BioMedInformatics

MDPI

Article

DualXBrNet: A Hybrid Dual-Path Cross-Level Fusion Framework for Multiclass Brain Tumor Classification Using MRI Images

Sarah Al-Azzawi and Ghaida A. Al-Suhail *

Department of Computer Engineering, College of Engineering, University of Basrah, Basrah 61004, Iraq; sarah.j.mohammed@uobasrah.edu.iq
* Correspondence: ghaida.suhail@uobasrah.edu.iq

Abstract

Brain tumor classification from magnetic resonance imaging (MRI) can support computer-aided diagnosis by enabling automated differentiation of tumor and non-tumor categories. This study presents DualXBrNet, a hybrid dual-level fusion framework for multiclass brain tumor classification using the original BRISC2025 dataset. Three pretrained convolutional neural networks (CNNs), ResNet50V2, MobileNetV2, and EfficientNetB0, were optimized to learn complementary representations of glioma, meningioma, pituitary tumor, and no-tumor MRI images. Their deep features were integrated through a learnable Feature Fusion network, while the class probability outputs of the three CNNs and the Feature Fusion model were subsequently combined using a leakage-aware five-fold out-of-fold (OOF) decision-level stacking strategy. A low-complexity Logistic Regression meta-classifier was trained exclusively on the resulting 16-dimensional OOF probability representations to prevent information leakage during meta-learning. On the independent official test set, DualXBrNet achieved 96.80% accuracy and a macro-average ROC-AUC of 99.77%. Oracle, statistical, computational complexity, feature space, and Occlusion Sensitivity analyses were further conducted to assess model complementarity, computational efficiency, representation quality, and interpretability. The results demonstrate that feature-level fusion provides the primary performance improvement while OOF-based decision stacking offers complementary refinement within a leakage-aware evaluation protocol. Overall, the proposed framework emphasizes methodological rigor through leakage-aware meta-learning, independent testing, and comprehensive model analysis, rather than relying on classification accuracy alone.

Keywords: brain tumor classification; MRI; DualXBrNet; transfer learning; feature fusion; out-of-fold stacking; logistic regression; occlusion sensitivity

Check for updates

Academic Editors: Saurav Mallik, Priyanka Roy and Zhongming Zhao

Received: 2 September 2026
Revised: 7 October 2026
Accepted: 12 October 2026
Published: 15 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Brain tumors represent a serious neurological health concern because of their heterogeneous appearance, variable progression, and potential impact on critical brain functions. Accurate identification of tumor type is, therefore, important for diagnosis, treatment planning, and patient management. Magnetic resonance imaging (MRI) is widely used for brain tumor assessment because it provides high soft-tissue contrast and detailed visualization of intracranial structures without ionizing radiation. However, manual interpretation can be challenging and time consuming, particularly when tumor classes exhibit overlapping visual characteristics or considerable intraclass variability. These limitations have motivated the development of computer-aided systems for more consistent and efficient MRI interpretation [1,2].

Artificial intelligence (AI), particularly deep learning (DL), has demonstrated considerable potential in medical image analysis. Convolutional neural networks (CNNs) can automatically learn hierarchical representations directly from medical images, reducing reliance on handcrafted features [3]. Transfer learning further improves their applicability by adapting representations learned from large-scale image datasets to medical imaging tasks with limited annotated data [4,5]. Architectures such as ResNet, MobileNet, and EfficientNet provide different representational and computational characteristics [6]. Nevertheless, the representations learned by an individual CNN remain influenced by its architectural characteristics and may not fully capture the diverse morphological patterns present across brain tumor classes [7]. This limitation has encouraged the use of complementary architectures and fusion strategies to exploit diverse learned representations.

On the other hand, Feature Fusion provides a mechanism for combining complementary information learned by different network layers, scales, or architectural branches. Feature representations may be integrated through concatenation, element-wise addition, fixed weighting, or learnable transformations. Concatenation preserves information from each branch by stacking feature vectors or maps, whereas element-wise addition produces a compact representation but requires compatible feature dimensions. Learnable fusion allows trainable layers to model nonlinear interactions among the combined representations. At the pipeline level, feature-level fusion integrates latent representations before classification, whereas decision-level fusion combines class scores or probability outputs generated by independently trained models [8,9]. Although both strategies have demonstrated advantages, using only one fusion level may leave complementary information unexploited, motivating the investigation of multi-level fusion.

Recent brain MRI studies have introduced transfer learning, hybrid architectures, attention mechanisms, and ensemble learning [10]. However, stacking needs a strict separation between the samples used to train the base models and those used to build the meta-classifier. The meta-level predictions derived from previously observed training samples may lead to information leakage and to overly optimistic estimates of performance. Furthermore, model complementarity, statistical reliability, computational complexity, and explainability are not always evaluated together within a unified framework [11].

To address these methodological considerations, this study proposes DualXBrNet, a dual-path cross-level fusion framework for multiclass brain tumor classification from MRI images. The methodological contribution of DualXBrNet does not lie in Feature Fusion, heterogeneous ensembling, or stacking individually. Instead, it establishes a hierarchical connection between representation-level fusion and decision-level stacking. Accordingly, DualXBrNet operates through two complementary paths: a representation-level Feature Fusion path and a decision-level stacking path. ResNet50V2, MobileNetV2, and EfficientNetB0 are employed as complementary CNN branches to generate heterogeneous deep representations and individual four-class probability outputs. Their latent representations are integrated through a learnable Feature Fusion pathway, which produces an additional four-class probability vector. This fusion-derived probability vector is then combined with the 12 probabilities produced by the three individual CNNs, forming a compact 16-dimensional cross-level meta-representation. Thus, the final meta-classifier receives both architecture-specific decisions and an additional decision derived from the fused feature representation.

To reduce information leakage during meta-learning, the 16-dimensional representation is constructed using a strict five-fold OOF protocol, whereby each meta-training prediction is generated for a sample that was excluded from training the corresponding base model. The contribution of the fusion-derived probability vector is explicitly evaluated by comparing the proposed 16-dimensional representation with conventional 12-dimensional

CNN-only stacking. Following OOF meta-learning, the constituent deep models are retrained using the complete development set, while the official test set remains isolated until final evaluation. This protocol separates base model training, meta-level learning, and final testing, thereby providing a controlled basis for assessing the contribution of cross-level fusion.

The main contributions of this study are summarized as follows:

1. A hierarchical cross-level fusion architecture is introduced in which the probability output learned from heterogeneous feature fusion is incorporated into decision-level stacking alongside the outputs of ResNet50V2, MobileNetV2, and EfficientNetB0.
2. A compact 16-dimensional meta-representation is constructed from 12 individual-CNN probabilities and four Feature Fusion probabilities. Its contribution is evaluated against 12-dimensional CNN-only stacking through paired ablation analysis.
3. A strict five-fold OOF protocol generates out-of-sample meta-features for all 5000 development images while maintaining complete isolation of the official test set.
4. A low-complexity multinomial Logistic Regression meta-classifier is employed as a conservative second-level learner, and its performance is compared with RBF-SVM, Random Forest, and MLP alternatives.
5. The framework is evaluated through classification metrics, controlled ablation, complementarity assessment, statistical testing, computational analysis, t-SNE/UMAP feature space visualization, and Occlusion Sensitivity. DualXBrNet achieves 96.80% accuracy on the isolated BRISC2025 official test set and 94.94% on an additional public-source dataset without retraining or fine-tuning.

The remainder of this paper is organized as follows. Section 2 reviews related studies on brain tumor classification and ensemble learning. Section 3 describes the dataset, preprocessing, OOF protocol, transfer learning models, and dual-level fusion framework. Section 4 presents the experimental results and complementary analyses. Section 5 discusses the findings, comparisons, and limitations, while Section 6 concludes the study and outlines future research directions.

2. Related Work

Recent brain MRI research has increasingly moved from single-backbone CNNs toward hybrid, attention-based, and ensemble architectures to improve multiclass tumor classification. Celik et al. [12] combined MobileNetV3 and EfficientNetB7 using a co-attention ensemble approach and achieved excellent results on Figshare and BraTS datasets, showing the benefit of fusing complementary feature maps from different networks.

Shaikh et al. [13] employed a stacking ensemble in which several CNNs, including MobileNetV3, VGG variants, ResNet50, AlexNet, and 3D CNN, served as base learners, with DenseNet201 as the meta-model. Haque et al. [14] further developed an explainable deep stacking framework integrating EfficientNetB0, MobileNetV2, GoogLeNet, and CapsuleNet with CatBoost as a meta-learner while also incorporating statistical validation and computational analysis.

More closely related hierarchical and meta-learning approaches have recently been reported. Ullah and Kim [15] proposed a dual-level ensemble that combines deep features extracted by pretrained vision transformers and aggregates the outputs of optimized machine learning classifiers. Kakon et al. [16] introduced an explainable deep ensemble meta-learning framework that combines multiple predictive branches and incorporates explainability analysis.

Most notably, Uthra and Sivashankari [17] employed four independent base architectures and concatenated their four-class probability outputs to construct a 16-dimensional meta-feature representation, which was generated using five-fold OOF prediction and classi-

fied using LightGBM. This approach closely resembles DualXBrNet in terms of meta-feature dimensionality and leakage-aware stacking. However, the fourth probability vector in DualXBrNet is not produced by an additional independent backbone. Instead, it is generated by a learnable Feature Fusion pathway that jointly integrates the latent representations of the same three heterogeneous CNNs. Consequently, the proposed meta-representation combines three architecture-specific probability vectors with an additional learned cross-model consensus. The incremental contribution of this fusion-derived probability vector is explicitly evaluated by comparing the 12-dimensional CNN-only representation with the proposed 16-dimensional cross-level representation.

Other recent approaches further illustrate the diversity of fusion and ensemble strategies. Chen et al. [18] fused aligned features extracted from ResNet101, DenseNet121, and Efficient-NetB0 before classification. Patro et al. [19] and Velpula et al. [20] employed ensembles of deep CNNs and voting-based decision aggregation, whereas Eone Etoua et al. [21] combined handcrafted local descriptors and lightweight CNN representations through a macro-F1-weighted ensemble.

Al-Azzawi and Al-Suhail [22] proposed BrTxNet, which integrates EfficientNetB0 and VGG16 through learnable feature-level fusion for binary brain tumor MRI classification. Although the framework demonstrated the effectiveness of learning complementary deep representations, it remained limited to feature-level fusion and did not incorporate OOF probability-level stacking.

Recent studies have also explored optimized ensemble weighting and adaptive fusion. Talukder et al. [23] used a combination of multiple transfer learning architectures using a genetic algorithm and grid search-based weight optimization and showed improved ensemble performance compared to individual CNNs. Singh et al. [24] proposed a dynamic ensemble of CNN, ResNet50, and EfficientNetB5 with adaptive model weighting and incorporated Grad-CAM, SHAP, SmoothGrad, and LIME to enhance interpretability. Panigrahi et al. [25] introduced a hybrid transfer learning and self-attention framework based on DenseNet201 and Transformer-style attention, while also comparing several pretrained CNNs, including MobileNetV2 and ResNet50V2.

Feature-level fusion has likewise become an important direction. Jayaraman et al. [26] proposed a CNN–ViT framework using cross-attention fusion to combine complementary local and global representations for multiclass MRI classification. Yonar [27] developed a hybrid framework that combined deep features from several pretrained CNNs with metric learning and swarm intelligence-based feature selection. Sanchez-Moreno et al. [28] investigated an ensemble of VGG16, DenseNet121, and Inception-ResNet-v2 with majority voting together with Grad-CAM++ and Integrated Gradients, showing the importance of combining ensemble prediction with explainability.

Furthermore, attention-based and explainable models have attained promising results in multiclass tumor recognition. Zarenia et al. [29] proposed a hierarchical deformable attention framework with saliency mapping for joint brain tumor classification and segmentation, and Mao et al. [30] enhanced DenseNet121 with dilated convolutions and Squeeze and Excitation attention to improve four-class MRI classification. Collectively, these studies confirm that feature fusion, heterogeneous ensembling, stacking, meta-learning, statistical evaluation, and explainability are established directions in brain tumor MRI classification. Accordingly, the contribution of DualXBrNet is not claimed to arise from any of these components in isolation. The specific methodological distinction is the OOF-based hierarchical connection between the two fusion paths: the learnable Feature Fusion branch produces a four-class probability representation that is incorporated alongside the 12 individual-CNN probabilities within a 16-dimensional OOF meta-representation.

This allows the meta-classifier to jointly model three architecture-specific decisions and an additional consensus decision learned from their fused latent features. The value of this connection is explicitly assessed through a paired ablation against 12-dimensional CNN-only stacking, rather than inferred solely from the final test accuracy. The framework is additionally examined through model complementarity, statistical, computational, representation-level, and explainability analyses to clarify both the benefits and limitations of the proposed integration.

3. Proposed Methodology

The overall methodological framework of DualXBrNet is illustrated in Figure 1. The workflow comprises five sequential stages: strict five-fold OOF model development, cross-level OOF meta-feature generation, Logistic Regression meta-classifier training, final retraining of the deep models on the complete development set, and independent evaluation on the official test set. This design ensures that the decision-level meta-classifier is trained exclusively on out-of-sample predictions while preserving complete isolation of the final test set.

