European Journal of Computer Science and Information Technology, 14(7),38-55, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Leakage-Aware Evaluation of a Dual-Backbone Feature-Fusion Ensemble for Multi-Class Brain Tumor MRI Classification

Md Selim Mia,[1] Md Nasir Uddin,[2] Sagor Anup Chandra,[3] Osman Goni,[4] Shaykat Purokayisto,[5*]

[1,2,5]School of Artificial Intelligence, Nanjing University of Information Science & Technology, Nanjing, China

[3,4]School of Computer Science & Technology, Nanjing University of Information Science & Technology, Nanjing, China

Corresponding Email: selim.shpi@gmail.com‖shaykatariyan@gmail.com

doi: https://doi.org/10.37745/ejcsit.2013/vol14n73855 Published November 01, 2026

Citation: Mia M.S, Uddin M.N., Chandra S.A., Osman Goni, Shaykat Purokayisto (2026) Leakage-Aware Evaluation of a Dual-Backbone Feature-Fusion Ensemble for Multi-Class Brain Tumor MRI Classification, European Journal of Computer Science and Information Technology, 14(7),38-55

Abstract: *Deep-learning studies of brain tumor MRI classification often report accuracies close to or above 99%, yet public benchmarks contain duplicate and near-duplicate images that can leak between training and test sets. This study audits the widely used 7,200-image Brain Tumor MRI Dataset and proposes a leakage-aware protocol that removes exact duplicates and groups near-duplicates, using perceptual hashing with pixel correlation, before a group-stratified split. The audit shows that 36.25% of the official test images have a near-duplicate in the official training set. Under the leakage-free protocol, an EfficientNetV2-B0 + ResNet50V2 feature-fusion network and a three-model ensemble were trained with test-time augmentation. The ensemble, selected on validation data alone, achieved 97.80% accuracy (95% CI 96.86–98.64%), macro-F1 of 0.978 and ROC-AUC of 0.998 on 954 unseen images, missing one of 740 tumors; the fusion model alone reached 98.32%. Rigorous de-duplication thus yields credible, still clinically promising, performance estimates.*

Keywords: Brain tumor classification, MRI, Data leakage, Feature fusion, Ensemble learning, Transfer learning

INTRODUCTION

Tumors of the brain and central nervous system are among the most consequential neoplasms in clinical practice: they are frequently aggressive, they arise in an organ with little tolerance for mass effect, and their management depends on an accurate diagnosis of tumor type. The 2021 World Health Organization classification of central nervous system tumors integrates histological and molecular features into a detailed taxonomy (Louis et al., 2021), and population registries continue to document a substantial burden of primary brain and other central nervous system tumors (Ostrom et al., 2023). Magnetic resonance imaging (MRI) is the primary non-invasive modality for detecting such lesions

and for planning further work-up, but visual interpretation of large MRI volumes is time-consuming and subject to inter-reader variability. Automated decision support that reliably distinguishes common tumor types from each other and from normal scans is therefore of considerable practical interest.Deep convolutional neural networks (CNNs) have transformed medical image analysis (Litjens et al., 2017). Because annotated medical data sets are small, most brain tumor classifiers rely on transfer learning from networks pre-trained on ImageNet (Deng et al., 2009), whose early-layer features transfer well across domains (Yosinski et al., 2014; Raghu et al., 2019). Using this strategy, published studies on public brain tumor MRI collections report accuracies of roughly 94–99% (Swati et al., 2019; Deepak and Ameer, 2019; Rehman et al., 2020), and even higher figures are often quoted for the popular Kaggle collections.

Such headline figures are only as trustworthy as the evaluation protocol behind them. *Data leakage*—any flow of information from the evaluation data into model development—systematically inflates performance (Kaufman et al., 2012) and has been identified as a central cause of irreproducible results in machine-learning-based science (Kapoor and Narayanan, 2023). In medical imaging, splitting at the level of images rather than patients, and reusing the test set for model selection, are recurrent methodological failures (Roberts et al., 2021; Varoquaux and Cheplygina, 2022). For brain tumor MRI the effect is large: with the same CNN and data set, Badža and Barjaktarović (2020) obtained 96.56% accuracy under record-wise ten-fold cross-validation but only 88.48% under subject-wise cross-validation. The widely used Brain Tumor MRI Dataset (Nickparvar, 2021) is itself a curated merger of the figshare collection (Cheng et al., 2015), the SARTAJ collection (Bhuvaji et al., 2020) and the Br35H collection (Hamada, 2020). It contains adjacent slices of the same patients and repeated images, while patient identifiers are not distributed, so a patient-level split cannot be performed directly.

This study addresses that gap by combining a duplicate-aware evaluation protocol with a strong yet transparent classification model. Its contributions are as follows. **First**, an image-fingerprinting audit (MD5 hashing, perceptual hashing and normalised cross-correlation) quantifies leakage in the official split of the Brain Tumor MRI Dataset: 36.25% of the official test images have an exact or near-duplicate in the official training set. **Second**, a leakage-aware protocol is proposed in which near-duplicate groups are kept within a single partition of a group-stratified split, the final model is selected on validation data only, and the test set is used once. **Third**, a dual-backbone feature-fusion network (EfficientNetV2-B0 and ResNet50V2, each with its own correct input preprocessing) and a three-model ensemble with test-time augmentation are developed; the validation-selected ensemble reaches 97.80% accuracy and the fusion model 98.32% on 954 strictly unseen images. **Fourth**, the evaluation reports confidence intervals, chance-corrected agreement, McNemar tests, Grad-CAM explanations and an error analysis, rather than a single accuracy value.

The remainder of the paper is organised as follows. Section 2 reviews related work and the theoretical basis of the approach. Section 3 describes the data, the leakage-aware protocol and the models. Section 4 reports the findings, Section 5 discusses them together with the study's limitations, Section 6 outlines implications for research and practice, and Sections 7 and 8 conclude and propose future research.

LITERATURE AND THEORETICAL UNDERPINNING

Brain tumor classification from MRI

Early work relied on handcrafted descriptors. Cheng et al. (2015) introduced the figshare data set of 3,064 contrast-enhanced T1-weighted slices from 233 patients (glioma, meningioma and pituitary tumor) and showed that augmenting the tumor region and partitioning it into ring-shaped sub-regions improved intensity-histogram, grey-level co-occurrence and bag-of-words features, reaching 91.28% accuracy. CNN-based approaches subsequently dominated. Abiwinanda et al. (2019) trained compact CNNs from scratch, Afshar, Mohammadi and Plataniotis (2018) explored capsule networks to better encode spatial relationships, and Badža and Barjaktarović (2020) designed a dedicated CNN and evaluated it under both record-wise and subject-wise cross-validation. Transfer learning produced the highest reported accuracies: Swati et al. (2019) used block-wise fine-tuning of VGG19 (94.82%, five-fold cross-validation), Deepak and Ameer (2019) used GoogLeNet features (98% mean accuracy) and Rehman et al. (2020) fine-tuned VGG16 (98.69%). Table 1 summarises these studies. A common thread is that evaluation protocols differ considerably, and only a few studies explicitly control for images of the same patient appearing on both sides of the split.

