IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151198

Brain Tumor MRI Classification Using ML and Medical Imaging

Manisha S. Patil[1], Saurabh D. Badgujar[2], Dhanashri P. Chavhan[3], Pranav N. Chopade[4], Riya M. Deshmukh[5]

Associate Professor, Department of Computer Science (Data Science),

R. C. Patel Institute of Technology, Shirpur, Maharashtra, India – 425405[1]

UG Students, Department of Computer Science (Data Science),

R. C. Patel Institute of Technology, Shirpur, Maharashtra, India – 425405[2-5]

*Corresponding author:saurabhbadgujar851@gmail.com

Abstract: Doctors often turn to brain magnetic resonance imaging (MRI) when an abnormality is suspected, but going through large numbers of scans manually takes time, and tumor types that look alike can lead to inconsistent readings. This study evaluates MobileNetV2, Swin-Tiny and EfficientNetV2-S, each pretrained on ImageNet, for sorting MRI images into four classes: glioma, meningioma, pituitary tumor and no tumor. Every model is trained with the same data split and the same two phase transfer learning procedure, built in PyTorch with the timm library. The dataset holds 7,200 images. Of these, 4,480 train the models, 1,120 support validation and 1,600 are set aside for the final test. Validation results decide which model is chosen, and the test set is used only after that choice. On validation data, EfficientNetV2-S achieved the best macro F1 score (0.9786), followed by Swin-Tiny (0.9556) and MobileNetV2 (0.9095). On the held out test set, EfficientNetV2-S reached 93.56% accuracy and 93.43% macro F1. Test recall by class was 79.75% for glioma, 96.00% for meningioma, 100.00% for no tumor and 98.50% for pituitary tumor, so glioma is the principal weakness. These results are encouraging for this dataset. The system remains a research prototype that has not been validated externally, and it is not meant to replace clinical assessment.

Keywords: Brain tumor classification, magnetic resonance imaging, transfer learning, EfficientNetV2-S, Swin Transformer, MobileNetV2, deep learning.

I. INTRODUCTION

A brain tumor forms when cells within or beside the brain grow in an uncontrolled way. Tumors differ greatly in position, size and appearance on scans. MRI shows soft tissue in fine detail, so it is usually among the first investigations ordered when a brain lesion is suspected. Even so, a radiologist who reviews hundreds of slices by eye becomes fatigued, and one slice from two different tumor types can look nearly the same. Software that groups images according to visual patterns can therefore give the radiologist a second reading.

Convolutional neural networks (CNNs) and, more recently, vision transformers learn image features directly from pixel values instead of depending on descriptors designed by hand. Transfer learning takes a network that has already learned from a large collection of natural images, such as ImageNet, and adapts it to a smaller medical dataset, which reduces the cost of training. A high score on one collection says little about performance with other scanners, hospitals or patient groups, though. One accuracy number can also conceal poor results on a class that matters clinically.

In this paper we assign brain MRI images to four categories: no tumor, glioma, meningioma and pituitary tumor. Three pretrained networks from different families are compared. They are a compact CNN (MobileNetV2), a CNN scaled by a compound rule (EfficientNetV2-S) and a hierarchical vision transformer (Swin-Tiny). The data partitions, augmentation and optimizer settings are identical for all three. We make sure the test images play no role in picking a checkpoint or ranking the networks, so the final figures describe images the models have never encountered.

The main contributions are as follows:

- A controlled comparison of three architectures from different families, run under one documented two phase transfer learning protocol and a split that avoids data leakage.
- Precision, recall and F1 reported for every class alongside overall accuracy, which exposes uneven performance that aggregate metrics hide.
- An error analysis of glioma, the weakest class, supported by a confusion matrix and Grad-CAM visualizations.
- A discussion of the difference between validation and test performance, and of the limits of the dataset.

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151198

The remaining sections are arranged as follows. Section II reviews related work. Section III explains the dataset and methods. Section IV describes the experimental setup. Section V presents the results. Section VI discusses the findings and compares them with existing studies. Section VII lists limitations, and Section VIII concludes the paper.

II. RELATED WORK

Solanki et al. [1] provide an overview of machine learning and deep learning methods applied to brain tumor detection and classification. Their review includes CNN pipelines, transfer learning from standard backbones, 3D networks and older classifiers such as SVM and k-NN. They point out that studies are hard to compare, since datasets, imaging types, preprocessing steps and evaluation rules are not the same. They also name limited data, heavy computation and the diversity of tumors as open problems.

Chattopadhyay and Maitra [2] describe a CNN that detects tumors in 2D MRI slices, with 99.74% accuracy under their own experimental conditions. Abdusalomov et al. [3] choose an improved YOLOv7 object detector instead. Their modifications include transfer learning, image enhancement, a Convolutional Block Attention Module (CBAM), an SPPF+ layer and a bidirectional feature pyramid network. They report 99.5% accuracy in locating glioma, meningioma and pituitary tumors among images that show no tumor. A detector must also state where the tumor is, so detection and image level classification are not the same task.