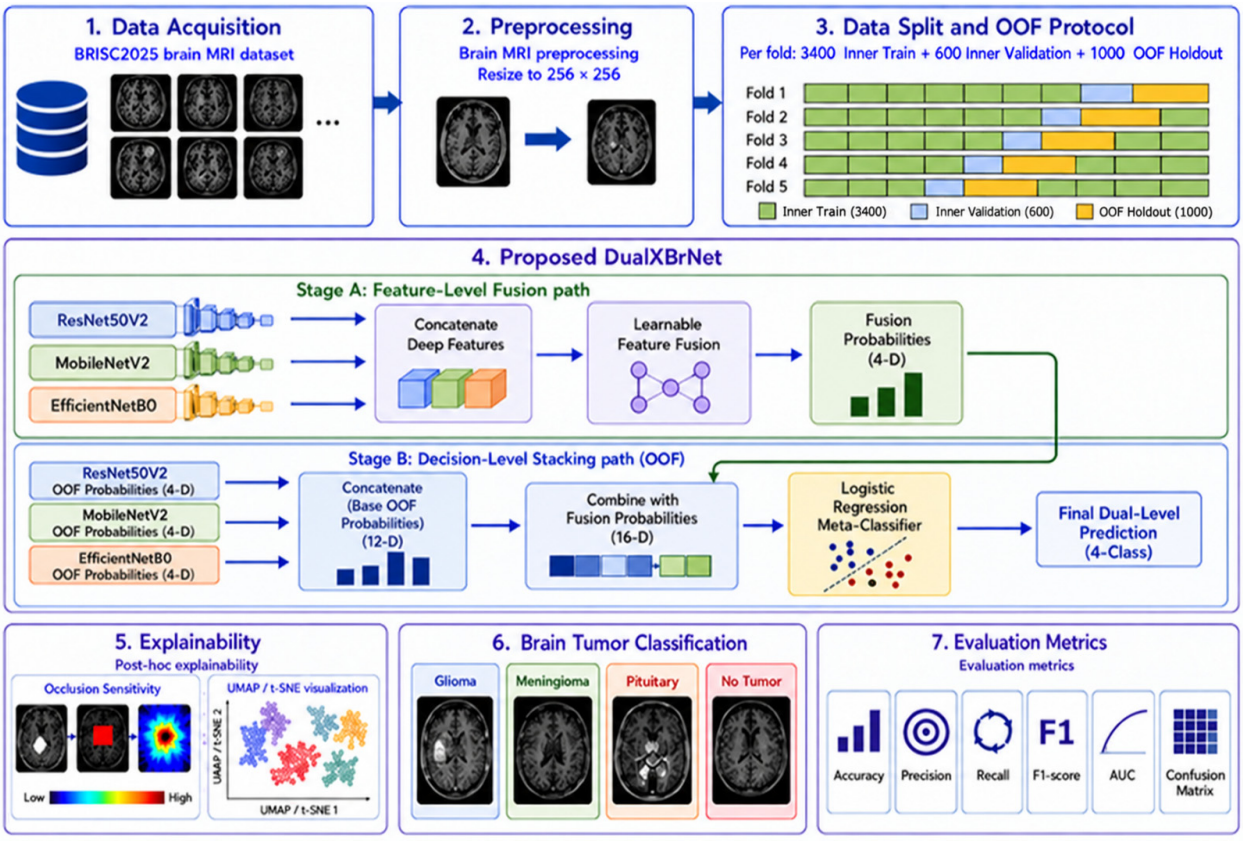


Figure 1. Overall workflow of the proposed DualXBrNet framework.

3.1. Datasets

3.1.1. BRISC2025 Dataset

The experiments were conducted using the original BRISC2025 brain MRI classification dataset [31], which contains 6000 MRI images distributed across four diagnostic categories: glioma, meningioma, no tumor, and pituitary tumor. The dataset provides an original 5000-image training partition and an independent 1000-image official test set. In this study, the original training partition was treated as the development set. The development set consisted of all samples available for model development, including CNN training, internal validation, five-fold OOF prediction generation, meta-classifier training, and final model retraining. In

contrast, the official test set remained completely isolated throughout model development and was used only for the final unbiased evaluation. No duplicate removal or additional dataset cleaning procedure was applied in order to preserve the original dataset composition.

As shown in Table 1, the class distribution is moderately imbalanced, particularly within the development set; therefore, class weighting was incorporated during model training to reduce bias toward the more represented classes.

Table 1. Distribution of the BRISC2025 dataset.

Class	Development Set	Official Test Set	Total
Glioma	1147	254	1401
Meningioma	1329	306	1635
No Tumor	1067	140	1207
Pituitary	1457	300	1757
Total	5000	1000	6000

3.1.2. Additional Public Dataset

To assess the behavior of the finalized framework on data obtained from another public source, an additional evaluation was conducted using the Brain Tumor MRI dataset published by Masoud Nickparvar [32]. The dataset contains 7200 MRI images evenly distributed across four classes, glioma, meningioma, no tumor, and pituitary tumor, with 1800 images per class. The original dataset includes 5600 images in its training partition and 1600 images in its testing partition.

The BRISC2025-trained ResNet50V2, MobileNetV2, EfficientNetB0, Feature Fusion, and Logistic Regression meta-classifier were applied directly to all 7200 images without retraining, fine-tuning, model selection, or parameter modification. Each image was resized to 256 × 256 pixels using the same bilinear interpolation and preprocessing pipeline employed in the original experiment. Performance was evaluated using accuracy, macro-precision, macro-recall, macro-F1, weighted F1, and one-versus-rest macro-AUC. The results were calculated for the complete dataset and separately for its original training and testing partitions. The Feature Fusion and DualXBrNet predictions were also compared using McNemar's exact test.

3.2. Preprocessing and Data Augmentation

All MRI images were resized to 256 × 256 pixels to provide a consistent input resolution for the three transfer learning models. The resized images were converted to floating-point tensors and supplied to the networks using the same input representation adopted during model development and final evaluation. No additional dataset cleaning or duplicate removal procedure was applied in this study.

To improve generalization and reduce overfitting, data augmentation was applied exclusively to the training subset of each fold. The augmentation pipeline consisted of random horizontal flipping, random rotation with a factor of 0.05 (corresponding to approximately −18° to +18°), random zoom within ±8%, horizontal and vertical translation within ±3% of the corresponding image dimension, and random contrast adjustment within ±5%. No vertical flipping or brightness augmentation was applied. The default reflection padding and bilinear interpolation settings of the TensorFlow/Keras augmentation layers were used. Augmentation was disabled for the internal validation, OOF hold-out, and official test samples to ensure that model selection and evaluation were performed on unaltered images.

Because the development set exhibits moderate class imbalance, class weights were calculated from the training data and incorporated into the loss function during CNN

training. The weighting scheme puts more emphasis on the under-represented classes during optimization, but it does not change the original distribution of classes.

3.3. Strict Five-Fold Out-of-Fold Protocol

To ensure leakage-aware stacking and complete separation between model development and final evaluation, a strict five-fold out-of-fold (OOF) protocol was applied exclusively to the 5000-image development set. Figure 2 explains the five-fold OOF-based cross-level fusion and meta-classifier training pipeline for DualXBrNet. The overall procedure consisted of five sequential stages as follows.

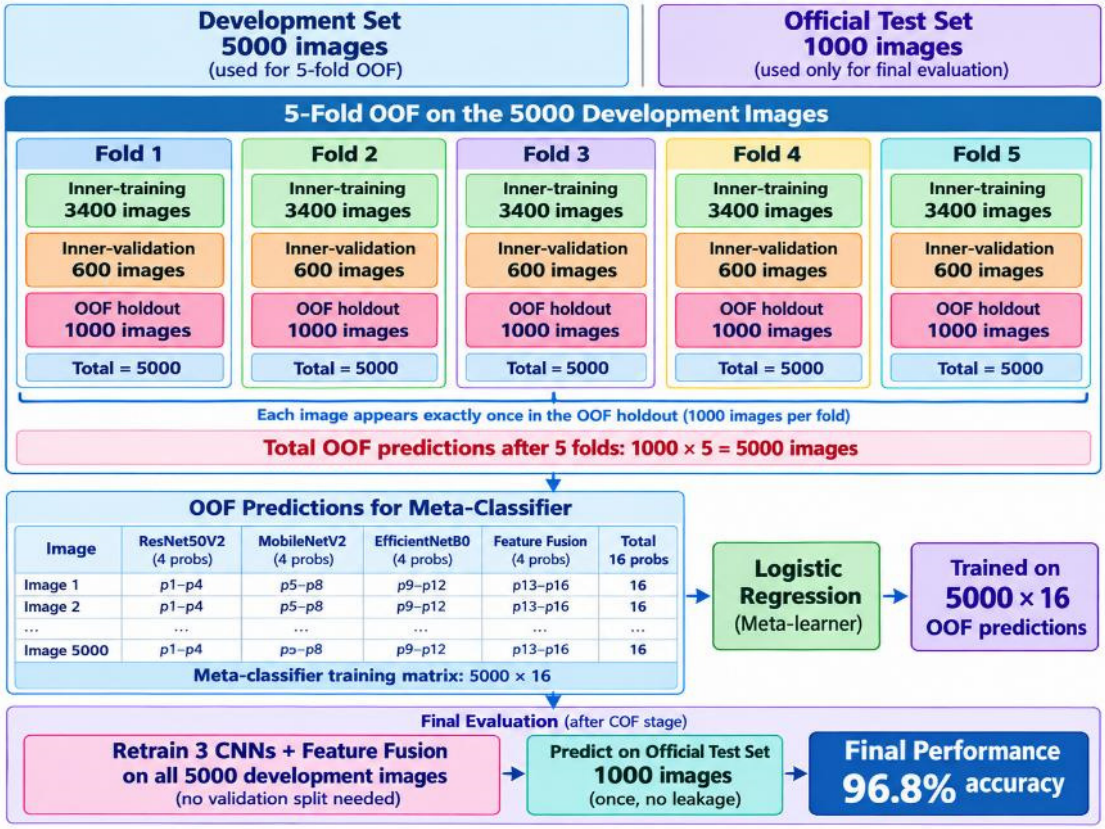


Figure 2. Five-fold out-of-fold training and decision-level stacking strategy of DualXBrNet.

Stage 1—OOF model development. The development set was partitioned into five stratified folds while preserving the original class distribution. In each iteration, one fold containing approximately 1000 images was held out as the OOF subset, while the remaining 4000 images formed the model development portion. Within each outer fold, the remaining 4000 images were further divided using a stratified 85:15 split, resulting in 3400 inner training images and 600 internal validation images. The split was generated using stratified random sampling to preserve the class proportions, with a fixed random seed of 42. The split indices were generated once for each outer fold and were consistently reused by ResNet50V2, MobileNetV2, EfficientNetB0, and the fold-specific Feature Fusion model. Consequently, all constituent models were trained and validated using identical samples within the corresponding fold, and checkpoint selection was based exclusively on the shared inner validation subset. The 1000-image OOF subset remained completely excluded from both inner training and internal validation and was used only to generate unseen predictions. Explicit index intersection checks confirmed the absence of overlap among the inner training, internal validation, and OOF subsets, and the complete fold assignments were stored to support reproducibility.

The 85:15 split was conceptually compared with the 80:20 alternative, as summarized in Table 2. The 85:15 configuration retains more samples for CNN training while providing a sufficient stratified validation subset for checkpoint selection, early stopping, and learning rate scheduling. Both configurations maintain complete isolation of the 1000-image OOF holdout and generate the same 5000 OOF predictions across the five folds.

Table 2. Comparison of candidate inner training and inner validation splits within each outer fold.

Criterion	Selected 85:15 Split	Alternative 80:20 Split
Inner training samples	3400 (85%)	3200 (80%)
Inner validation samples	600 (15%)	800 (20%)
Outer OOF holdout	1000	1000
Total samples per outer fold	5000	5000
Training data availability	Higher	Lower
Validation set size	Adequate	Larger
Checkpoint monitoring	Supported	Supported
OOF isolation	Complete	Complete
OOF coverage after five folds	5000 predictions	5000 predictions
Main consideration	Preserves more samples for CNN training while retaining a stratified validation subset	Provides a larger validation subset but reduces the samples available for model fitting

Stage 2—OOF meta-feature generation. The trained fold-specific models generated predictions only for the corresponding held-out OOF samples. For each image, ResNet50V2, MobileNetV2, EfficientNetB0, and the Feature Fusion pathway each produced a four-class probability vector. These outputs were concatenated to form a 16-dimensional OOF meta-feature vector. Repeating the procedure across all five folds ensured that every development image received exactly one prediction from models that had not been trained on that image.

The complete OOF meta-training matrix was constructed by concatenating the fold-specific representations:

$$Z_{OOF} = [Z^{(1)};Z^{(2)};Z^{(3)};Z^{(4)};Z^{(5)}] \in R^{5000\times16} \quad (1)$$

where $Z^{(k)}$ denotes the 16-dimensional OOF meta-feature matrix generated for the held-out samples of fold k, and the semicolon denotes row-wise concatenation. Accordingly, $Z_{OOF} \in R^{5000\times16}$ represents the complete OOF meta-training matrix obtained by concatenating the outputs of all five folds.

Stage 3—Meta-classifier training. After all five folds were completed, the resulting OOF meta-features from the full 500- image development set were used to train a multinomial Logistic Regression meta-classifier. Because the meta-classifier was trained exclusively on predictions generated from unseen samples, training to meta-level leakage was avoided.

Stage 4—Final retraining. After OOF generation and meta-classifier training were completed, ResNet50V2, MobileNetV2, EfficientNetB0, and the Feature Fusion network were retrained on the complete 5000-image development set using fixed training schedules selected from the fold-validation results. The trained Logistic Regression parameters were then kept fixed.

Stage 5—Independent final evaluation. The 1000-image official test set remained completely isolated during training, validation, OOF generation, checkpoint selection, and meta-classifier fitting. During final evaluation, the retrained CNNs and Feature Fusion model generated the same 16-dimensional probability representation for each test image, which was then passed to the previously trained Logistic Regression meta-classifier to obtain the final four-class prediction.

3.4. Transfer Learning Models

Transfer learning was employed to adapt pretrained convolutional neural networks to the four-class brain MRI classification task, as shown in Table 3. Three heterogeneous ImageNet pretrained architectures, ResNet50V2, MobileNetV2, and EfficientNetB0, were selected because they provide complementary representational characteristics and different computational profiles. Each backbone was independently optimized using the same overall training protocol. In the first training stage, the pretrained convolutional backbone was frozen while the newly added task-specific classification layers were trained. In the second stage, selected upper backbone layers were unfrozen and fine-tuned using a reduced learning rate to further adapt the learned representations to the MRI domain.

Table 3. Task-specific classification heads used for the transfer learning models.

Model	Added Layers
ResNet50V2	Global Average Pooling → Batch Normalization → Dense (256, ReLU) → Dropout (0.40) → Softmax (4 Classes)
MobileNetV2	Global Average Pooling → Dense (256, ReLU) → Dropout (0.25) → Softmax (4 Classes)
EfficientNetB0	Global Average Pooling → Batch Normalization → Dense (384, Swish) → Dropout (0.30) → Softmax (4 Classes)

ResNet50V2 was selected as a deep residual backbone with strong capability for MRI-based brain tumor representation learning, consistent with its recent use in hybrid brain tumor classification frameworks [33]. MobileNetV2 was included as a lightweight architecture that provides computationally efficient feature extraction and has demonstrated effectiveness in MRI-based brain tumor classification [34]. EfficientNetB0 was selected as an efficient transfer learning backbone that balances representational capability and computational cost and has recently been applied successfully to brain MRI analysis [35].

The penultimate representations of the three models were subsequently used as inputs to the proposed learnable Feature Fusion module.

The customized backbone architectures are illustrated separately in Figure 3a–c to provide a clear view of the task-specific modifications applied to each transfer learning model. The subsequent fusion stages are presented independently in later figures to distinguish the individual CNN adaptation from the proposed feature-level and decision-level fusion mechanisms.