Table 1. Selected prior studies on brain tumor MRI classification. Accuracies are as reported by the authors and were obtained under different data sets and protocols, so they are not directly comparable.

Study	Data	Approach	Reported accuracy
Cheng et al. (2015)	figshare, 3 classes	Tumor-region augmentation and ring-form partition; BoW features	91.28%
Swati et al. (2019)	figshare, 3 classes	Block-wise fine-tuned VGG19, 5-fold CV	94.82%
Deepak and Ameer (2019)	figshare, 3 classes	GoogLeNet transfer-learned features	98% (mean)
Rehman et al. (2020)	figshare, 3 classes	Fine-tuned VGG16	98.69%
Badža and Barjaktarović (2020)	figshare, 3 classes	Custom CNN, 10-fold CV: record-wise / subject-wise	96.56% / 88.48%
This study	**Nickparvar v2, 4 classes, de-duplicated**	**Fusion network and 3-model TTA ensemble; group-aware split**	**97.80% (ensemble), 98.32% (fusion)**

Transfer learning and complementary backbones

Residual networks made very deep CNNs trainable through identity shortcuts (He et al., 2016a); the pre-activation variant used here, ResNet50V2, moves batch normalisation and activation before each convolution, which improves gradient flow (He et al., 2016b; Ioffe and Szegedy, 2015). EfficientNet scales depth, width and resolution jointly (Tan and Le, 2019), and EfficientNetV2 adds fused-MBConv blocks and training-aware architecture search to obtain accurate yet compact models (Tan and Le, 2021). The two families therefore embody different inductive biases plain residual bottlenecks versus depth-wise, squeeze-and-excitation blocks and can be expected to produce partly complementary representations. Two established ways of exploiting such complementarity are *feature-level fusion*, in which pooled embeddings are concatenated before a shared classifier, and *decision-level ensembling*, in which the predicted probabilities of independently trained models are averaged. Deep ensembles improve both accuracy and the quality of predictive uncertainty (Lakshminarayanan, Pritzel and Blundell, 2017).

Data leakage and the validity of evaluation

Kaufman et al. (2012) formalised leakage as the use of information at training or selection time that would not be available at deployment. Kapoor and Narayanan (2023) catalogued leakage across scientific fields and showed that it repeatedly produced over-optimistic conclusions. In medical imaging, Roberts et al. (2021) found that most COVID-19 imaging models suffered from such methodological flaws, and Varoquaux and Cheplygina (2022) argued that test sets are often too small, reused for model selection or not independent of the training data. When patient identifiers are unavailable, image similarity offers a practical surrogate: cryptographic hashes such as MD5 detect byte-identical files, whereas perceptual hashes based on the discrete cosine transform are robust to re-encoding, resizing and small intensity changes (Zauner, 2010). Credible reporting further requires uncertainty estimates and paired statistical tests. Bootstrap confidence intervals (Efron and Tibshirani, 1993), Cohen's kappa (Cohen, 1960), the Matthews correlation coefficient, which is informative under class imbalance (Chicco and Jurman, 2020), and McNemar's test for paired classifiers (McNemar, 1947; Dietterich, 1998) are adopted in this study.

Research gap

Although high accuracies have been reported on public brain tumor MRI data, few studies (i) quantify duplicate and near-duplicate leakage in the benchmark itself, (ii) enforce a split in which near-identical images cannot cross partitions, and (iii) select the final model without consulting the test set. This study addresses all three while still pursuing a high-performing, explainable architecture.

METHODOLOGY

Overview

Figure 1 summarises the workflow. All images were cropped and resized, fingerprinted and grouped into near-duplicate families; the families were then partitioned by a group-stratified split. Three models were trained with an identical two-phase transfer-learning recipe, combined by test-time augmentation and soft voting, and the final model was chosen on the validation set before a single evaluation on the test set.

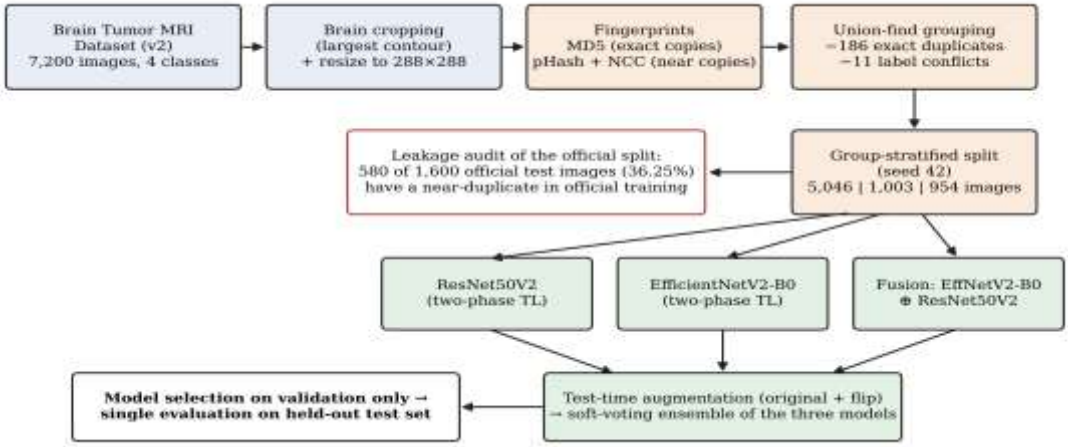


Figure 1. Overview of the leakage-aware workflow. Orange boxes denote the de-duplication and splitting stages; green boxes denote model training and inference.

Data set

The Brain Tumor MRI Dataset, version 2 (Nickparvar, 2021), was used. It contains 7,200 two-dimensional MRI slices in four classes—glioma, meningioma, pituitary tumor and no tumor—distributed as 1,400 training and 400 testing images per class. According to its documentation, the collection combines the figshare (Cheng et al., 2015), SARTAJ (Bhuvaji et al., 2020) and Br35H (Hamada, 2020) data sets; images of the no-tumor class originate from Br35H, and glioma images of SARTAJ with inconsistent labels were replaced by figshare images. The slices include axial, coronal and sagittal views and several MRI sequences, and image sizes vary.

