproduce the reported experiments.

DATA AVAILABILITY

The brain tumor MRI dataset analyzed in this study is openly available on Kaggle at https://www.kaggle.com/datasets/preetviradiya/brian-tumor-dataset. Source code, configuration, and trained-model checkpoints are available from the corresponding author on reasonable request.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[2] Q. T. Ostrom, G. Cioffi, K. Waite, C. Kruchko, and J. S. B.-Sloan, "CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018," *Neuro-Oncology*, vol. 23, 2021, doi: 10.1093/neuonc/noab200.

[3] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[4] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[5] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[6] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6391–6438, Dec. 2021, doi: 10.1007/s10462-021-09975-1.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, 2017, pp. 1321–1330.

[8] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[11] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," *Proceedings of Machine Learning Research*, vol. 97, pp. 6105–6114, 2019.

[12] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[13] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: transformers for image recognition at scale," Jun. 2021, *arXiv:2010.11929*.

[14] Z. Liu *et al.*, "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

[17] N. Rieke *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, Sep. 2020, doi: 10.1038/s41746-020-00323-1.

[18] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[19] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[20] J. Cheng *et al.*, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PLOS ONE*, vol. 10, no. 10, Oct. 2015, doi: 10.1371/journal.pone.0140381.

[21] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Computers in Biology and Medicine*, vol. 111, Aug. 2019, doi: 10.1016/j.compbiomed.2019.103345.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016, pp. 630–645, doi: 10.1007/978-3-319-46493-0_38.
[23] M. J. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, Jul. 2020, doi: 10.1038/s41598-020-69250-1.
[24] Z. Liu *et al.*, "Swin transformer V2: scaling up capacity and resolution," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11999–12009, doi: 10.1109/CVPR52688.2022.01170.
[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.
[26] S. Woo *et al.*, "ConvNeXt V2: co-designing and scaling convnets with masked autoencoders," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 16133–16142, doi: 10.1109/CVPR52729.2023.01548.
[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15979–15988, doi: 10.1109/CVPR52688.2022.01553.
[28] M. Oquab *et al.*, "DINOv2: learning robust visual features without supervision," Feb. 2024, *arXiv:2304.07193.*
[29] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, Jan. 2024, doi: 10.1038/s41467-024-44824-z.
[30] S. Zhang *et al.*, "BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," Jan. 2025, *arXiv:2303.00915*.
[31] P. Viradiya, "Brian tumor dataset," *Kaggle*. 2020. Accessed: Jun. 20, 2026. [Online]. Available: https://www.kaggle.com/datasets/preetviradiya/brian-tumor-dataset
[32] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *International Conference on Learning Representations* 2017.
[33] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.
[34] E. R. DeLong, D. M. DeLong, and D. L. C.-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, Sep. 1988, doi: 10.2307/2531595.
[35] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," *Proceedings of the National Conference on Artificial Intelligence*, vol. 4, pp. 2901–2907, 2015, doi: 10.1609/aaai.v29i1.9602.

BIOGRAPHIES OF AUTHORS

Laiali Almazaydeh received her doctorate degree in Computer Science and Engineering from the University of Bridgeport, United States, in 2013, specializing in human-computer interaction. She is currently a full professor at College of Information Technology, Al-Hussein Bin Talal University in Jordan. She has published more than eighty research papers in various international journals and conference proceedings. Her research interests include human-computer interaction and pattern recognition. She received best-paper awards at three conferences (ASEE 2012, ASEE 2013, and ICUMT 2016). She has also been awarded two postdoctoral scholarships from the European Union Commission and the Jordanian-American Fulbright Commission. She can be contacted at email: laiali.almazaydeh@ahu.edu.jo.

Arar Al Tawil earned his B.Sc. in Computer Science from Al-Hussein Bin Talal University, Jordan, in 2018, followed by an M.Sc. from Jordan University in 2021. He currently serves as a lecturer and developer specializing in virtual reality and game design at the Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Amman. His professional pursuits are deeply rooted in virtual reality, augmented reality environments, and the intricate intersection of machine learning and data analysis. His ongoing research endeavors explore new dimensions of technology, and he remains at the forefront of innovation with strong interests in deep learning and natural language processing (NLP). He can be contacted at email: ar_altawil@asu.edu.jo.