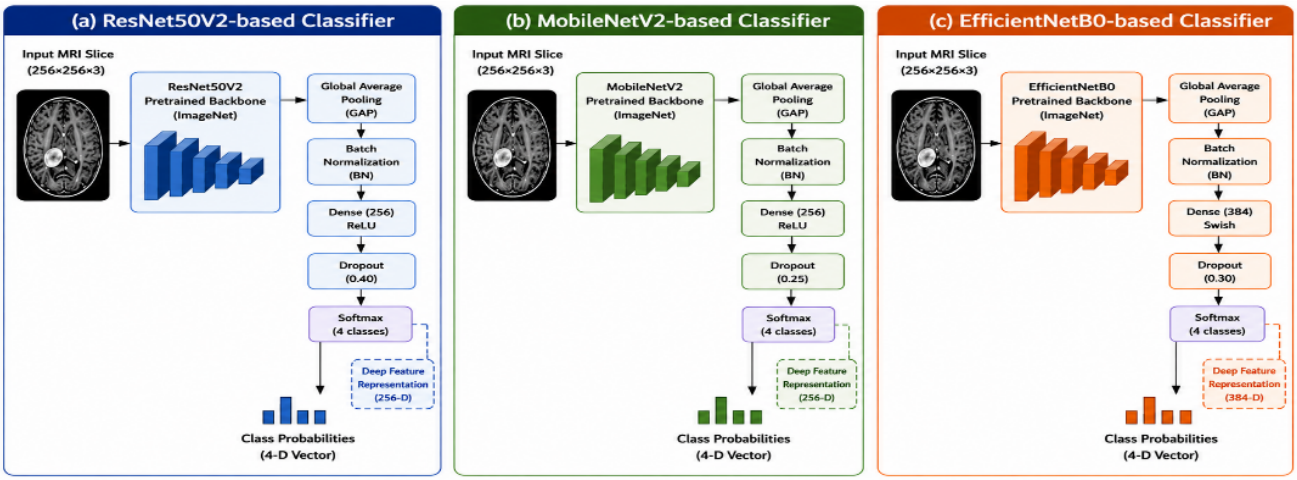


Figure 3. Customized transfer learning architectures used in the proposed framework: (**a**) ResNet50V2-based classifier, (**b**) MobileNetV2-based classifier, and (**c**) EfficientNetB0-based classifier.

3.5. Proposed Dual-Level Fusion Framework

The proposed DualXBrNet framework integrates complementary information extracted by ResNet50V2, MobileNetV2, and EfficientNetB0 through two successive fusion levels: learnable feature-level fusion and decision-level stacking. The first level combines the deep representations produced by the three heterogeneous CNN backbones, whereas the second level integrates the class probability outputs of the individual CNNs with those generated by the Feature Fusion classifier. A strict five-fold out-of-fold (OOF) strategy is used during framework development to generate out-of-sample meta-features for the final Logistic Regression meta-classifier, thereby avoiding training-to-meta-level leakage.

This dual-level design enables the framework to exploit complementary information both in the latent feature space and at the prediction level.

3.5.1. Feature-Level Fusion

At the first fusion level, deep feature representations extracted from the penultimate task-specific layers of the three CNN branches were combined. ResNet50V2 and MobileNetV2 each produced a 256-dimensional feature vector, whereas EfficientNetB0 generated a 384-dimensional representation. The three feature vectors were concatenated to form a unified 896-dimensional representation:

$$h_F = [h_R \| h_M \| h_E] \in R^{896} \quad (2)$$

where $h_R \in R^{256}$, $h_M \in R^{256}$, and $h_E \in R^{384}$ denote the deep feature representations extracted from ResNet50V2, MobileNetV2, and EfficientNetB0, respectively; ‖ denotes vector concatenation; and $h_F \in R^{896}$ represents the resulting concatenated feature vector.

The concatenated representation was processed by a learnable fusion network consisting of Batch Normalization, a 256-unit fully connected layer with ReLU activation, a dropout of 0.40, a 64-unit fully connected layer with ReLU activation, and a dropout of 0.25. A four-unit softmax output layer subsequently produced the feature fusion probability vector:

$$p_F = Softmax\left(f_{fusion(h_F)}\right) \in R^4 \quad (3)$$

where $P_F \in R^4$ denotes the four-class probability vector generated by the learnable Feature Fusion network, with each element representing the predicted probability of one diagnostic class.

Unlike fixed averaging or direct voting, the learnable fusion module is able to model nonlinear interactions among the heterogeneous deep representations before producing the final four-class Feature Fusion probability output. The architecture of this fusion stage is illustrated in Figure 4.

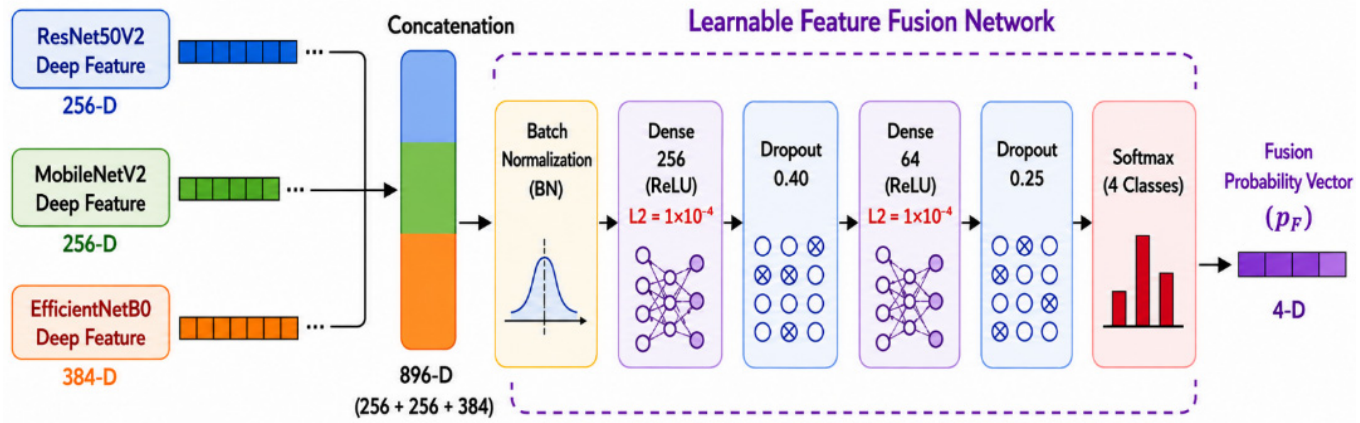


Figure 4. Architecture of the proposed learnable feature-level fusion module.

3.5.2. Decision-Level Stacking

At the second fusion level, the four-class probability vectors generated by the three individual CNN classifiers were combined with the probability vector produced by the Feature Fusion model.

The three individual CNN classifiers, ResNet50V2, MobileNetV2, and EfficientNetB0, provide three four-class probability vectors, yielding 12 probability values. These values are concatenated with the four probabilities generated by the Feature Fusion model to construct a 16-dimensional meta-feature vector:

$$z = [P_R \| P_M \| P_E \| P_F] \in \{R\}^{\{16\}} \quad (4)$$

where P_R, P_M, and P_E denote the four-class probability vectors generated by ResNet50V2, MobileNetV2, and EfficientNetB0, respectively; P_F denotes the Feature Fusion probability vector; $\|$ represents vector concatenation; and $z \in R^{16}$ denotes the resulting cross-level meta-feature vector used as input to the Logistic Regression meta-classifier.

The resulting vector combines decisions produced directly by the individual CNN classifiers with a decision derived from their jointly fused deep representations. This cross-level integration allows the meta-classifier to learn complementary prediction relationships rather than applying a fixed averaging rule.

A multinomial Logistic Regression model was used as the final meta-classifier. Logistic Regression was deliberately selected because the stacking input consists of only 16 structured OOF class probability features rather than high-dimensional deep representations. Its low model complexity and L2 regularization help limit overfitting at the meta-learning stage, while its linear formulation provides an interpretable mechanism for learning complementary contributions across the three CNN outputs and the Feature Fusion probabilities.

Importantly, the meta-classifier is trained exclusively on OOF predictions generated for samples excluded from the corresponding base model training folds, ensuring that the learned decision rule reflects the out-of-sample behavior of the constituent models. The multinomial Logistic Regression meta-classifier maps the 16-dimensional meta-feature vector to the final four-class probability output using its learned weight and bias parameters. The complete decision-level stacking mechanism is illustrated in Figure 5.

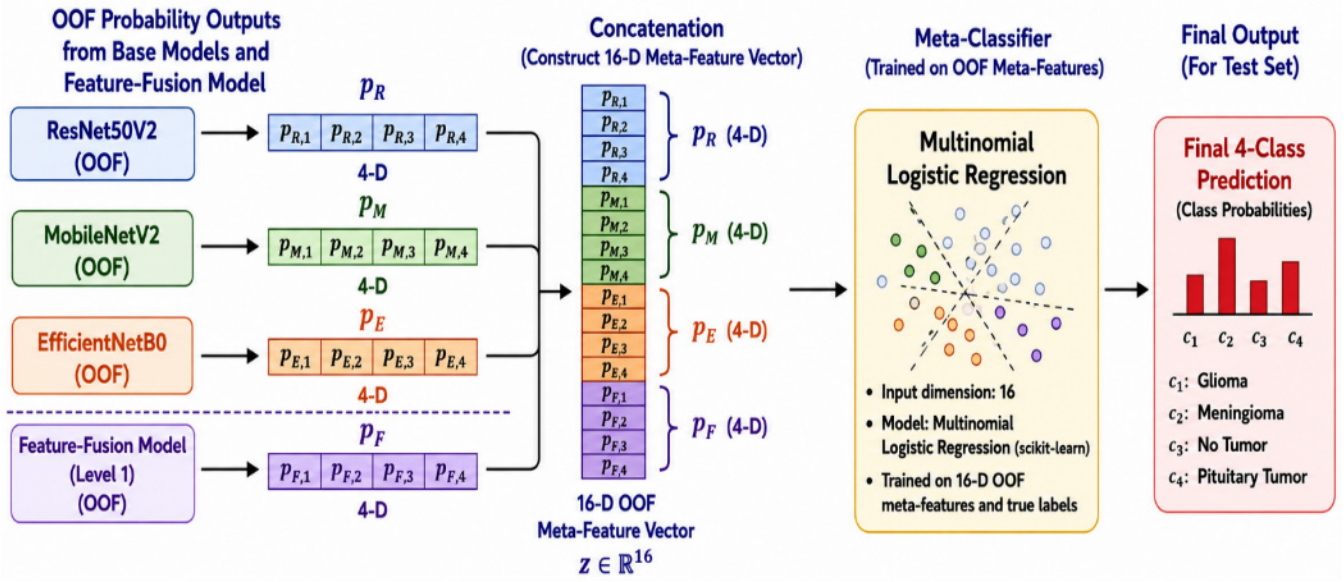


Figure 5. Decision-level stacking architecture using the 16-D OOF meta-feature vector and Logistic Regression meta-classifier.

3.5.3. Out-of-Fold Meta-Learning

The decision-level stacking stage was trained using the strict five-fold OOF protocol described in Section 3.3.

- For each held-out sample, ResNet50V2, MobileNetV2, and EfficientNetB0 generated four class probability vectors together with their deep feature representations. The corresponding fold-specific Feature Fusion model processed these representations and produced an additional four-class probability vector. These outputs were concatenated to form the 16-dimensional OOF meta-feature vector defined in Equation (4).
- The fold-specific OOF representations were concatenated to construct the complete meta-training matrix defined in Equation (1).
- The multinomial Logistic Regression meta-classifier was then trained exclusively on these out-of-sample OOF meta-features and their corresponding class labels.
- After meta-classifier training, the three CNN models and the Feature Fusion network were retrained on the complete 5000-image development set using the fixed schedules selected from the fold-validation stage. The independent 1000-image official test set remained completely isolated throughout model development. During final evaluation, the retrained models generated the same 16-dimensional representations for each test image, which were then passed to the fixed Logistic Regression meta-classifier to obtain the final four-class prediction.

3.6. Evaluation Metrics and Analysis Protocols

The proposed DualXBrNet framework was evaluated using complementary quantitative, statistical, computational, representational, and explainability analyses. Classification performance was assessed using accuracy, macro-precision, macro-recall, macro-F1-score, specificity, Matthews correlation coefficient (MCC), Cohen's kappa, and multiclass receiver operating characteristic area under the curve (ROC-AUC). Additional analyses were conducted to assess model complementarity, statistical reliability, computational complexity, feature space organization, and model interpretability.

3.6.1. Classification Performance Metrics

For each diagnostic class, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) were derived using a one-vs-rest formulation.

- Accuracy was calculated as the proportion of correctly classified samples among all evaluated samples [36].
- Precision was used to measure the proportion of positive predictions that were correctly classified.
- Recall, also referred to as sensitivity, was used to measure the proportion of actual positive samples that were correctly identified.
- Specificity was used to quantify the proportion of actual negative samples that were correctly identified.
- The F1-score was calculated as the harmonic mean of precision and recall, providing a balanced measure of both metrics.
- Macro-averaged metrics: For the four-class classification task, macro-averaged precision, recall, and F1-score were calculated by computing each metric separately for each class and then taking the unweighted mean across all four classes. This approach assigns equal importance to each diagnostic category regardless of its sample size.
- We also employed the Matthews correlation coefficient (MCC) as a balanced measure of the agreement between the predicted and true multiclass labels. The MCC considers the complete confusion matrix and provides values ranging from −1 to 1, where 1 indicates

perfect classification, 0 indicates prediction no better than random agreement, and −1 indicates complete disagreement.

- Cohen's kappa was used to quantify agreement between the predicted and true class labels after accounting for agreement expected by chance [37].
- Multiclass ROC-AUC was used to evaluate the ability of each model to discriminate among the four diagnostic categories across different classification thresholds. Macro-ROC-AUC was calculated by giving equal contribution to the class-wise ROC-AUC values.

3.6.2. Oracle and Pairwise Complementarity Analysis

Oracle analysis was used to quantify the complementary predictive behavior of ResNet50V2, MobileNetV2, and EfficientNetB0. The oracle does not represent a deployable classifier; instead, it provides a diagnostic upper bound by considering a sample correctly classified whenever at least one constituent CNN predicts its true class.

Oracle accuracy was calculated as the proportion of evaluated samples for which at least one of the three constituent CNNs predicted the correct class. Thus, a sample was counted as correctly classified by the oracle whenever ResNet50V2, MobileNetV2, or EfficientNetB0 produced the true class label.