Preprocessing

Each image was converted to greyscale, smoothed with a 5 × 5 Gaussian filter and thresholded at an intensity of 45; after two erosion and two dilation iterations, the bounding box of the largest external contour was used to crop away the empty background (the original image was kept if the contour covered less than 20% of either dimension). Cropped images were resized with area interpolation to 288 × 288 pixels for training and inference, and separately to 224 × 224 pixels for fingerprinting, so that duplicate detection is independent of the training resolution.

Duplicate-aware grouping

Three fingerprints were computed for every image. An MD5 hash of the original file identifies byte-identical copies. A 64-bit perceptual hash h was obtained by resizing the greyscale image to 32 × 32 pixels, applying a two-dimensional discrete cosine transform, retaining the 8 × 8 lowest-frequency coefficients and thresholding them at their median (Zauner, 2010). The dissimilarity of two images a and b is the Hamming distance between their hashes:

$$d(a, b) = \Sigma_k \left[h_k(a) \neq h_k(b) \right], \quad k = 1, \ldots, 64 \qquad (1)$$

Because perceptual hashes of MRI slices with similar global anatomy can collide, a second, pixel-level criterion was required. Each image was resized to 64 × 64 pixels and flattened into a vector x; with $\bar{x}$ denoting its mean, the normalised cross-correlation is

$$r(a, b) = (x_a - \bar{x}_a) \cdot (x_b - \bar{x}_b) / \left(|x_a - \bar{x}_a| \cdot |x_b - \bar{x}_b| \right) \qquad (2)$$

Two images were linked when they had the same MD5 hash, or when $d(a, b) \leq 8$ **and** $r(a, b) \geq 0.90$. These thresholds were calibrated by visual inspection of borderline pairs on the independent SARTAJ collection, where the perceptual hash alone chained hundreds of unrelated images into a single group, whereas the joint criterion produced compact groups of at most ten images consisting of adjacent slices of the same examination. Connected components of the resulting graph, found with a union-find structure, define near-duplicate *groups*. Groups whose members carried different class labels were removed as label noise, and only one copy of each set of byte-identical files was kept. Figure 3 (Section 4.1) shows typical linked pairs.

Leakage audit and group-stratified split

Before any splitting, the official partition was audited by counting the official test images whose group also contains an official training image. The de-duplicated images were then split with a stratified group k-fold procedure (random seed 42): one of seven folds formed the test set, and one of six folds of the remainder formed the validation set, giving an approximately 72/14/14 split in which no group and therefore no exact or near-duplicate is shared between partitions. This property was verified programmatically.

Models

Three models were trained (Figure 2). The *ResNet50V2* and *EfficientNetV2-B0* models each consist of one ImageNet-pretrained backbone followed by global average pooling (GAP) and a classification head. The proposed *fusion* model passes the same input through both backbones in parallel and concatenates the 1,280-dimensional EfficientNetV2-B0 embedding with the 2,048-dimensional ResNet50V2 embedding into a 3,328-dimensional vector. Correct input preprocessing is applied to each branch: EfficientNetV2-B0 receives raw pixel values in [0, 255] and rescales them internally, while the ResNet50V2 branch receives values linearly mapped to [−1, 1]. All three models share the same head: dropout (0.4), a 512-unit dense layer with ReLU activation and L2 weight decay (10^{-4}), dropout (0.3) and a four-way softmax output (Srivastava et al., 2014). The models contain 24.62 M (ResNet50V2), 6.58 M (EfficientNetV2-B0) and 31.19 M (fusion) parameters.

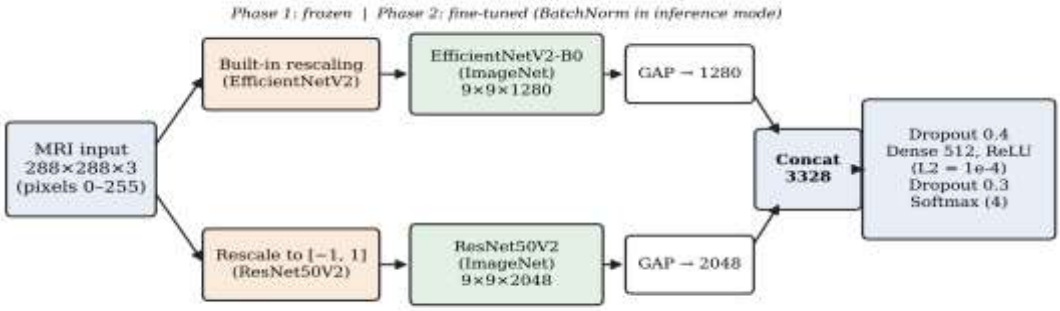


Figure 2. Architecture of the proposed dual-backbone feature-fusion network. Each backbone receives its own input normalisation; at 288 × 288 input both backbones yield 9 × 9 feature maps.

Training procedure

Training followed two phases. In phase 1 the backbones were frozen and only the head was trained with Adam (Kingma and Ba, 2015) at a learning rate of 10^{-3} for up to 12 epochs, with the learning rate reduced by a factor of 0.3 after two epochs without improvement in validation loss and early stopping

after four. In phase 2 all layers were unfrozen while batch-normalisation layers were kept in inference mode, and the model was fine-tuned for up to 35 epochs with a one-epoch linear warm-up from 5×10^{-7} to 5×10^{-5} followed by cosine decay (Loshchilov and Hutter, 2017); early stopping monitored validation loss with a patience of eight epochs and restored the best weights. Categorical cross-entropy with label smoothing of 0.1 was used (Szegedy et al., 2016). On-line augmentation (Shorten and Khoshgoftaar, 2019) comprised horizontal flips, rotations of up to ±18°, zoom of ±10%, translations of ±5%, and contrast and brightness changes of ±10%. Mixed-precision arithmetic was used to accelerate training (Micikevicius et al., 2018). Table 2 lists the settings.

Inference, ensembling and model selection

At inference, test-time augmentation (TTA) averaged the softmax outputs p for the original image x and its horizontal mirror x'. The ensemble then averaged the TTA outputs of the $M = 3$ models:

$$P(y \mid x) = (1/M)\, \Sigma_m \tfrac{1}{2} [\, p_m(y \mid x) + p_m(y \mid x') \,] \qquad (3)$$

The candidate models—the three single models and the ensemble—were ranked by validation accuracy, with ties broken by validation log-loss. This rule was fixed before training, and the top-ranked candidate was declared the final model before any test prediction was inspected.