Considered together, these works show that automated MRI analysis is achievable. A high number in a paper does not prove that a model generalizes or is ready for clinical use, however. Since datasets, label definitions, preprocessing and evaluation rules differ, lining up published percentages to rank them is unreliable.

TABLE I. SUMMARY OF RELATED WORK ON BRAIN TUMOR MRI CLASSIFICATION

Reference	Approach and Task	Reported Result / Contribution	Relevant Limitation
Solanki et al. [1]	Survey of ML, CNN, transfer learning, and segmentation methods	Reviews ~100 studies; summarizes methods, datasets, limitations, research directions	A review, not a single controlled experiment; no new model accuracy to compare
Chattopadhyay & Maitra [2]	CNN-based segmentation/detection using 2D MRI images	Reports 99.74% accuracy in its experimental setting	Task and protocol differ; accuracy is not directly comparable
Abdusalomov et al. [3]	Enhanced YOLOv7 for tumor object detection and localization	Reports 99.5% accuracy; uses attention and multi-scale feature fusion	Object detection / localization differs from image-level 4-class classification
Present work	Transfer learning comparison: MobileNetV2, Swin-Tiny, EfficientNetV2-S	EfficientNetV2-S: 93.56% test accuracy; 93.43% test macro F1	Uses one Kaggle dataset; external-dataset and clinical validation remain future work

III. MATERIALS AND METHODS

A. Dataset

The study uses the Brain Tumor MRI Dataset that M. Nickparvar published on Kaggle [4]. To our knowledge it was compiled from earlier public sources, including figshare, SARTAJ and Br35H. The copy we used has 7,200 images spread evenly over four classes: glioma, meningioma, no tumor and pituitary. The Training folder contains 1,400 images per class, 5,600 in total, and the Testing folder contains 400 per class, 1,600 in total. The Kaggle listing was last updated on 13 February 2026 and carries a CC BY 4.0 license. All images are RGB files of varying resolution. An automated script checked that each file was intact and removed any that could not be read.

- The Kaggle listing was last updated on **13 Feb 2026**, and the license is **CC BY 4.0**.
- The counts are exactly 1,400 per class in Training and 400 per class in Testing. This resolves the version question I raised earlier, so state the update date.
- The download date in the log is 11 Sep 2026. Use that date, or the V3 notebook's download date, as the access date in reference [4].

All images are RGB files at varying resolutions. An automated script verified image integrity and removed unreadable files

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151198

B. Preprocessing and Data Split

Each image is resized, center cropped to 224 × 224 pixels, converted into a tensor and normalized with the ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). Images enter the networks as three channel inputs so that they match the pretrained input layers. All architectures share the same 224 × 224 input size for consistency, although EfficientNetV2-S was originally designed for larger resolutions.

Stratified 80/20 sampling divided the Training folder into training and validation subsets. The Testing folder was kept as the reserved test set and was evaluated once, after the final model had been selected. The dataset has no patient identifiers, so the split is made at image level. Slices from the same patient may therefore appear in more than one subset (Section VII).

TABLE I. DATASET PARTITION

Subset	Number of Images	Per Class	Purpose
Training	4,480	1,120	Fit model parameters
Validation	1,120	280	Compare architectures, monitor metrics, select checkpoint
Testing	1,600	400	Final evaluation after model selection
Total	7,200	1,800	—

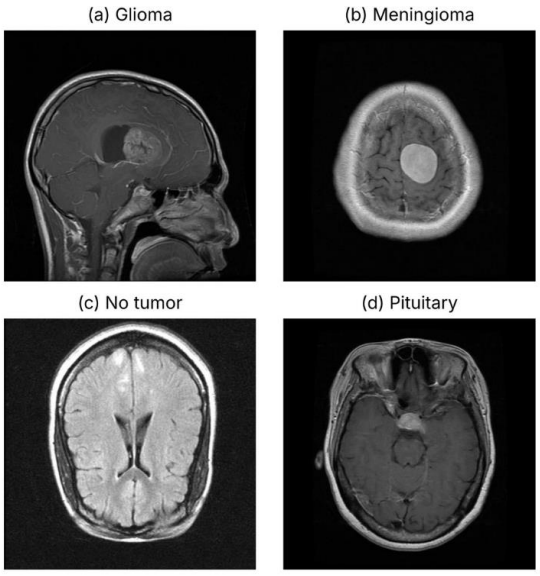


Figure 1. Representative mri slices from the four classes of the brain tumor mri dataset [4] (Testing folder): (a) glioma, (b) meningioma, (c) no tumor, and (d) pituitary tumor.