Pairwise oracle accuracy was similarly calculated for each pair of CNNs to determine the maximum accuracy attainable when either member of the pair correctly classified a sample.

Model error diversity was further quantified using shared errors and Jaccard error similarity. For two models A and B, let E_A and E_B denote their respective sets of misclassified samples.

Jaccard error similarity was calculated as the number of samples misclassified by both models divided by the number of samples misclassified by at least one of them.

Lower Jaccard values indicate less overlap between the models' error sets and, therefore, greater error diversity.

Unique-correct predictions were also examined. A prediction was considered unique to one CNN when that CNN correctly classified a sample while the other two CNNs simultaneously misclassified it. This analysis was used to assess whether a lower-performing backbone could still contribute non-redundant predictive information to the ensemble.

3.6.3. Statistical Validation

Statistical reliability of the final classification performance was assessed using bootstrap confidence interval estimation and McNemar's exact test [38]. Bootstrap resampling of the independent official test set was employed to estimate the sampling variability of the final accuracy and its 95% confidence interval.

McNemar's exact test was used for the paired comparison between the Feature Fusion model and the final DualXBrNet framework. The test evaluates discordant predictions between two classifiers on the same samples. Let b denote the number of samples correctly classified only by the first model and c the number correctly classified only by the second model. The exact McNemar test evaluates whether the observed imbalance between b and c is statistically significant. A significance level of $\alpha < 0.05$ was adopted.

To examine the robustness of the meta-classifier choice, the original multinomial Logistic Regression model (L2 regularization, C = 0.1, the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) solver, and a maximum of 5000 iterations) was compared with a radial basis function support vector machine (RBF-SVM), a Random Forest containing 500 trees, and a one-hidden-layer multilayer perceptron (MLP) containing 32 units.

The inputs to the RBF-SVM and MLP were standardized, whereas Logistic Regression and Random Forest were fitted directly to the OOF probability features. All meta-classifiers were evaluated using the same 16-dimensional OOF representation and the five original held-out folds; the independent official test set was not used for meta-classifier selection. The fold-wise results were summarized using the mean and sample standard deviation. Overall differences among the four meta-classifiers were assessed using the Friedman test, followed by exploratory two-sided Wilcoxon signed-rank comparisons between Logistic Regression and each alternative. An additional ablation compared Logistic Regression trained on the 12 individual CNN probabilities with the proposed 16-dimensional cross-level representation using McNemar's exact test.

3.6.4. Computational Complexity Analysis

Computational complexity was evaluated to characterize the efficiency of the individual CNN backbones and the proposed fusion components. The analysis included total parameters, trainable and non-trainable parameters, estimated parameter memory, saved model size, and floating point operations (FLOPs) required for a forward pass.

Parameter memory was estimated from the number of stored parameters and their numerical precision. FLOPs were measured under the same 256 × 256 input resolution used during model evaluation. These measures were used to examine the trade-off between predictive performance and computational demand across the constituent models and fusion stages.

3.6.5. Feature Space Visualization

We investigated the organization and separability of the learned deep representations using t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) [39,40]. Both methods project high-dimensional feature representations into a two-dimensional space while preserving different aspects of the underlying feature structure.

t-SNE focuses primarily on local neighborhood relationships, whereas UMAP aims to preserve both local structure and broader global manifold organization.

The two techniques were applied to the deep representations learned by ResNet50V2, MobileNetV2, EfficientNetB0, and the learnable Feature Fusion pathway to visually compare class compactness and separation among the four diagnostic categories.

3.6.6. Occlusion Sensitivity Explainability Analysis

Occlusion Sensitivity was employed as a post hoc explainability method to identify image regions that contributed most strongly to the predicted class [41]. A localized region of the input MRI was systematically replaced by an occlusion patch, and the resulting change in the predicted probability of the target class was measured.

For an image x, predicted class y, and an occluded version $x^{(r)}$ in which region r is masked, the occlusion sensitivity score can be expressed as:

$$S(r) = p_{y(x)} - p_{y(x^r)} \quad (5)$$

where S(r) denotes the sensitivity associated with region r, $p_y(x)$ represents the predicted probability of class y for the original image, and $p_y(x^{(r)})$ represents the corresponding probability after occluding region r. A larger positive value of S(r) indicates that the occluded region had a stronger positive influence on the original prediction.

Operationally, Occlusion Sensitivity was computed using a 32 × 32 pixel occlusion patch moved across each 256 × 256 input image with a stride of 24 pixels in both spatial directions. The occluded region was replaced by the per-channel mean intensity calculated

from the corresponding input image, thereby avoiding the introduction of an artificial black region. For each occluded image, the decrease in the target class probability relative to the original unoccluded prediction was calculated. Overlapping contributions were averaged, negative probability changes were set to zero, and the resulting sensitivity map was normalized to the range $[0, 1]$. Occluded images were processed in batches of 32 for computational efficiency. For the no-tumor example, the unchanged image was displayed as a reference without a tumor-focused occlusion map.

The resulting sensitivity values were spatially visualized as heatmaps and overlays to provide qualitative insight into whether the constituent models and Feature Fusion pathway relied on diagnostically informative brain regions.

4. Experimental Results

The experimental evaluation was designed to assess the proposed DualXBrNet framework from multiple perspectives, including predictive performance, cross-fold stability, computational efficiency, model complementarity, statistical reliability, feature space separability, and explainability. The results are reported for the three individual CNN models, the learnable Feature Fusion model, and the final dual-level stacking framework. The independent official test set was used only for final evaluation, while the OOF predictions generated from the development set were used for meta-classifier training and cross-fold analysis.

4.1. Training and Optimization Strategy

All experiments used 256 × 256-pixel inputs, a batch size of 16, and a random seed of 42. Implementation was performed in Google Colab using Python 3.13.15, TensorFlow 2.20.0, Keras 3.13.2, scikit-learn 1.6.1, NumPy 2.1.3, pandas 2.2.3, and SciPy 1.16.3. The three CNNs were trained in two stages: optimization of the classification heads with frozen backbones, followed by fine-tuning of selected upper layers at lower learning rates. AdamW with a weight decay of 1×10^{-4} was used for the CNNs, whereas Feature Fusion used Adam; sparse categorical cross-entropy was used throughout.

Class imbalance was addressed using weights of 1.089799, 0.940557, 1.171509, and 0.857927 for glioma, meningioma, no-tumor, and pituitary tumor, respectively. Within each OOF fold, model selection relied exclusively on the inner validation subset. Median validation-selected epochs determined the final schedules: 14 + 4 epochs for ResNet50V2, 4 + 1 for MobileNetV2, 8 + 2 for EfficientNetB0, and six epochs for Feature Fusion.

During fine-tuning, the final ResNet50V2 residual stage was unfrozen with Batch Normalization layers kept frozen; the upper 20 and 24 layers of MobileNetV2 and EfficientNetB0 were unfrozen, respectively. After OOF generation, the deep models were retrained on all 5000 development images, and the 16-dimensional OOF representation trained the Logistic Regression meta-classifier. Table 4 summarizes these settings. Figure 6 shows that ResNet50V2 and EfficientNetB0 maintained relatively close training and validation trajectories, whereas MobileNetV2 exhibited greater validation fluctuations and a wider late-stage gap, indicating less stable generalization. Feature Fusion converged rapidly without marked late-epoch divergence. Logistic Regression is excluded because it is not trained through epoch-based optimization.

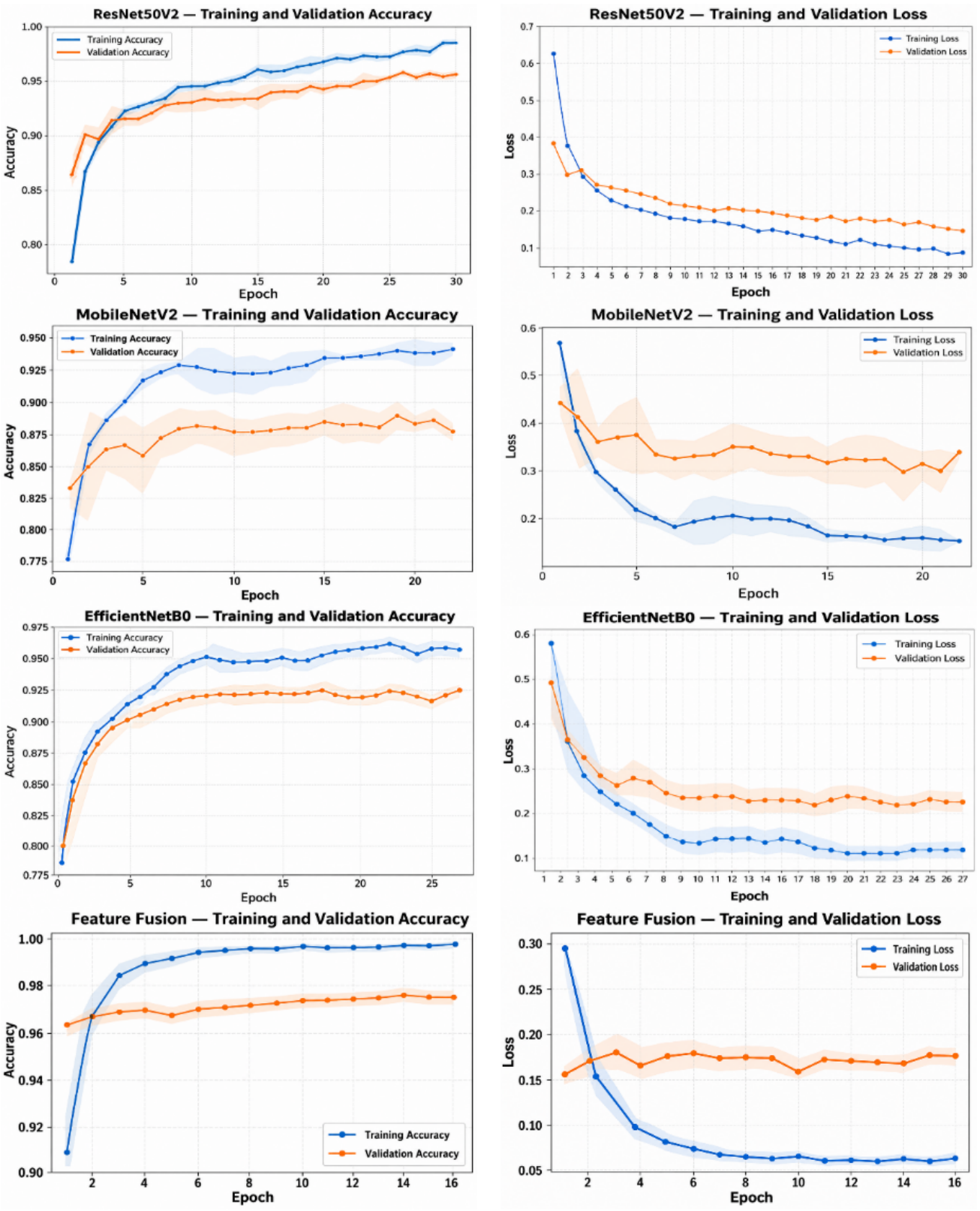


Figure 6. Training and validation accuracy and loss curves of ResNet50V2, MobileNetV2, EfficientNetB0, and the learnable Feature Fusion model.

Table 4. Training and optimization settings of the proposed framework.

Component	Stage 1 LR	Stage 2 LR	Optimizer	Weight Decay	Final Epochs	Fine-Tuning/Training Strategy
ResNet50V2	9×10^{-5}	2×10^{-6}	AdamW	1×10^{-4}	14 + 4	Final residual stage; BN layers frozen
MobileNetV2	9×10^{-5}	9×10^{-7}	AdamW	1×10^{-4}	4 + 1	Upper 20 backbone layers
EfficientNetB0	3×10^{-4}	1×10^{-6}	AdamW	1×10^{-4}	8 + 2	Upper 24 backbone layers
Feature Fusion	1×10^{-3}	N/A	Adam	N/A	6	Fusion MLP training
Meta-Classifier	N/A	N/A	Logistic Regression	L2, (C = 0.1)	N/A	LBFGS, max_iter = 5000

4.2. Five-Fold Out-of-Fold Performance

To evaluate the stability and generalization behavior of the constituent models before final testing, five-fold out-of-fold (OOF) performance was examined over the complete 5000-image development set. In each fold, predictions were generated exclusively for samples that were not used for training the corresponding fold-specific models. Table 5 summarizes the fold-wise classification accuracy together with the mean and standard deviation across the five folds.

Table 5. Five-fold OOF accuracy of the constituent models and evaluated meta-classifier configurations.

Model	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Mean Accuracy (%)	SD (%)
ResNet50V2	93.80	96.10	95.50	93.70	94.40	94.70	1.06
MobileNetV2	88.90	86.40	89.10	86.70	86.50	87.52	1.36
EfficientNetB0	91.70	91.60	91.10	91.00	90.10	91.10	0.64
Feature Fusion	97.30	97.20	96.60	96.40	96.20	96.74	0.49
LR (12-D CNN)	95.10	96.30	96.30	95.10	95.30	95.62	0.63
LR (16-D cross-level)	96.30	97.00	97.10	95.70	96.60	96.54	0.57
RBF-SVM (16-D cross-level)	96.80	97.40	97.00	96.60	96.40	96.84	0.38
RF (16-D cross-level)	96.80	97.40	97.10	96.20	96.30	96.76	0.51
MLP (16-D cross-level)	96.30	97.10	96.50	95.70	96.30	96.38	0.50

Among the individual CNN models, ResNet50V2 achieved the highest mean OOF accuracy of 94.70%, followed by EfficientNetB0 with 91.10% and MobileNetV2 with 87.52%. In the five folds, the learnable Feature Fusion model consistently outperformed all three individual CNNs, with a mean OOF accuracy of 96.74% and a standard deviation of only 0.49%. The Feature Fusion stage outperformed the strongest individual CNN, ResNet50V2, by an average of 2.04 percentage points and also had lower fold-to-fold variability.