Evaluation

Performance on the held-out test set was summarised by accuracy, balanced accuracy, macro-averaged precision, recall (sensitivity), F1-score and specificity, Cohen's kappa, the Matthews correlation coefficient (MCC) and the macro one-versus-rest area under the ROC curve (AUC). Ninety-five per cent confidence intervals for accuracy and macro-F1 were obtained from 1,000 bootstrap resamples of the test set. Paired comparisons between models used the exact (binomial) McNemar test on discordant predictions. Gradient-weighted class activation maps (Grad-CAM) were computed separately for the last convolutional feature maps of both fusion branches (Selvaraju et al., 2017).

Implementation

The pipeline was implemented in Python 3.13 with TensorFlow 2.20, Keras 3.13.2, OpenCV 4.14 and scikit-learn, and was run on a Kaggle notebook with an NVIDIA Tesla T4 GPU (16 GB). Training the three models took 60.1, 65.2 and 69.8 minutes, respectively. All random number generators were seeded, and the split manifest, training histories and metrics were exported to enable full reproduction.

Table 2. Training and inference configuration (identical for all three models).

Setting	Value
Input size / batch size	288 × 288 × 3 / 32
Phase 1 (frozen backbones)	Adam, lr 10^{-3}, ≤ 12 epochs; ReduceLROnPlateau (×0.3, patience 2); early stop (patience 4)
Phase 2 (full fine-tuning)	Adam, 1-epoch warm-up to 5×10^{-5}, cosine decay to 1%; ≤ 35 epochs; early stop (patience 8)
Loss	Categorical cross-entropy, label smoothing 0.1
Regularisation	Dropout 0.4 / 0.3; L2 10^{-4}; BatchNorm in inference mode
Augmentation	Flip (horizontal), rotation ±18°, zoom ±10%, shift ±5%, contrast and brightness ±10%
Inference	TTA (original + horizontal flip); soft-voting ensemble of 3 models
Hardware / precision	NVIDIA Tesla T4; mixed precision (float16)

RESULTS AND FINDINGS

Leakage audit and de-duplication

Table 3 summarises the audit. Among the 7,200 images, 187 byte-identical links and 3,391 near-duplicate links were found. No official test image had a byte-identical copy in the official training set, confirming the exact de-duplication performed by the data set's curator; nevertheless, **580 of the 1,600 official test images (36.25%) belong to a near-duplicate group that also contains official training images**. After removing 186 redundant exact copies and 11 images in label-conflicting groups, 7,003 images in 5,364 groups (largest group: 18 images) remained. Figure 3 illustrates near-duplicate pairs; most are adjacent slices from the same examination.

Table 3. Results of the duplicate audit of the Brain Tumor MRI Dataset (version 2).

Quantity	Value
Images in the data set (official train / test)	7,200 (5,600 / 1,600)
Exact-duplicate links (MD5) / near-duplicate links	187 / 3,391
Redundant exact copies removed	186
Images removed from label-conflicting groups	11
Images retained / near-duplicate groups / largest group	7,003 / 5,364 / 18
Official test images with an exact copy in official training	0 (0.00%)
Official test images with an exact or near-duplicate in official training	**580 (36.25%)**

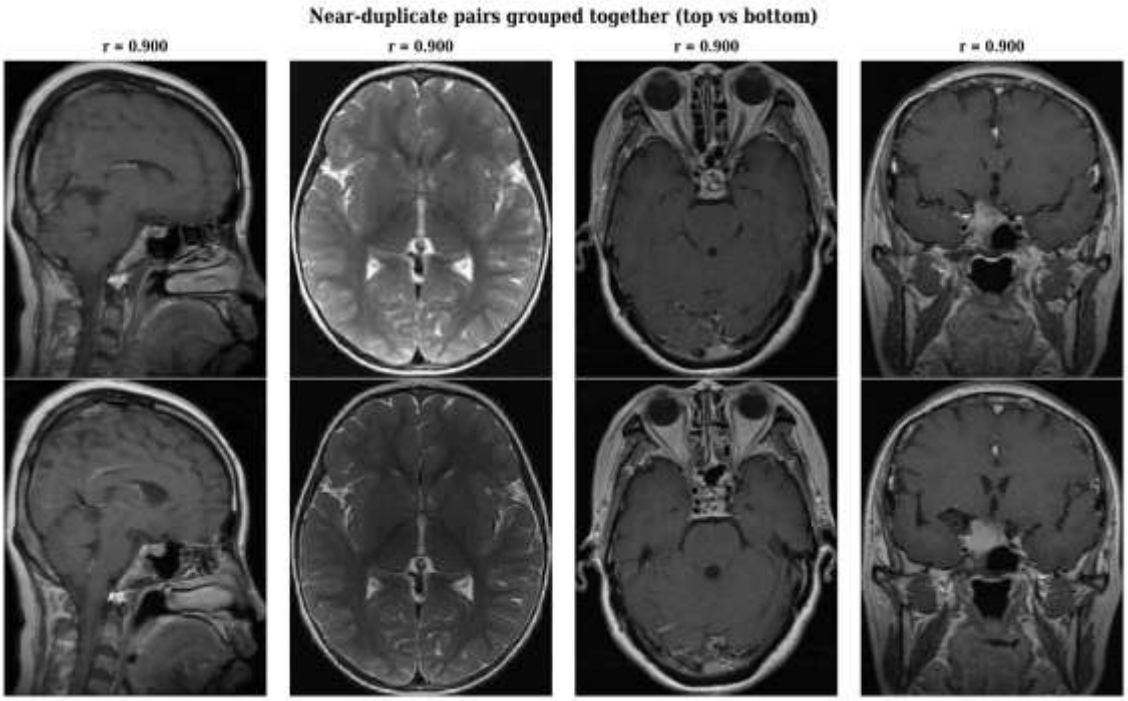


Figure 3. Examples of near-duplicate pairs (top versus bottom) linked by the joint perceptual-hash and correlation criterion; r denotes the normalized cross-correlation.

Table 4 gives the resulting group-stratified split. Class proportions are preserved in all partitions, and the test set contains 954 images, none of which has an exact or near-duplicate in the training or validation data.

Table 4. Leakage-free group-stratified split of the de-duplicated data.