The expanded results in Table 5 show that the RBF-SVM obtained the highest mean meta-classifier accuracy ($96.84 \pm 0.38\%$), followed by Random Forest ($96.76 \pm 0.51\%$), Logistic Regression ($96.54 \pm 0.57\%$), and MLP ($96.38 \pm 0.50\%$). However, the overall Friedman test did not indicate a statistically significant difference among the four 16-dimensional meta-classifiers ($p = 0.1083$). Consistently, the paired Wilcoxon comparisons between Logistic Regression and RBF-SVM, Random Forest, and MLP were not statistically significant ($p = 0.3125$, 0.2500, and 0.5000, respectively). Logistic Regression was, therefore, retained as the prespecified low-complexity and interpretable meta-learner rather than selecting a more complex alternative on the basis of a small fold-wise numerical difference. The meta-feature ablation further showed that adding the four Feature Fusion probabilities to the 12 individual CNN probabilities increased the mean OOF accuracy of Logistic Regression from $95.62 \pm 0.63\%$ to $96.54 \pm 0.57\%$. The paired OOF comparison included 26 samples correctly classified only by the 12-dimensional configuration and 72 samples correctly classified only by the 16-dimensional configuration. McNemar's exact test confirmed that this difference was statistically significant ($p = 0.000004$).

4.3. Final Official Test Performance

The final performance of the three individual CNN models, the learnable Feature Fusion model, and the proposed DualXBrNet framework was evaluated on the independent official test set of 1000 MRI images. In Table 6, the proposed dual-level framework achieved the highest overall classification performance with an accuracy of 96.80%, a macro-precision of 97.19%, a macro-recall of 97.03%, and a macro-F1-score of 97.08%. It also obtained the highest MCC (0.9565), Cohen's kappa (0.9563), and macro-ROC-AUC (99.77%).

Table 6. Performance comparison on the independent official test set.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MCC	Kappa	ROC-AUC (%)
ResNet50V2	94.20	95.02	94.58	94.75	0.9211	0.9207	99.49
MobileNetV2	86.00	87.75	87.06	86.99	0.8124	0.8085	97.83
EfficientNetB0	93.20	94.00	94.10	93.90	0.9087	0.9072	99.49
Feature Fusion	96.50	96.81	96.76	96.74	0.9525	0.9522	99.76
DualXBrNet (Dual-Level)	96.80	97.19	97.03	97.08	0.9565	0.9563	99.77

The comparative trends of the primary classification metrics and the secondary agreement and discrimination metrics are further illustrated in Figures 7 and 8, respectively.

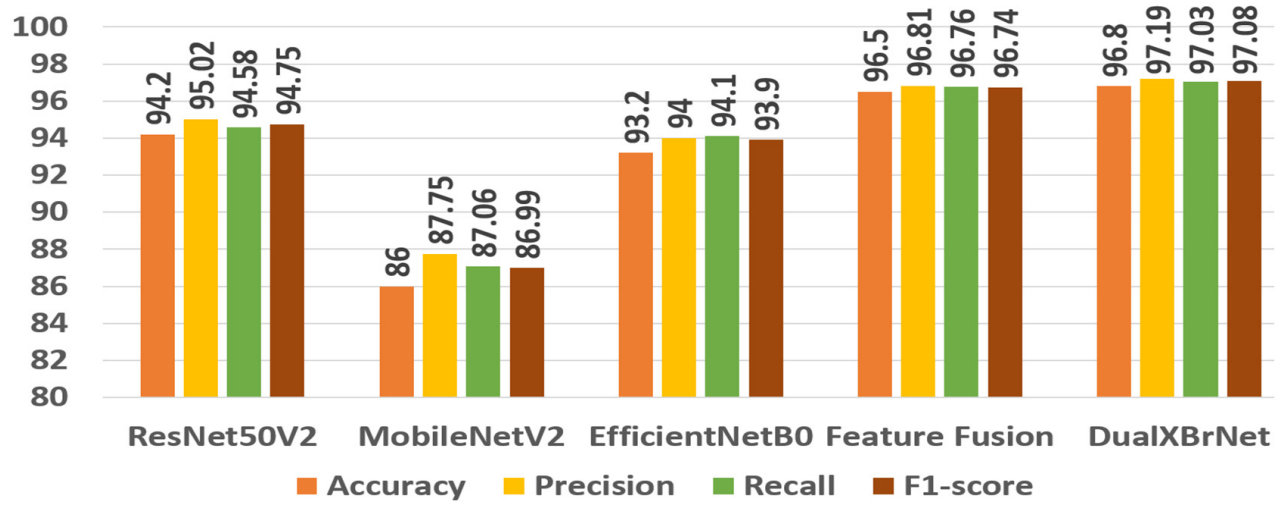


Figure 7. Comparison of accuracy, precision, recall, and F1-score across the evaluated models.

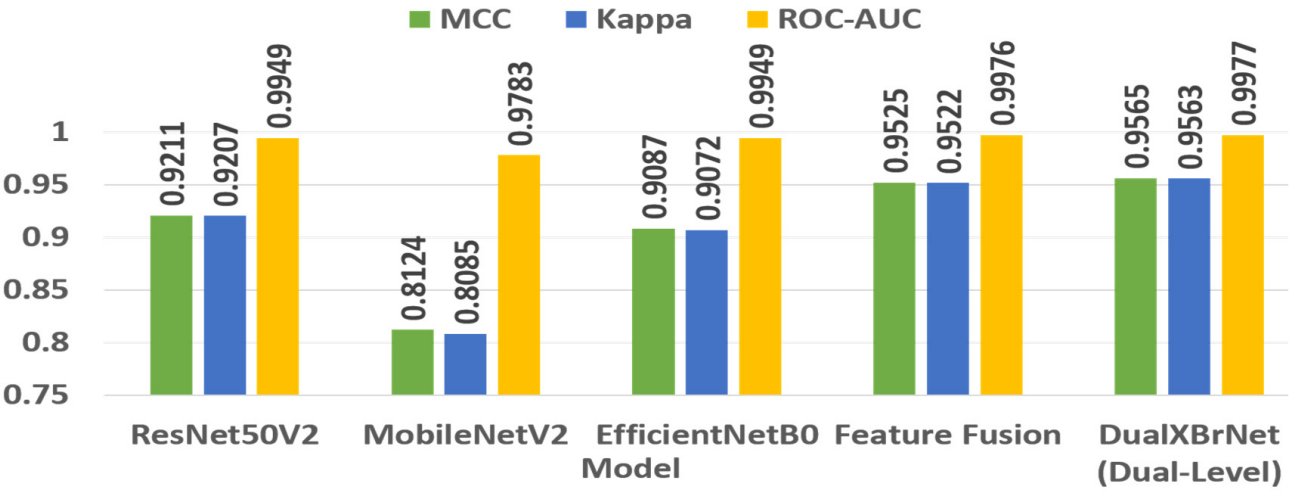


Figure 8. Comparison of MCC, Cohen's kappa, and ROC-AUC across the evaluated models.

Among the individual CNNs, ResNet50V2 achieved the strongest test accuracy at 94.20%, followed by EfficientNetB0 at 93.20%, while MobileNetV2 achieved 86.00%. As shown in Table 6 and Figure 7, the learnable Feature Fusion model increased the test accuracy to 96.50%, providing the primary performance gain of 2.30 percentage points over the strongest individual CNN, ResNet50V2. The subsequent OOF-based decision-level stacking stage further refined the fused predictions, increasing the accuracy to 96.80%. Accordingly, feature-level fusion provided the principal representational improvement, while decision-level stacking contributed an additional complementary gain of 0.30 percentage points.

This behavior is also reflected in Figure 8, where DualXBrNet achieved the highest MCC, Cohen's kappa, and ROC-AUC among all evaluated models.

4.4. Class-Wise Performance

A class-wise analysis was conducted to examine the diagnostic behavior of the final DualXBrNet model across the four tumor categories. As summarized in Table 7, the proposed model achieved strong performance for all classes. The no-tumor class obtained perfect sensitivity of 100.00% and an ROC-AUC of 100.00%, while the pituitary class also achieved 100.00% sensitivity with an F1-score of 98.85%. Meningioma achieved a sensitivity of 96.41% and an F1-score of 95.32%, whereas glioma showed the lowest sensitivity among the four classes at 91.73%, despite maintaining high precision (97.49%) and specificity (99.20%).

Table 7. Class-wise performance of the final DualXBrNet model on the independent official test set.

Class	Precision (%)	Sensitivity/Recall (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Glioma	97.49	91.73	99.20	94.52	0.9963
Meningioma	94.25	96.41	97.41	95.32	0.9948
No Tumor	99.29	100.00	99.88	99.64	1.0000
Pituitary	97.72	100.00	99.00	98.85	0.9998

The corresponding confusion matrix further confirms that the majority of errors occurred between glioma and meningioma, while the no-tumor and pituitary classes were classified without false negative errors. Overall, the class-wise results demonstrate that the proposed dual-level framework maintained a high level of discrimination across all four diagnostic categories.

4.5. Confusion Matrix and ROC Analysis

Figure 9 illustrates the confusion matrices of the five evaluated models, which allows us to directly compare their class-wise prediction patterns on the independent official test set. Among the individual CNN models, ResNet50V2 and EfficientNetB0 showed a better classification behavior in general than MobileNetV2, which showed more misclassifications comparatively. The learnable Feature Fusion model further enhanced the prediction consistency by combining complementary deep representations from the three backbone networks. The final DualXBrNet model showed the most balanced overall confusion pattern among the four diagnostic categories.

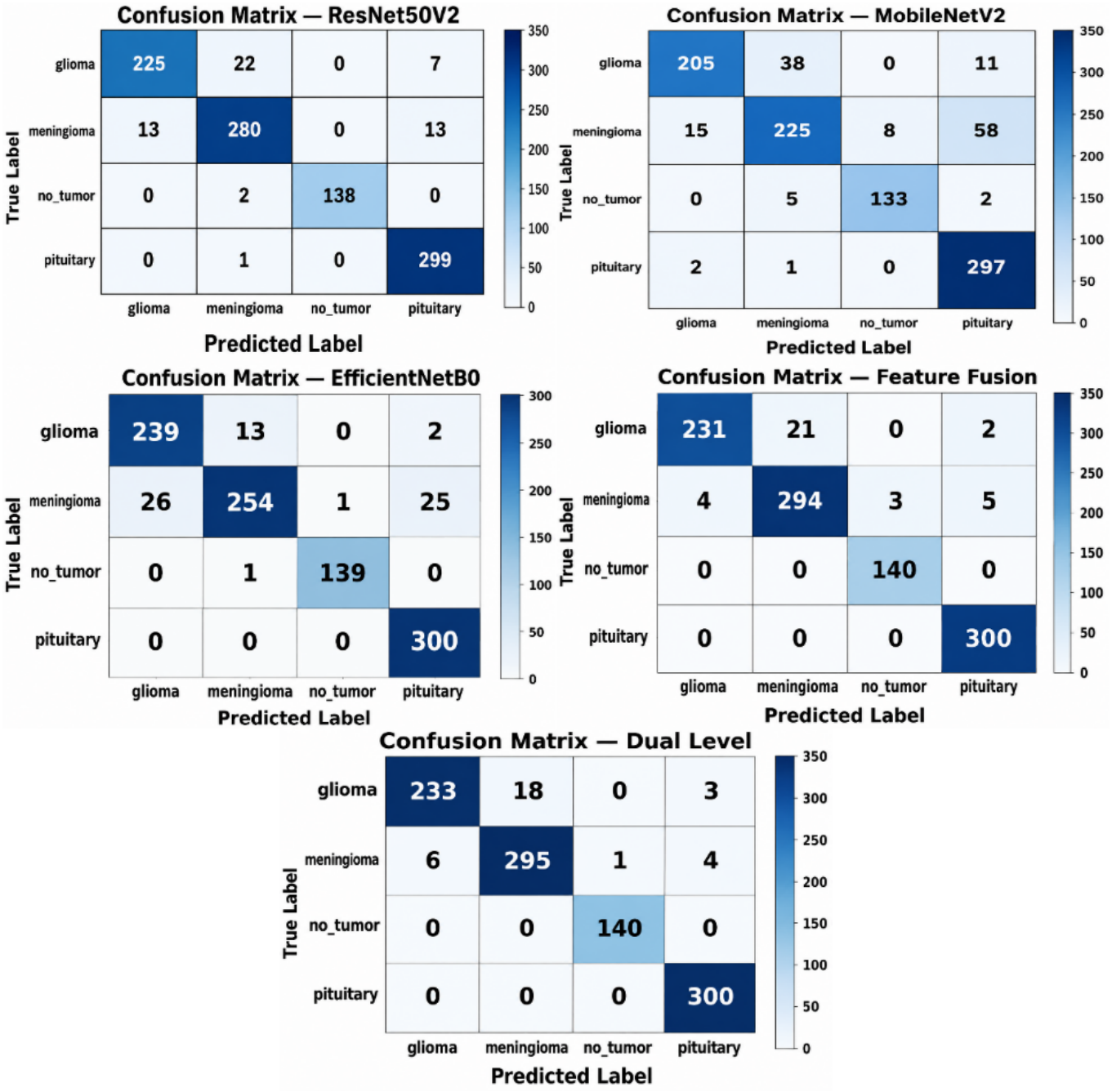


Figure 9. Confusion matrices of the five evaluated models on the independent official test set.

Corresponding receiver operating characteristic (ROC) curves are separately shown in Figure 10. All evaluated models exhibited good discriminative power, with fusion-based approaches demonstrating the best overall class separation. In agreement with the quantitative outcomes presented in Table 6, DualXBrNet achieves the highest macro-ROC-AUC among all the models evaluated. The overall results of the confusion matrix and ROC analyses validate the superiority of the combination of feature-level fusion and OOF-based decision-level stacking.

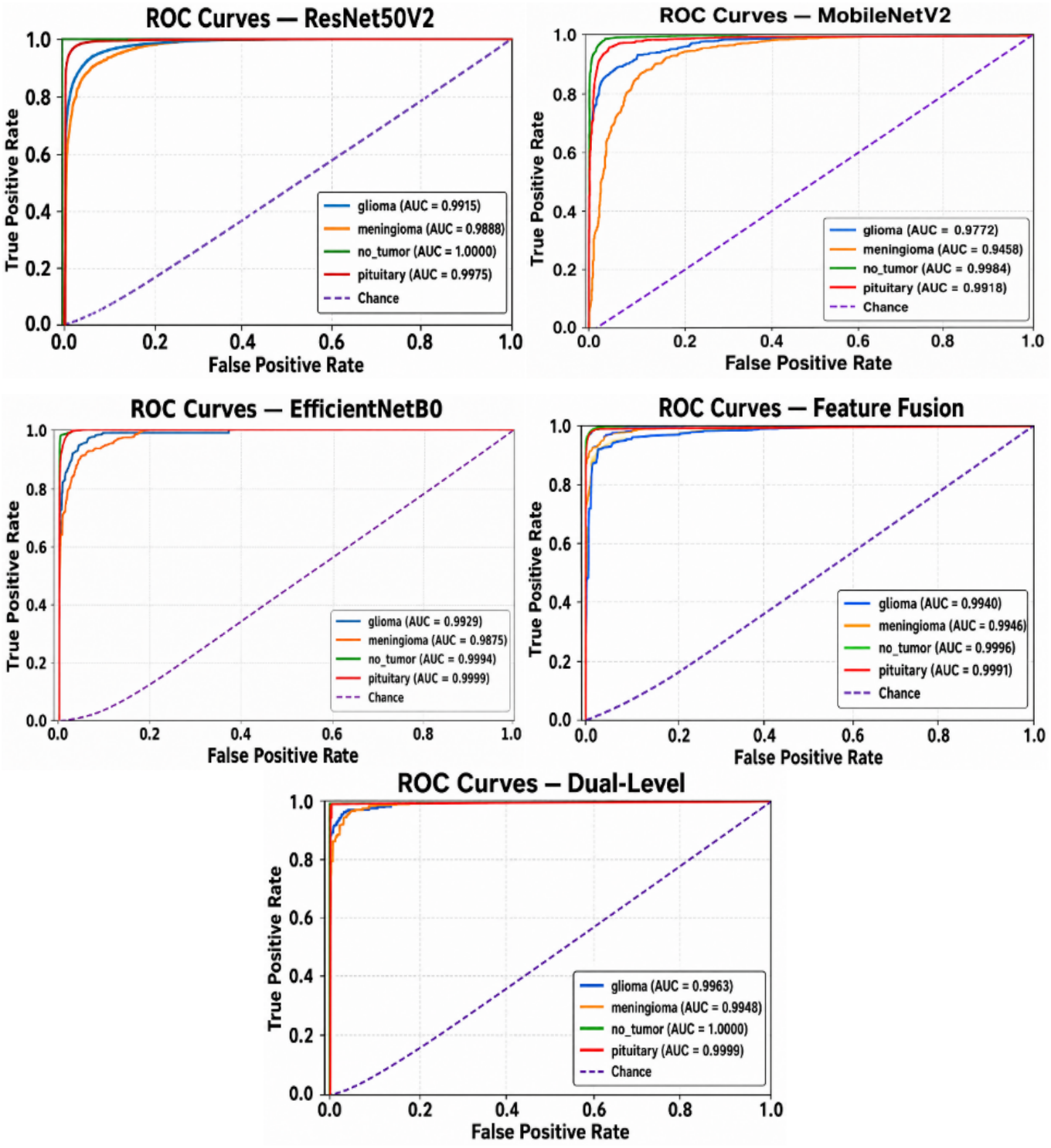


Figure 10. ROC curves of the five evaluated models on the independent official test set.

4.6. Oracle Complementarity Analysis

Under the strict five-fold OOF protocol, the three-CNN oracle accuracy reached 97.88%. Among the 5000 development samples, all three CNNs correctly classified 4132 cases, exactly two models were correct for 508 cases, and exactly one model was correct for 254 cases, while all three models failed simultaneously on only 106 samples. The uniquely correct predictions were distributed across all three backbones: 157 cases were correctly identified only by ResNet50V2, 28 only by MobileNetV2, and 69 only by EfficientNetB0. These results confirm that the prediction errors of the three CNNs were not fully overlapping.

A similar pattern was observed on the independent official test set. The individual CNN accuracies ranged from 86.00% for MobileNetV2 to 94.20% for ResNet50V2, with EfficientNetB0 achieving 93.20%. The oracle accuracy reached 98.40%, which is 4.20 percentage points higher than the strongest individual CNN. Specifically, 984 of the 1000 test images were correctly classified by at least one CNN, while only 16 were misclassified by all three models. Among the uniquely correct predictions, ResNet50V2 contributed 24 cases, EfficientNetB0 contributed 22, and MobileNetV2 contributed five. The overall oracle results for the OOF development set and the independent official test set are summarized in Table 8 and visually illustrated in Figure 11.

Table 8. Oracle complementarity analysis under the OOF and official test protocols.

Metric	OOF (5000 Images)	Official Test (1000 Images)
Oracle Accuracy (%)	97.88%	98.40%
All Three Correct	4132	817
Exactly Two Correct	508	116
Exactly One Correct	254	51
All Three Wrong	106	16
ResNet50V2 Only Correct	157	24
MobileNetV2 Only Correct	28	5
EfficientNetB0 Only Correct	69	22

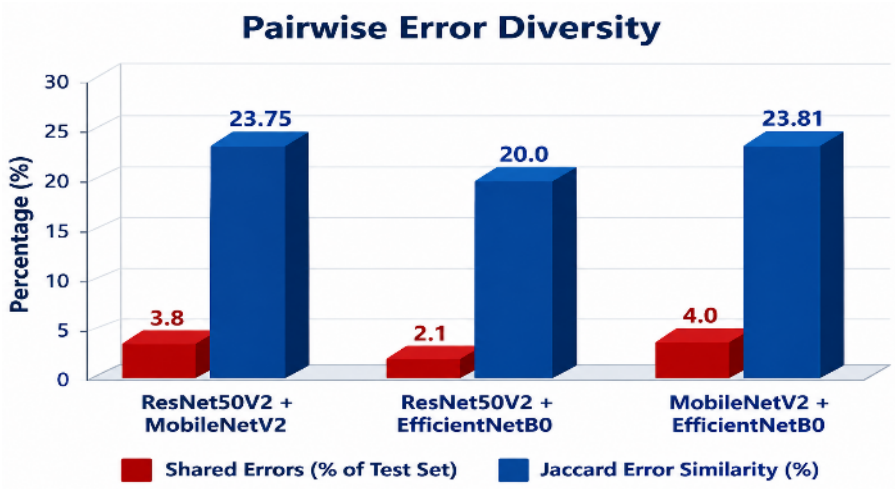


Figure 11. Oracle complementarity profile under the OOF development protocol and the independent official test set.

Despite its lower standalone accuracy, MobileNetV2 exhibited measurable complementary predictive behavior. When paired with ResNet50V2, the oracle accuracy increased from 94.20% to 96.20%, while the ResNet50V2–MobileNetV2 error overlap was 38 samples, with a Jaccard error similarity of 0.2375. Similarly, the MobileNetV2 EfficientNetB0 pair achieved an oracle accuracy of 96.00%, with 40 shared errors and a Jaccard similarity of 0.2381. These results indicate only partial overlap between the corresponding error sets and support the inclusion of MobileNetV2 as a complementary backbone despite its lower standalone accuracy. The pairwise complementarity results among the three CNN backbones are visualized in Figure 12.

The proposed Feature Fusion model increased the test accuracy to 96.50%, and the final DualXBrNet framework further improved it to 96.80%. Relative to the 98.40% oracle upper bound, the remaining oracle gap was 1.60 percentage points, indicating that DualXBrNet captures a substantial portion of the complementary predictive information available across the constituent CNNs, while some complementary predictive potential remains unexploited.

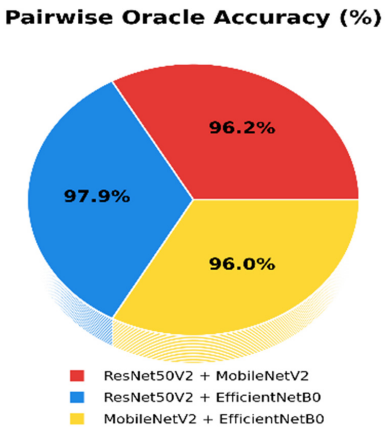


Figure 12. Pairwise error diversity of the three CNN backbones on the independent official test set, expressed in terms of shared errors and Jaccard error similarity.

4.7. Computational Complexity Analysis

The computational complexity of the proposed framework and its constituent models was evaluated in terms of the total number of parameters, trainable parameters, parameter memory, saved model size, and floating point operations (FLOPs), as summarized in Table 9. Among the three backbone CNNs, ResNet50V2 exhibited the highest computational demand, with approximately 24.10 million parameters, 14.96 million trainable parameters, and 9.13 GFLOPs. In contrast, MobileNetV2 was the most lightweight backbone, requiring only 2.59 million parameters and approximately 0.80 GFLOPs, while EfficientNetB0 provided an intermediate computational profile with 4.55 million parameters and 1.05 GFLOPs.

Table 9. Computational complexity of the backbone and fusion components.

Model	Total Parameters	Trainable Parameters	Parameter Memory (MB)	Saved Model Size (MB)	GFLOPs
ResNet50V2	24,098,564	14,955,780	91.93	206.70	9.1291
MobileNetV2	2,586,948	1,839,684	9.87	24.48	0.8013
EfficientNetB0	4,548,135	1,979,364	17.35	33.28	1.0465
Feature Fusion	249,924	248,132	0.95	2.88	0.0005
Meta-Logistic Regression	68	68	0.0005	0.0013	0.00000013

The learnable Feature Fusion module introduced only a modest additional overhead, containing approximately 0.25 million parameters and requiring less than 0.001 GFLOPs. Similarly, the decision-level meta-classifier was computationally negligible, consisting of only 68 trainable parameters and 132 FLOPs. These findings indicate that the major computational cost of the proposed framework is associated with the pretrained CNN backbones, whereas the additional feature and decision-level fusion stages contribute minimal complexity. Therefore, the improved diagnostic performance of DualXBrNet is achieved without introducing a substantial computational burden beyond that of the underlying backbone networks.

4.8. Additional Evaluation on the Masoud Dataset

The finalized BRISC2025 trained models were additionally evaluated on all 7200 images of the Masoud dataset without retraining, fine-tuning, or parameter modification. As summarized in Table 10, DualXBrNet achieved an accuracy of 94.94%, a macro-F1-score of 94.96%, and a macro-AUC of 99.41%, demonstrating that the framework maintained strong classification performance when applied to data obtained from another public source. Feature Fusion achieved the highest accuracy of 95.35% and a macro-F1-score of 95.37%, whereas DualXBrNet achieved the highest macro-AUC of 99.41%.

Table 10. Performance of the finalized BRISC2025-trained models on the complete Masoud dataset without retraining or fine-tuning.

Model	Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Weighted F1 (%)	Macro-AUC (%)
ResNet50V2	90.56	91.39	90.56	90.47	90.47	98.72
MobileNetV2	83.24	84.57	83.24	83.14	83.14	97.50
EfficientNetB0	88.33	89.00	88.33	88.22	88.22	98.53
Feature Fusion	95.35	95.61	95.35	95.37	95.37	99.34
DualXBrNet	94.94	95.27	94.94	94.96	94.96	99.41

On the original 1600 image testing partition of the Masoud dataset, both Feature Fusion and DualXBrNet achieved an accuracy of 89.69%. Their macro-F1-scores were 89.64% and 89.67%, respectively, while DualXBrNet achieved a macro-AUC of 98.03%. The paired prediction analysis showed that Feature Fusion exclusively corrected 15 samples missed by DualXBrNet, while DualXBrNet exclusively corrected 15 samples missed by Feature Fusion. Accordingly, McNemar's exact test indicated no statistically significant difference between the two models ($p = 1.000$).

Because a post hoc overlap audit identified shared images between the BRISC2025 and Masoud collections, these results are interpreted as a supplementary cross-dataset consistency assessment rather than fully independent external validation.

4.9. Statistical Validation

To assess the reliability of the final DualXBrNet performance and determine whether its improvement over the Feature Fusion model was statistically meaningful, bootstrap confidence interval estimation and McNemar's test were applied to the independent official test set. The final dual-level model achieved an observed accuracy of 96.80%, with a bootstrap mean accuracy of 96.81% and a 95% confidence interval of 95.70–97.80%. The corresponding bootstrap standard deviation was 0.55 percentage points, indicating limited variability in the estimated test accuracy.

As reported in Table 11, the paired comparison between the Feature Fusion and final dual-level models showed that three cases were correctly classified only by the Feature Fusion model, whereas six cases were correctly classified only by the dual-level model. Although the final dual-level framework improved accuracy from 96.50% to 96.80%, corresponding to a numerical gain of 0.30 percentage points, the exact McNemar test produced a p-value of 0.5078. Therefore, the observed improvement was numerical but not statistically significant at the conventional 0.05 significance level. This result indicates that the final decision-level stacking stage provided a modest performance gain over feature-level fusion, but the difference was not sufficiently large to establish statistical superiority on the current 1000-image test set.

Table 11. Statistical validation of the final DualXBrNet model.

Statistical Measure	Result
Observed accuracy (%)	96.80
Bootstrap mean accuracy (%)	96.81
Bootstrap 95% CI (%)	95.70–97.80
Bootstrap SD (%)	0.55
Feature Fusion only correct	3
Dual-level only correct	6
Accuracy difference (percentage points)	+0.30
McNemar exact p-value	0.5078
Statistical significance	Not significant

4.10. Feature Space Visualization

To examine the discriminative structure of the learned representations, t-SNE and UMAP were used to visualize the feature spaces of the three CNN backbones and Feature Fusion. The resulting two-dimensional projections are presented side by side in Figure 13.