Class	Train	Validation	Test	Total
Glioma	1,290	240	252	1,782
Meningioma	1,277	258	243	1,778
No tumor	1,220	247	214	1,681
Pituitary	1,259	258	245	1,762
Total	**5,046**	**1,003**	**954**	**7,003**

Training behaviour and model selection

All models converged smoothly. The best fine-tuning epochs (by validation loss) were 26 of 34 for ResNet50V2, 34 of 35 for EfficientNetV2-B0 and 29 of 35 for the fusion model. Figure 4 shows the learning curves of the fusion model: validation accuracy rises sharply once the backbones are unfrozen and then plateaus near 98.5%, while the validation loss decreases monotonically without the divergence that would indicate harmful over-fitting. On the validation set, the ensemble achieved the highest accuracy (98.80%), followed by the fusion model (98.60%), ResNet50V2 (98.50%) and EfficientNetV2-B0 (97.81%). According to the pre-specified rule, the **ensemble was selected as the final model** before the test set was examined.

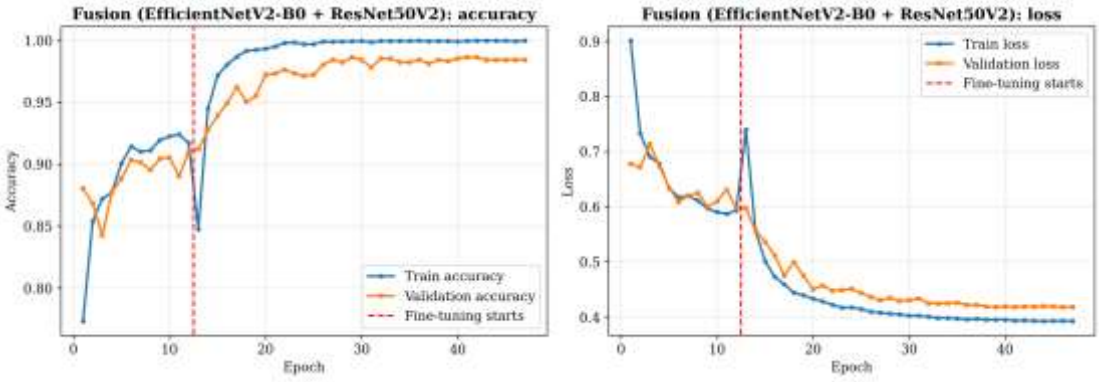


Figure 4. Training and validation accuracy (left) and loss (right) of the fusion model; the dashed line marks the start of full fine-tuning.

Test-set performance

Table 5 reports the test-set results of all candidates. The selected ensemble achieved **97.80% accuracy (95% CI 96.86–98.64%)**, a macro-F1 of 0.9781, Cohen's kappa and MCC of 0.9706 and a macro AUC of 0.9983. The fusion model obtained the highest single-model accuracy of 98.32% (95% CI 97.48–99.06%) with only 16 errors, ahead of ResNet50V2 (97.59%) and EfficientNetV2-B0 (96.96%). All candidates exceed 96.9% accuracy and 0.997 AUC, indicating that strong performance persists after leakage has been removed.

Table 5. Validation and test performance of all candidate models (macro averages; TTA applied; 954 test images). The ensemble was selected on validation accuracy. Cohen's kappa (0.959–0.978) differs from the MCC by at most 0.001 for every model.

Model	Val acc. (%)	Test acc. (%) [95% CI]	Prec.	Recall	F1	MCC	AUC	Params (M)
ResNet50V2	98.50	97.59 [96.65–98.53]	0.977	0.976	0.976	0.968	0.998	24.62
EfficientNetV2-B0	97.81	96.96 [95.70–98.01]	0.971	0.969	0.970	0.960	0.998	6.58
Fusion (proposed)	98.60	98.32 [97.48–99.06]	0.984	0.983	0.984	0.978	0.998	31.19
Ensemble (selected)	**98.80**	**97.80 [96.86–98.64]**	**0.979**	**0.978**	**0.978**	**0.971**	**0.998**	**62.38**

Class-wise and clinical performance

Table 6 and Figure 5 detail the selected ensemble. Pituitary tumors (recall 99.18%) and gliomas (98.41%) were recognised most reliably, whereas meningiomas were the most difficult class (recall 96.30%), mainly because four meningiomas were labelled glioma and five pituitary tumor. From a screening perspective, only **one of the 740 tumor images** was classified as *no tumor*, corresponding to a tumor-versus-normal sensitivity of 99.86%; six of the 214 normal scans were flagged as tumors (specificity 97.20%). The ROC curves in Figure 6 are close to the upper-left corner for every class, with per-class AUCs above 0.99.

Class	Precision	Recall	F1-score	Specificity	Support
Glioma	0.9764	0.9841	0.9802	0.9915	252
Meningioma	0.9710	0.9630	0.9669	0.9902	243
No tumor	0.9952	0.9720	0.9835	0.9986	214
Pituitary	0.9720	0.9918	0.9818	0.9901	245
Macro average	**0.9786**	**0.9777**	**0.9781**	**0.9926**	**954**

Table 6. Per-class test performance of the selected ensemble.

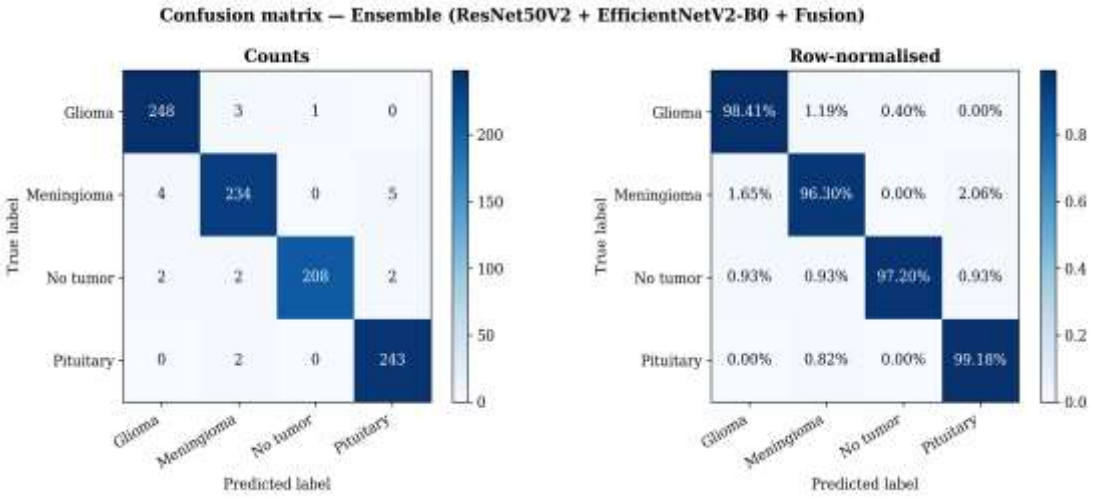
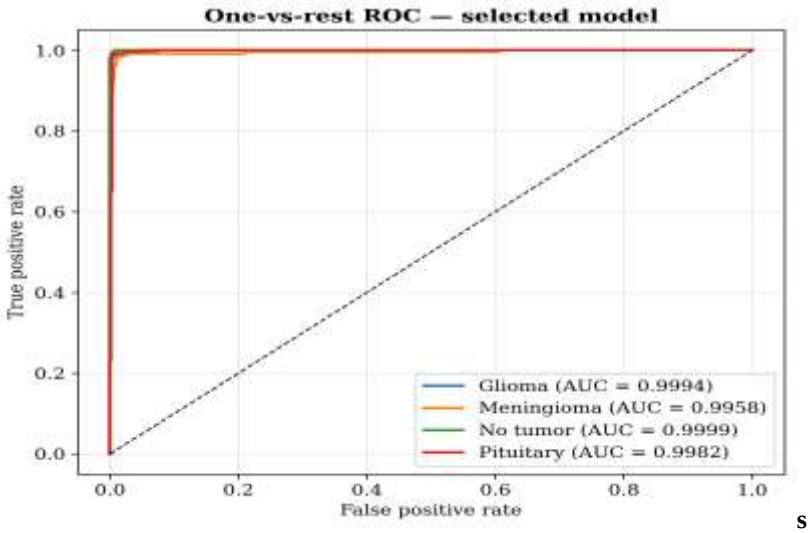