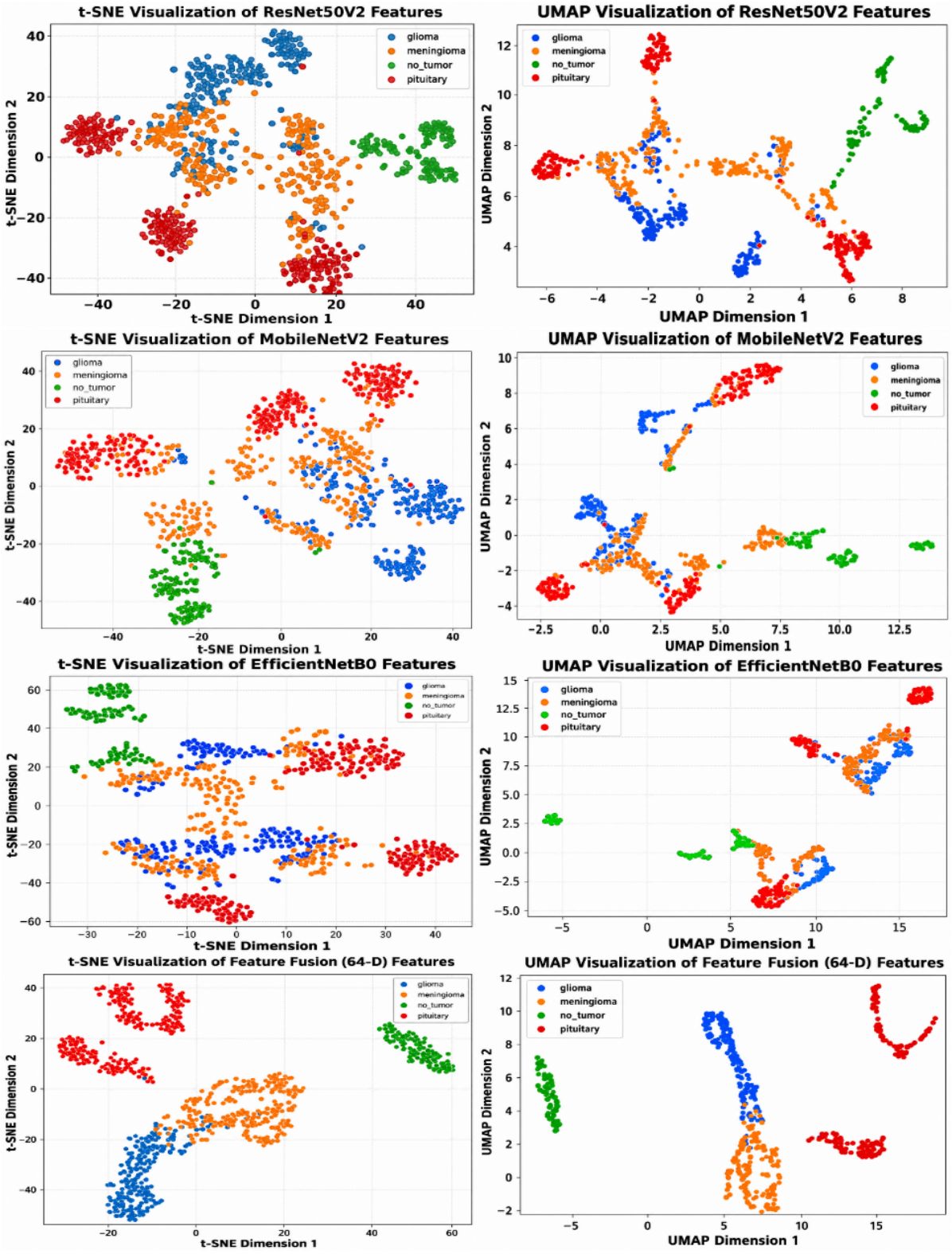


Figure 13. t-SNE and UMAP visualizations of ResNet50V2, MobileNetV2, EfficientNetB0, and the learnable Feature Fusion representation on the independent official test set.

The backbones exhibited different levels of intraclass compactness and interclass separation. ResNet50V2 and EfficientNetB0 showed more structured class distributions than MobileNetV2, whereas Feature Fusion produced tighter clusters and clearer class separation. These observations are consistent with the quantitative results and indicate that combining complementary deep representations improves feature space discrimination.

4.11. Explainability Analysis

The Occlusion Sensitivity visualizations for representative samples from the four diagnostic classes are presented in Figure 14. For the tumor classes, occluding regions located within or near the visible lesion generally produced the largest reduction in the predicted class probability, resulting in concentrated high-sensitivity regions around the abnormal tissue. This behavior was particularly evident in the glioma, meningioma, and pituitary examples, although the spatial extent of the response varied according to the size, appearance, and location of the lesion. In contrast, the no-tumor example produced a less localized sensitivity pattern because the corresponding decision was based on the absence of a focal abnormality rather than the presence of a single discriminative lesion.

These observations suggest that the predictions were influenced primarily by diagnostically relevant intracranial regions rather than by image borders or unrelated background content. Nevertheless, Occlusion Sensitivity provides a qualitative, model-dependent explanation and should not be interpreted as a precise tumor segmentation result or clinical localization map. The visualizations, therefore, complement the quantitative findings by illustrating the spatial evidence associated with individual predictions while also highlighting the need for broader validation of explanation consistency across additional samples.

4.12. Comparison with Recent Studies

To contextualize the performance and methodological contribution of DualXBrNet, Table 12 compares the proposed framework with representative brain tumor MRI studies published between 2024 and 2026, including Feature Fusion, ensemble, stacking, meta-learning, and explainable classification approaches. Particular attention is given to how predictive information is combined and whether out-of-fold (OOF) predictions are used for meta-learning. Because the compared studies employ different datasets, preprocessing procedures, data partitions, and evaluation protocols, their reported accuracies should be interpreted as contextual benchmarks rather than direct performance comparisons. For DualXBrNet, the 96.80% accuracy corresponds to the BRISC2025 official test set, whereas the 94.94% result was obtained on the complete Masoud dataset without retraining. Since shared images were identified between the BRISC2025 and Masoud collections, this result is considered a supplementary cross-dataset consistency assessment rather than fully independent external validation.

The closest methodological comparison is the approach of Uthra and Sivashankari [17], which also constructs a 16-dimensional OOF meta-representation. Their representation concatenates probability outputs from four independent base architectures, whereas DualXBrNet combines the 12 probabilities generated by three heterogeneous CNNs with four additional probabilities produced through learnable fusion of their latent representations. Thus, the proposed representation integrates architecture-specific decisions with a fusion-derived prediction, and its incremental contribution is evaluated through the controlled comparison between 12-dimensional CNN-only stacking and the proposed 16-dimensional representation.

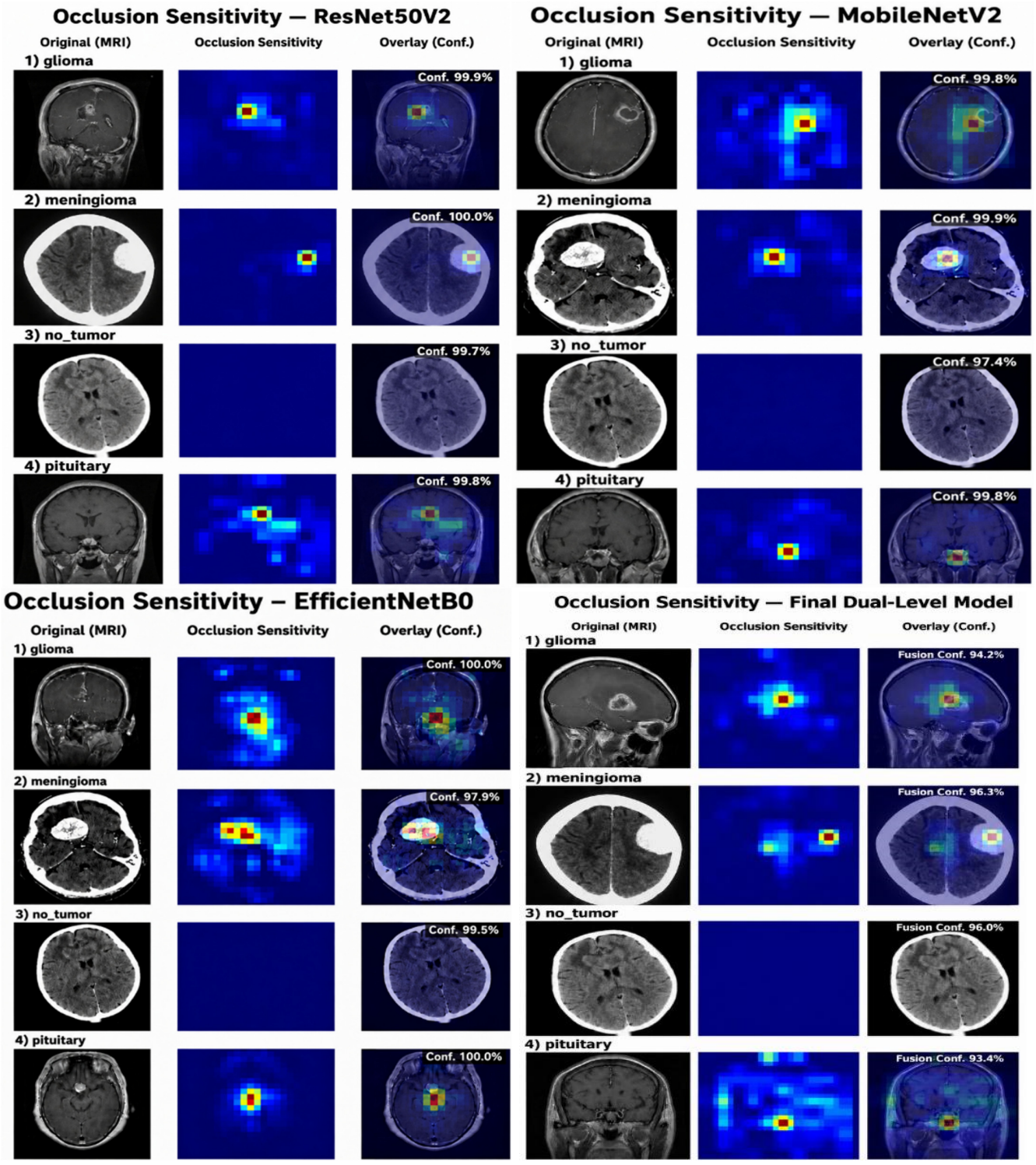


Figure 14. Occlusion Sensitivity maps and overlays for representative MRI samples from the glioma, meningioma, no-tumor, and pituitary classes.

Table 12. Performance comparison of DualXBrNet with related studies.

Study	Year	Fusion/Ensemble Strategy	Meta-Learning and Validation Design	Dataset/ Setting	Accuracy (%)
Almnawer et al. [42]	2026	CrossTaskFormer With cross-task attention	Multi-task ensemble with test time augmentation; no OOF stacking	BRISC four-class MRI dataset	99.30
Alkharaan et al. [43]	2026	EfficientNet-B1 with U-Net++	Transfer learning; no OOF stacking	BRISC2025 classification and segmentation dataset	99.70
Bouguerra et al. [44]	2025	FALCON fusion of custom CNNs and EfficientNet-B0–B7	65%/17%/18% train–validation–test split; no OOF stacking	Masoud dataset (7023 images)	100.00
Chen et al. [18]	2024	Feature summation fusion of three CNNs	Direct fused feature classification; no meta-learner	Figshare and Kaggle datasets	99.18/97.24
Haque et al. [14]	2025	Heterogeneous stacking of four deep models	CatBoost meta-learner with complementary analyses	Two four-class datasets from four public sources	98.09 ± 0.36/ 98.99 ± 0.41
Uthra & Sivashankari [17]	2026	Probability stacking of four architectures	Five-fold OOF 16-D representation with LightGBM	Masoud dataset; 5600/1600 split	95.50
Eone Etoua et al. [21]	2026	Macro-F1-weighted fusion of LBP and three CNNs	Five-fold CV with internal weighting; no OOF stacking	Masoud four-class MRI dataset	97.93 ± 0.18
This Study DualXBrNet	2026	Three-CNN feature fusion with cross-level integration	Five-fold OOF stacking; 12-D vs. 16-D ablation	BRISC2025 test set; Masoud consistency assessment	96.80/94.94

5. Discussion

The results indicate that the primary benefit of DualXBrNet arises from learnable feature-level fusion. By combining heterogeneous representations from ResNet50V2, MobileNetV2, and EfficientNetB0, the Feature Fusion network learns complementary interactions before generating class probabilities. Decision-level stacking provides a smaller additional refinement because its inputs already summarize the information learned by the individual CNNs and their fused representation.

5.1. Predictive Complementarity of the CNN Backbones

Oracle accuracy reached 97.88% under the strict OOF protocol and 98.40% on the independent official test set, confirming that the three CNNs did not consistently fail on the same samples. Pairwise and unique-correct analyses further showed that MobileNetV2 contributed non-redundant predictions despite its lower standalone accuracy. The results are in favor of using heterogeneous backbones instead of only relying on the strongest individual CNN.

5.2. Statistical Significance of the Decision-Level Refinement

DualXBrNet increased Feature Fusion accuracy from 96.50% to 96.80%. However, this difference was not statistically significant according to McNemar's exact test ($p = 0.5078$).

The stacking stage should, therefore, be interpreted as a modest refinement rather than a statistically confirmed improvement. Nevertheless, the strict OOF procedure ensured that the meta-classifier was trained on out-of-sample predictions, thereby avoiding training to meta-level leakage.

5.3. Class-Wise Performance and Challenging Classes

Strong discrimination was achieved across the four diagnostic categories, although glioma remained the most challenging class, with a sensitivity of 91.73%. Most residual errors involved confusion between glioma and meningioma. Although class weighting addressed class imbalance, it could not resolve similarities in intensity, texture, shape, and anatomical appearance. Future work could investigate focal or class-balanced loss, targeted augmentation, hard-example mining, and region-aware segmentation to improve performance on visually ambiguous cases.

5.4. Explainability and Feature Space Organization

The t-SNE and UMAP visualisations showed better class organization after feature fusion, indicating that the fused representation captured more discriminative structure. Occlusion Sensitivity maps further showed that predictions were generally influenced by informative intracranial regions, with tumor class responses concentrated near visible lesions and a more diffuse pattern for the no-tumor example. However, these maps provide qualitative model-dependent explanations and should not be interpreted as precise tumor-segmentation or clinical localization results.

5.5. Meta-Classifier Selection and the 16-Dimensional Representation

None of the evaluated nonlinear or tree-based meta-classifiers produced a statistically significant fold-wise improvement over Logistic Regression. This supports retaining Logistic Regression as a parsimonious and interpretable second-level model.

In contrast, adding the four Feature Fusion probabilities to the 12 individual CNN probabilities increased mean OOF accuracy from 95.62% to 96.54%. McNemar's exact test confirmed that this improvement was statistically significant ($p = 0.000004$). Therefore, the principal contribution of the second-level design lies in enriching the meta-representation with fusion-derived information rather than increasing meta-classifier complexity.

5.6. Computational Considerations

The fusion components introduced limited additional complexity compared with the backbone CNNs. The Feature Fusion network contained approximately 0.25 million parameters, while the Logistic Regression meta-classifier contained only 68 trainable parameters. Thus, the main computational cost remained associated with the pretrained CNN backbones, and cross-level integration was achieved without introducing a computationally expensive second-level architecture.

5.7. Cross-Dataset Evaluation and Study Limitations

DualXBrNet achieved 94.94% accuracy on the complete Masoud dataset without retraining or fine-tuning. The reduction from 96.80% on the BRISC2025 official test set suggests some sensitivity to dataset distribution. Moreover, a post hoc overlap audit identified shared images between the BRISC2025 and Masoud collections. This experiment is, therefore, interpreted as a supplementary cross-dataset consistency assessment rather than fully independent external validation.

Other limitations include the absence of patient-level metadata, evaluation on a single primary benchmark, the non-significant improvement of stacking over Feature Fusion, and

the qualitative nature of the explainability analysis. Future work should evaluate institutionally independent cohorts with confirmed patient-level separation and investigate calibration, uncertainty estimation, alternative cross-level fusion strategies, and complementary explainability methods.

6. Conclusions

This study proposed DualXBrNet, a dual-level fusion framework that integrates ResNet50V2, MobileNetV2, and EfficientNetB0 through learnable feature-level fusion and five-fold OOF decision-level stacking. The framework achieved 96.80% accuracy and a macro-ROC-AUC of 99.77% on the independent official BRISC2025 test set.

The Feature Fusion path provided the largest performance gain, whereas the subsequent decision-level stacking produced a smaller and statistically non-significant improvement. Oracle analysis indicated that the heterogeneous CNN backbones contributed partially non-redundant predictions, while the significant improvement of the 16-dimensional representation over the 12-dimensional CNN-only representation demonstrated the complementary contribution of the fusion-derived probabilities.

Furthermore, DualXBrNet also achieved 94.94% accuracy on the Masoud dataset without retraining. However, because shared images were identified between the two collections, this result should be regarded as a supplementary cross-dataset consistency assessment rather than fully independent external validation. Future work should evaluate the framework on institutionally independent datasets with patient-level separation, investigate alternative meta-fusion strategies, and conduct broader quantitative evaluations of explanation consistency.

Author Contributions: Conceptualization, G.A.A.-S.; methodology, all authors; data curation, S.A.-A.; formal analysis, all authors; investigation, all authors; project administration, G.A.A.-S.; software, S.A.-A.; supervision, G.A.A.-S.; validation, all authors; visualization, S.A.-A.; writing—original draft preparation, all authors; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The BRISC2025 dataset used for model development and official testing is publicly available at Kaggle: https://www.kaggle.com/datasets/briscdataset/brisc2025 (accessed on 16 August 2026). The additional Masoud Brain Tumor MRI dataset used for the supplementary cross-dataset consistency assessment is publicly available at Kaggle: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (accessed on 25 September 2026). The implementation code may be made available by the corresponding author upon reasonable request, subject to institutional and publication-related considerations.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5, OpenAI) to assist with language refinement, text organization, and improving the clarity of the manuscript. All AI-assisted content was critically reviewed, verified, and revised by the authors, who take full responsibility for the accuracy, originality, and integrity of the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AI	Artificial Intelligence
AUC	Area Under the Curve
BN	Batch Normalization
CI	Confidence Interval
CNN	Convolutional Neural Network
CV	Cross-Validation
DL	Deep Learning
FLOPs	Floating Point Operations
FN	False Negative
FP	False Positive
GAP	Global Average Pooling
LBFGS	Limited memory Broyden–Fletcher–Goldfarb–Shanno
LBP	Local Binary Pattern
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
OOF	Out-of-Fold
RBF-SVM	Radial Basis Function Support Vector Machine
ReLU	Rectified Linear Unit
RF	Random Forest
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
t-SNE	t-Distributed Stochastic Neighbor Embedding
TN	True Negative
TP	True Positive
UMAP	Uniform Manifold Approximation and Projection
XAI	Explainable Artificial Intelligence

References

1. Dorfner, F.J.; Patel, J.B.; Kalpathy-Cramer, J.; Gerstner, E.R.; Bridge, C.P. A review of deep learning for brain tumor analysis in MRI. *npj Precis. Oncol.* **2025**, *9*, 2. [CrossRef]
2. Bouhafra, S.; El Bahi, H. Deep Learning Approaches for Brain Tumor Detection and Classification Using MRI Images (2020 to 2024): A Systematic Review. *J. Imaging Inform. Med.* **2025**, *38*, 1403–1433. [CrossRef]
3. Reyes, D.; Sánchez, J. Performance of convolutional neural networks for the classification of brain tumors using magnetic resonance imaging. *Heliyon* **2024**, *10*, e25468. [CrossRef]
4. Rasa, S.M.; Islam, M.M.; Talukder, M.A.; Uddin, M.A.; Khalid, M.; Kazi, M.; Kazi, M.Z. Brain tumor classification using fine-tuned transfer learning models on magnetic resonance imaging (MRI) images. *Digit. Health* **2024**, *10*, 20552076241286140. [CrossRef]
5. Alahmed, H.A.; Al-Suhail, G.A. Exploring Transfer Learning Techniques for Brain Tumor Diagnosis in MRI Data. In Proceedings of the 2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI), Sana'a, Yemen, 26–27 November 2024; pp. 1–8. [CrossRef]
6. Mohammed, S.J.; Al-Suhail, G.A. Explainable AI-Based Deep Learning for Brain Tumor MRI Analysis: A Review. In Proceedings of the 2025 6th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), Ruse, Bulgaria, 26–28 November 2025; pp. 1–7. [CrossRef]
7. Jozi, N.S.; Al-Suhail, G.A.; Pham, V.-T. LCD-VRD: An Explainable Ensemble Deep Learning Framework for Lung Cancer Detection from CT Scans. *BioMedInformatics* **2026**, *6*, 36. [CrossRef]
8. Bouguerra, O.; Attallah, B.; Brik, Y. MRI-based brain tumor ensemble classification using two stage score level fusion and CNN models. *Egypt. Inform. J.* **2024**, *28*, 100565. [CrossRef]
9. Verma, A.; Yadav, A.K. FusionNet: Dual input feature fusion network with ensemble based filter feature selection for enhanced brain tumor classification. *Brain Res.* **2025**, *1852*, 149507. [CrossRef]
10. Goni, M.O.; Joy, D.D.; Ahmed, M.F.; Islam, S.; Nahid, N.; Mojumder, M.A.N. A Hybrid Stacking Ensemble of Deep Learning and Machine Learning Models for Brain Tumor Classification from MRI Scans. In Proceedings of the 2025 28th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 19–21 December 2025; pp. 2366–2371. [CrossRef]

11. Napa, K.K.; Murugan, S.; Murugan, J.S.; Jayanthi, A. BRAIN-META: A reproducible CNN–vision transformer meta-ensemble pipeline for explainable brain tumor classification. *MethodsX* **2026**, *16*, 103769. [CrossRef]
12. Celik, F.; Celik, K.; Celik, A. Enhancing brain tumor classification through ensemble attention mechanism. *Sci. Rep.* **2024**, *14*, 22260. [CrossRef]
13. Shaikh, A.; Amin, S.; Zeb, M.A.; Sulaiman, A.; Al Reshan, M.S.; Alshahrani, H. Enhanced brain tumor detection and segmentation using densely connected convolutional networks with stacking ensemble learning. *Comput. Biol. Med.* **2025**, *186*, 109703. [CrossRef]
14. Haque, R.; Khan, M.A.; Rahman, H.; Khan, S.; Siddiqui, M.I.H.; Limon, Z.H.; Swapno, S.M.M.R.; Appaji, A. Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. *Comput. Biol. Med.* **2025**, *191*, 110166. [CrossRef]
15. Ullah, Z.; Kim, J. Hierarchical Deep Feature Fusion and Ensemble Learning for Enhanced Brain Tumor MRI Classification. *Mathematics* **2025**, *13*, 2787. [CrossRef]
16. Kakon, S.C.; Sazid, Z.A.; Begum, I.A.; Samad, M.A.; Hosen, A.S.M.S. Explainable Deep Ensemble Meta-Learning Framework for Brain Tumor Classification Using MRI Images. *Cancers* **2025**, *17*, 2853. [CrossRef]
17. Uthra, B.; Sivashankari, R. Query-Guided Learning for Efficient and Interpretable Multi-Class Brain Tumor Classification in MRI. *Front. Artif. Intell.* **2026**, *9*, 1849051. [CrossRef]
18. Chen, W.; Tan, X.; Zhang, J.; Du, G.; Fu, Q.; Jiang, H. A Robust Approach for Multi-Type Classification of Brain Tumor Using Deep Feature Fusion. *Front. Neurosci.* **2024**, *18*, 1288274. [CrossRef]
19. Patro, S.G.K.; Govil, N.; Saxena, S.; Mishra, B.K.; Zamani, A.T.; Ben Miled, A.; Parveen, N.; Elshafie, H.; Hamdan, M. Brain Tumor Classification Using an Ensemble of Deep Learning Techniques. *IEEE Access* **2024**, *12*, 162094–162106. [CrossRef]
20. Velpula, V.K.; Vadlamudi, J.S.; Janapati, M.; Kasaraneni, P.P.; Kumar, Y.V.P.; Challa, P.R.; Mallipeddi, R. Enhanced Brain Tumor Classification Using Convolutional Neural Networks and Ensemble Voting Classifier for Improved Diagnostic Accuracy. *Comput. Electr. Eng.* **2025**, *123*, 110124. [CrossRef]
21. Eone Etoua, O.V.; Tapamo Kenfack, H.M.; Tapamo, J.-R. F1-Weighted Ensemble of CNN and Local Features for Explainable Brain Tumor MRI Classification. *Appl. Comput. Intell. Soft Comput.* **2026**, *2026*, 7961939. [CrossRef]
22. Al-Azzawi, S.J.M.; Al-Suhail, G.A. BrTxNet: A Hybrid EfficientNetB0–VGG16 Model with Learnable Feature Fusion for Brain Tumor MRI Classification. *Eng. Proc.* **2026**, *154*, 56. [CrossRef]
23. Talukder, M.A.; Islam, M.M.; Uddin, M.A.; Layek, M.A.; Acharjee, U.K.; Bhuiyan, T.; Moni, M.A. A deep ensemble learning framework for brain tumor classification using data balancing and fine-tuning. *Sci. Rep.* **2025**, *15*, 35251. [CrossRef]
24. Singh, R.; Gupta, S.; Ibrahim, A.O.; Gabralla, L.A.; Bharany, S.; Rehman, A.U.; Hussen, S. Advanced dynamic ensemble framework with explainability driven insights for precision brain tumor classification across datasets. *Sci. Rep.* **2025**, *15*, 29090. [CrossRef]
25. Panigrahi, S.; Das Adhikary, D.R.; Pattanayak, B.K. Hybrid transfer learning and self-attention framework for robust MRI-based brain tumor classification. *Sci. Rep.* **2025**, *15*, 21343. [CrossRef]
26. Jayaraman, G.; Meganathan, S.; Shah, S.S.M.; Anuradha, M.; Sundararajan, R.K.; Rajakumar, R. A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification. *Sci. Rep.* **2025**, *15*, 45769. [CrossRef]
27. Yonar, A. A swarm intelligence-driven hybrid framework for brain tumor classification with enhanced deep features. *Sci. Rep.* **2025**, *15*, 37543. [CrossRef]
28. Sánchez-Moreno, L.; Perez-Peña, A.; Duran-Lopez, L.; Dominguez-Morales, J.P. Ensemble-based Convolutional Neural Networks for brain tumor classification in MRI: Enhancing accuracy and interpretability using explainable AI. *Comput. Biol. Med.* **2025**, *195*, 110555. [CrossRef]
29. Zarenia, E.; Akhlaghi Far, A.; Rezaee, K. Automated multi-class MRI brain tumor classification and segmentation using deformable attention and saliency mapping. *Sci. Rep.* **2025**, *15*, 8114. [CrossRef]
30. Mao, Y.; Kim, J.; Podina, L.; Kohandel, M. Dilated SE-DenseNet for brain tumor MRI classification. *Sci. Rep.* **2025**, *15*, 3596. [CrossRef]
31. Fateh, A.; Rezvani, Y.; Moayedi, S.; Rezvani, S.; Fateh, F.; Fateh, M.; Abolghasemi, V. BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification. *Sci. Data* **2026**, *13*, 361. [CrossRef]
32. Nickparvar, M. *Brain Tumor MRI Dataset*; Kaggle: San Francisco, CA, USA, 2021. Available online: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (accessed on 16 August 2026).
33. Al Bataineh, A.F.; Nahar, K.M.O.; Khafajeh, H.; Samara, G.; Alazaidah, R.; Nasayreh, A.; Bashkami, A.; Gharaibeh, H.; Dawaghreh, W. Enhanced Magnetic Resonance Imaging-Based Brain Tumor Classification with a Hybrid Swin Transformer and ResNet50V2 Model. *Appl. Sci.* **2024**, *14*, 10154. [CrossRef]
34. Adamu, M.J.; Kawuwa, H.B.; Qiang, L.; Nyatega, C.O.; Younis, A.; Fahad, M.; Dauya, S.S. Efficient and Accurate Brain Tumor Classification Using Hybrid MobileNetV2–Support Vector Machine for Magnetic Resonance Imaging Diagnostics in Neoplasms. *Brain Sci.* **2024**, *14*, 1178. [CrossRef]
35. Mahesh, T.R.; Gupta, M.; Anupama, T.A.; Kumar, V.V.; Geman, O.; Kumar, V.D. An XAI-enhanced EfficientNetB0 framework for precision brain tumor detection in MRI imaging. *J. Neurosci. Methods* **2024**, *410*, 110227. [CrossRef]

36. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. Erratum in *Sci. Rep.* **2024**, *14*, 15724. [CrossRef]
37. Aguilar-Ruiz, J.S.; Michalak, M. Classification performance assessment for imbalanced multiclass data. *Sci. Rep.* **2024**, *14*, 10759. [CrossRef]
38. Rink, P.; Brannath, W. Post-selection confidence bounds for prediction performance. *Mach. Learn.* **2025**, *114*, 82. [CrossRef]
39. Jung, S.; Dagobert, T.; Morel, J.-M.; Facciolo, G. A Review of t-SNE. *Image Process. Line* **2024**, *14*, 250–270. [CrossRef]
40. Healy, J.; McInnes, L. Uniform manifold approximation and projection. *Nat. Rev. Methods Primers* **2024**, *4*, 82. [CrossRef]
41. Ali, S.A.; Arain, K.R.; Mushtaq, N.A.; ul Rehman, O. Interpretable Deep Learning for Brain Tumor Diagnosis: Occlusion Sensitivity-Driven Explainability in MRI Classification. *VFAST Trans. Softw. Eng.* **2025**, *13*, 135–146. [CrossRef]
42. Almnawer, R.; Alhaj, Z.; Ozturk, M. CrossTaskFormer: Bidirectional cross-task attention for joint brain tumor segmentation and classification. *J. King Saud Univ. Comput. Inf. Sci.* **2026**, *38*, 542. [CrossRef]
43. Alkharaan, R.; Alobaidi, J.; Bakarman, J.; Alshamlan, H. Brain Tumor Classification and Segmentation in MR Images Using EfficientNet and U-Net++ Models. *Diagnostics* **2026**, *16*, 1745. [CrossRef]
44. Bouguerra, O.; Brik, Y.; Alzahrani, S.I.; Attallah, B.; Djerioui, M.; Shayea, I. FALCON: A multi-layered fusion ensemble for high-certainty brain tumor diagnosis from MRI. *J. Radiat. Res. Appl. Sci.* **2025**, *18*, 101940. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.