Figure 5. Confusion matrix of the selected ensemble on the test set: absolute counts (left) and row-normalized percentages (right).



s

Figure 6. One-versus-rest ROC curves of the selected ensemble.

Effect of the design choices

A preliminary configuration (224 × 224 input, fine-tuning learning rate 3×10^{-5} without warm-up, at most 25 epochs, no TTA), evaluated on the identical split, provides an ablation of the final design (Table 7). In that configuration EfficientNetV2-B0 was under-trained its best epoch was the 24th of

25 and reached only 93.92%, which in turn limited the fusion model (96.75%) to below ResNet50V2 alone (97.48%). The final design raised EfficientNetV2-B0 by 3.04 points and the fusion model by 1.57 points, after which fusion became the best single model. This shows that feature fusion only pays off when *both* branches are adequately trained.

Model	Preliminary (224 px, no TTA)	Final (288 px, warm-up, TTA)	Change (points)
ResNet50V2	97.48%	97.59%	+0.11
EfficientNetV2-B0	93.92%	96.96%	+3.04
Fusion	96.75%	98.32%	+1.57
Ensemble	—	97.80%	—

Table 7. Test accuracy of the preliminary and final configurations on the same leakage-free split.

Statistical comparison

Exact McNemar tests compared the selected ensemble with the other candidates on the 954 paired test predictions. Against ResNet50V2 the ensemble was correct on 3 images that ResNet50V2 misclassified and wrong on 1 that ResNet50V2 classified correctly ($p = 0.625$); against EfficientNetV2-B0 the counts were 12 and 4 ($p = 0.077$); against the fusion model they were 0 and 5 ($p = 0.063$). None of the differences is significant at the 5% level, so the four candidates should be regarded as statistically comparable, with the fusion model and the ensemble forming the strongest group.

Explainability and error analysis

Grad-CAM maps (Figure 7) show that, for meningioma and pituitary cases, both branches of the fusion model concentrate on the lesion and the sellar region respectively, while for normal scans the attention is distributed over the parenchyma, as expected in the absence of a focal lesion. The ResNet50V2 branch generally produces more focal maps than the EfficientNetV2-B0 branch, which is consistent with the two backbones contributing different evidence. The misclassified cases (Figure 8) are dominated by meningioma–glioma confusions involving large, heterogeneous or atypically enhancing masses, and by a meningioma located near the sella that was labelled pituitary tumor; several of these images are ambiguous even on visual inspection, and some may reflect label noise inherited from the source collections.

European Journal of Computer Science and Information Technology, 14(7),38-55, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

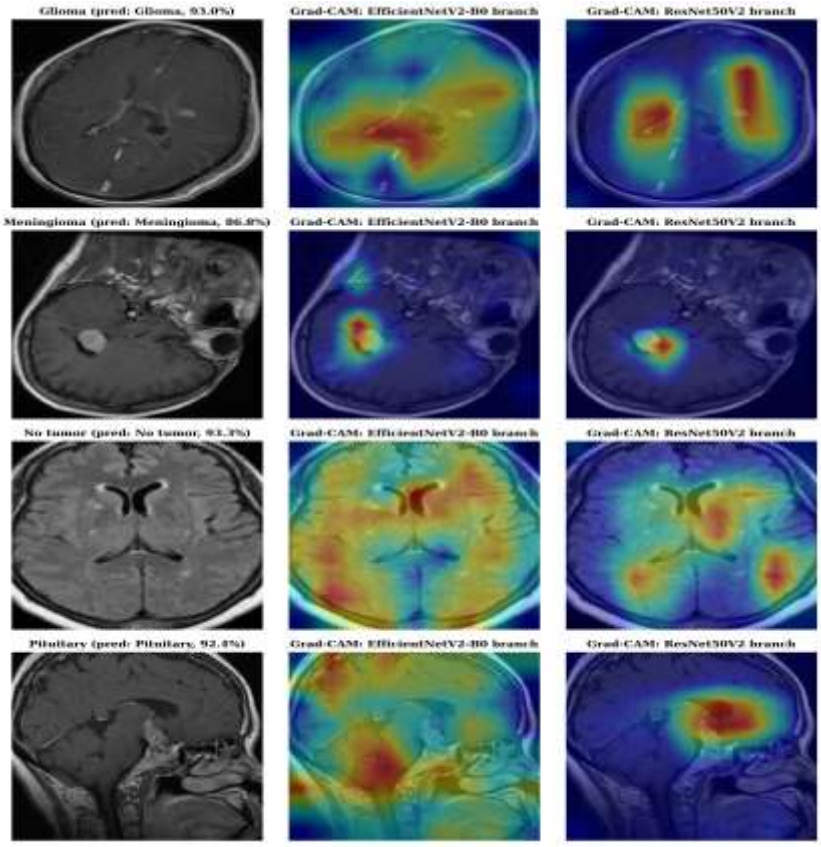


Figure 7. Grad-CAM explanations of the fusion model for one correctly classified test image per class, computed separately for the EfficientNetV2-B0 and ResNet50V2 branches.

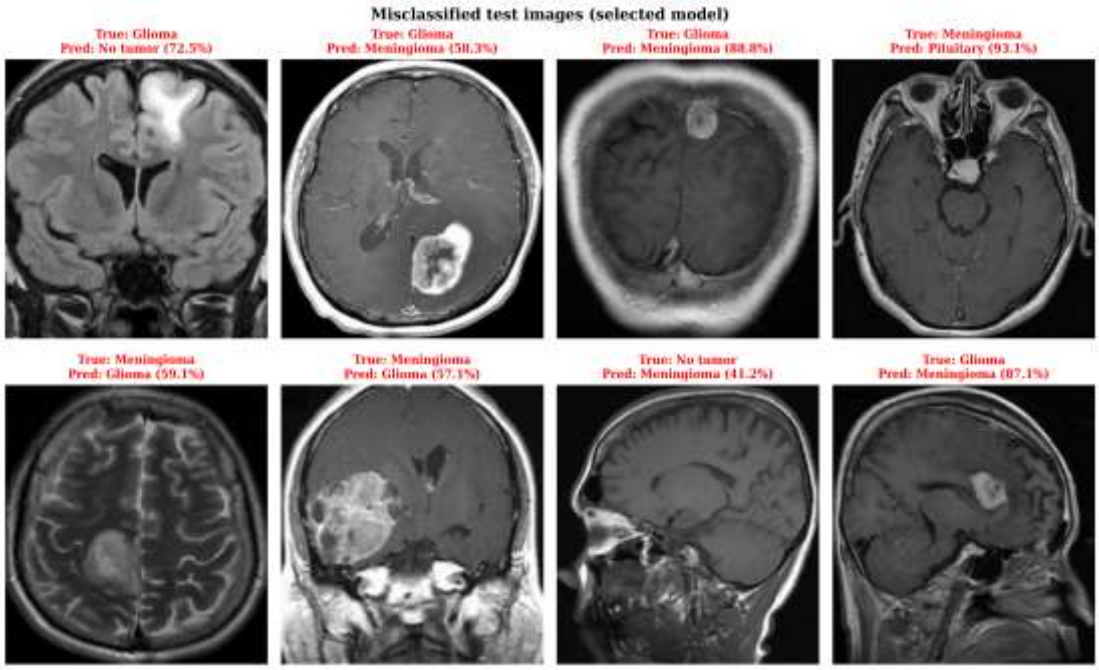


Figure 8. Examples of misclassified test images of the selected ensemble, with true labels, predicted labels and predicted probabilities.

DISCUSSION

The central finding of this study is that more than one third of the official test images of a widely used brain tumor MRI benchmark have a near-duplicate in its training set, even though no byte-identical copies remain. Accuracies measured on that split therefore partly reflect the recognition of previously seen slices rather than generalization, consistent with the record-wise versus subject-wise gap reported by Badža and Barjaktarović (2020) and with broader warnings about leakage (Kapoor and Narayanan, 2023; Varoquaux and Cheplygina, 2022).

Under the leakage-aware protocol the proposed models still reached 97–98% accuracy. Although lower than the figures often quoted for such data, these results are more credible, because near-duplicate groups never cross partitions and the test set was used only once. The bootstrap confidence intervals (about ±1 point) and the non-significant McNemar tests further show that small accuracy differences on a test set of this size should not be over-interpreted.

The ablation shows that feature fusion helps only when both branches are adequately trained: once EfficientNetV2-B0 was properly fine-tuned, the fusion model made the fewest errors. The ensemble was selected because it had the highest validation accuracy, although the fusion model scored 0.52 points higher on the test set. Keeping the pre-specified choice avoids post-hoc selection on the test set; as the difference corresponds to five images and is not significant, the lighter fusion model (31.2 M versus 62.4 M parameters) is an equally valid alternative. Clinically, only one of 740 tumor images was classified as normal, and most residual errors were confusions between tumor types, an error profile suitable for triage or second-reader support.

The study has limitations. Near-duplicate grouping is only a surrogate for patient-level splitting, and its thresholds were calibrated visually. All experiments used a single public data set of two-

dimensional slices without external validation, each model was trained with one random seed, Grad-CAM evidence is qualitative, and the ensemble roughly triples inference cost.

IMPLICATIONS TO RESEARCH AND PRACTICE

Implications for research. Studies that use the Brain Tumor MRI Dataset or similar merged collections should report a duplicate audit and use group-aware splits, because the official split alone does not guarantee independence. The fingerprinting procedure described here—MD5 for exact copies combined with perceptual hashing and pixel correlation for near copies—is simple, fast (about 2.4 minutes for 7,200 images with parallel processing) and transferable to other imaging benchmarks. Model selection should be pre-specified and based on validation data, and results should be accompanied by confidence intervals, chance-corrected metrics such as MCC and kappa, and paired significance tests. The release of split manifests would further allow other researchers to reproduce and compare results fairly.

Implications for practice. The proposed models identify tumor-bearing slices with very high sensitivity and distinguish the three tumor types with high accuracy, which makes them suitable candidates for decision support—for example, prioritising MRI studies for radiologist review or providing a second opinion on tumor type. However, they are not intended for stand-alone diagnosis. Deployment would require prospective, multi-centre validation, calibration of predicted probabilities, integration with full three-dimensional examinations and compliance with regulatory requirements. Where computational resources are limited, the single fusion model offers nearly identical accuracy at half the cost of the ensemble.

CONCLUSION

This paper presented a leakage-aware evaluation of brain tumor MRI classification. An audit based on exact and perceptual fingerprints showed that 36.25% of the official test images of a widely used benchmark have a near-duplicate in its training set. After de-duplication and a group-stratified split, a dual-backbone EfficientNetV2-B0 + ResNet50V2 feature-fusion network and a three-model ensemble with test-time augmentation were trained with a carefully designed two-phase transfer-learning recipe. The validation-selected ensemble achieved 97.80% accuracy (95% CI 96.86–98.64%), a macro-F1 of 0.978, an MCC of 0.971 and an AUC of 0.998 on 954 strictly unseen images, and missed only one of 740 tumors; the fusion model alone reached 98.32%. These results demonstrate that high and clinically promising performance remains attainable when evaluation is rigorous, and that transparent protocols are as important as model architecture for trustworthy medical AI.

FUTURE RESEARCH

Future work should first validate the models on external, multi-centre data sets that provide patient identifiers, enabling true patient-level splitting and an assessment of robustness to scanner and protocol differences. Second, extending the approach from two-dimensional slices to three-dimensional volumes and to multi-sequence inputs may resolve the remaining meningioma–glioma confusions. Third, repeating training with several random seeds and with nested or group-wise cross-validation would quantify variability more fully. Fourth, newer backbones such as vision transformers or ConvNeXt, as well as tumor-region guidance from segmentation, could further improve discrimination. Fifth, probability calibration and uncertainty estimation, for example through deep-ensemble variance (Lakshminarayanan, Pritzel and Blundell, 2017), would allow uncertain cases to be

referred to experts. Finally, quantitative evaluation of Grad-CAM maps against radiologist annotations and reader studies are needed to establish clinical usefulness.

DATA AND CODE AVAILABILITY

The Brain Tumor MRI Dataset (Version 2) analysed in this study is publicly available on Kaggle under a CC BY 4.0 licence at:

https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

The complete source code, including pre-processing, duplicate detection, model training and evaluation, is publicly available at:

https://www.kaggle.com/code/mdselimmia/leakage-aware-evaluation-of-a-dual-backbone-featur

The split manifest listing the train, validation and test assignment of every image, together with the training histories and evaluation outputs, is provided with the code to enable full reproduction of the reported results.

REFERENCES

Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A. and Mengko, T. R. (2019) *Brain tumor classification using convolutional neural network*, In World Congress on Medical Physics and Biomedical Engineering 2018, IFMBE Proceedings, Vol. 68/1, Springer, Singapore, pp. 183-189.

Afshar, P., Mohammadi, A. and Plataniotis, K. N. (2018) *Brain tumor type classification via capsule networks*, In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), IEEE, Athens, Greece, pp. 3129-3133.

Badža, M. M. and Barjaktarović, M. Č. (2020) *Classification of brain tumors from MRI images using a convolutional neural network*, Applied Sciences, 10(6), 1999.

Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S. and Kanchan, S. (2020) *Brain Tumor Classification (MRI)*, Kaggle dataset, https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z. and Feng, Q. (2015) *Enhanced performance of brain tumor classification via tumor region augmentation and partition*, PLoS ONE, 10(10), e0140381.

Chicco, D. and Jurman, G. (2020) *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation*, BMC Genomics, 21, 6.

Cohen, J. (1960) *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, 20(1), 37-46.

Deepak, S. and Ameer, P. M. (2019) *Brain tumor classification using deep CNN features via transfer learning*, Computers in Biology and Medicine, 111, 103345.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009) *ImageNet: A large-scale hierarchical image database*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Miami, USA, pp. 248-255.

Dietterich, T. G. (1998) *Approximate statistical tests for comparing supervised classification learning algorithms*, Neural Computation, 10(7), 1895-1923.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Hamada, A. (2020) *Br35H :: Brain Tumor Detection 2020*, Kaggle dataset, https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection.

He, K., Zhang, X., Ren, S. and Sun, J. (2016a) *Deep residual learning for image recognition*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, USA, pp. 770-778.

He, K., Zhang, X., Ren, S. and Sun, J. (2016b) *Identity mappings in deep residual networks*, In Computer Vision – ECCV 2016, Lecture Notes in Computer Science, Vol. 9908, Springer, Cham, pp. 630-645.

Ioffe, S. and Szegedy, C. (2015) *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, In Proceedings of the 32nd International Conference on Machine Learning (ICML), PMLR 37, pp. 448-456.

Kapoor, S. and Narayanan, A. (2023) *Leakage and the reproducibility crisis in machine-learning-based science*, Patterns, 4(9), 100804.

Kaufman, S., Rosset, S., Perlich, C. and Stitelman, O. (2012) *Leakage in data mining: Formulation, detection, and avoidance*, ACM Transactions on Knowledge Discovery from Data, 6(4), Article 15.

Kingma, D. P. and Ba, J. (2015) *Adam: A method for stochastic optimization*, In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, USA.

Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017) *Simple and scalable predictive uncertainty estimation using deep ensembles*, In Advances in Neural Information Processing Systems 30 (NeurIPS), Curran Associates, Long Beach, USA, pp. 6402-6413.

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

Loshchilov, I. and Hutter, F. (2017) *SGDR: Stochastic gradient descent with warm restarts*, In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France.

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A. and Ellison, D. W. (2021) *The 2021 WHO classification of tumors of the central nervous system: A summary*, Neuro-Oncology, 23(8), 1231-1251.

McNemar, Q. (1947) *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika, 12(2), 153-157.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G. and Wu, H. (2018) *Mixed precision training*, In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, Canada.

Nickparvar, M. (2021) *Brain Tumor MRI Dataset* (Version 2), Kaggle dataset, https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

Ostrom, Q. T., Price, M., Neff, C., Cioffi, G., Waite, K. A., Kruchko, C. and Barnholtz-Sloan, J. S. (2023) *CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2016-2020*, Neuro-Oncology, 25(Supplement 4), iv1-iv99.

Raghu, M., Zhang, C., Kleinberg, J. and Bengio, S. (2019) *Transfusion: Understanding transfer learning for medical imaging*, In Advances in Neural Information Processing Systems 32 (NeurIPS), Curran Associates, Vancouver, Canada, pp. 3347-3357.

Rehman, A., Naz, S., Razzak, M. I., Akram, F. and Imran, M. (2020) *A deep learning-based framework for automatic brain tumors classification using transfer learning*, Circuits, Systems, and Signal Processing, 39(2), 757-775.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L. et al. (2021) *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans*, Nature Machine Intelligence, 3, 199-217.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) *Grad-CAM: Visual explanations from deep networks via gradient-based localization*, In Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, pp. 618-626.

Shorten, C. and Khoshgoftaar, T. M. (2019) *A survey on image data augmentation for deep learning*, Journal of Big Data, 6, 60.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) *Dropout: A simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research, 15, 1929-1958.

Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S. and Lu, J. (2019) *Brain tumor classification for MR images using transfer learning and fine-tuning*, Computerized Medical Imaging and Graphics, 75, 34-46.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) *Rethinking the Inception architecture for computer vision*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, USA, pp. 2818-2826.

Tan, M. and Le, Q. V. (2019) *EfficientNet: Rethinking model scaling for convolutional neural networks*, In Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97, pp. 6105-6114.

Tan, M. and Le, Q. V. (2021) *EfficientNetV2: Smaller models and faster training*, In Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139, pp. 10096-10106.

Varoquaux, G. and Cheplygina, V. (2022) *Machine learning for medical imaging: Methodological failures and recommendations for the future*, npj Digital Medicine, 5, 48.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) *How transferable are features in deep neural networks?*, In Advances in Neural Information Processing Systems 27 (NeurIPS), Curran Associates, Montréal, Canada, pp. 3320-3328.

Zauner, C. (2010) *Implementation and Benchmarking of Perceptual Image Hash Functions*, Master's thesis, Upper Austria University of Applied Sciences, Hagenberg, Austria.