C. Data Augmentation

Only training images are augmented. The changes are made on the fly and consist of a random horizontal flip and a random rotation of up to ±15°. Validation and test images are only resized and normalized in a fixed way, so randomness never affects the evaluation scores.

D. Network Architectures

Three architectures pretrained on ImageNet are compared. In each one the original classification layer is replaced with a linear layer that outputs four class scores.

MobileNetV2 [5] lowers computation by using depthwise separable convolutions and inverted residual blocks that end in linear bottlenecks. In this study it acts as the small, inexpensive reference model.



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151198

EfficientNetV2-S [6] blends Fused-MBConv and MBConv blocks. It was found through a training aware architecture search and grows depth, width and resolution together.

Swin-Tiny [7] is a hierarchical vision transformer. Its self attention operates inside shifted local windows, so it captures both nearby and distant relationships while its computational cost rises only linearly with image size. It represents the transformer family here.

TABLE. MODEL COMPLEXITY

E. Two-Phase Training Strategy

All three architectures were trained with identical settings. In Phase 1 the pretrained backbone stays frozen, and only the new head is trained for 20 epochs. This phase uses Cross Entropy Loss, AdamW [10] with a learning rate of 1×10^{-3}, a batch size of 32 and CosineAnnealingLR [11]. In Phase 2 the backbone and head are both unfrozen and fine tuned together for 10 epochs, using AdamW with a learning rate of 1×10^{-4}. For each model the checkpoint with the highest validation macro F1 is kept. A fixed random seed of 42 is applied.

F. Evaluation Metrics

Performance is reported as accuracy, precision, recall and F1 for each class, macro F1, and the confusion matrix:

$$Precision = TP / (TP + FP), \quad Recall = TP / (TP + FN),$$
$$F1 = 2 \times Precision \times Recall / (Precision + Recall).$$

Macro F1 is the plain average of the four class F1 values, with each class weighted equally. It is helpful for medical images, because overlooking one tumor type can be serious even when overall accuracy appears high.

IV. Experimental Setup and Implementation

All models were coded in Python 3.10 with PyTorch, timm [8], torchvision, scikit-learn and OpenCV. Training ran on Google Colab using an NVIDIA Tesla T4 GPU. Each backbone received the same data partitions, preprocessing, augmentation and optimizer settings (Table III). Checkpoints were stored according to validation macro F1, and the best checkpoint of every model was loaded for evaluation.

The procedure had eight steps: (1) load the dataset and inspect it; (2) divide the Training folder into stratified training and validation parts; (3) resize, normalize and augment the training images; (4) load each pretrained backbone and replace its classifier; (5) train only the head (Phase 1); (6) fine tune the entire network (Phase 2); (7) select the model with the best validation macro F1; and (8) score that checkpoint once on the reserved test set.

Early experiments used the test folder to select models, which can bias the final estimate. The log from that stage listed 5,600 training files, 1,600 validation files and 175 steps per epoch (5,600 ÷ 32). The 1,600 files labeled as validation were actually the Testing folder, and this finding led to the three way split. The present protocol keeps validation data for checkpoint choice and architecture comparison only.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151198

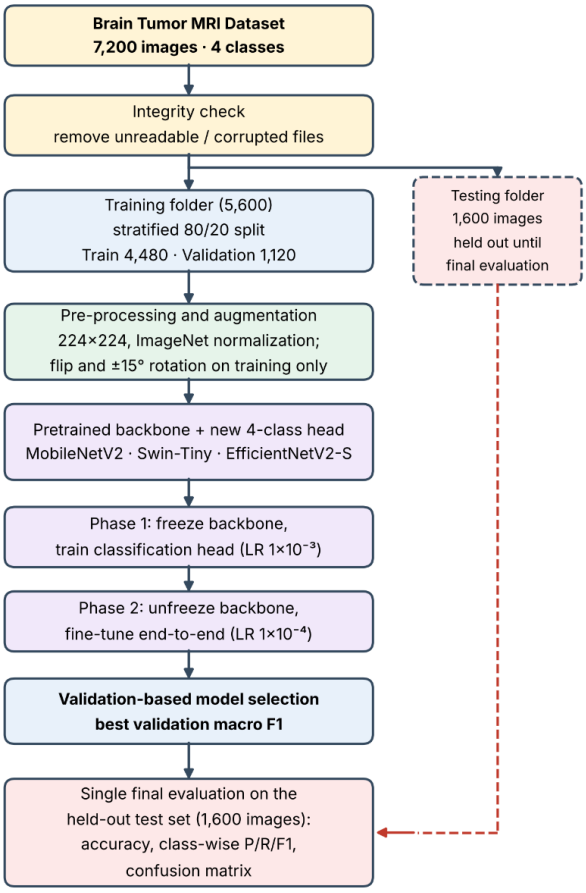


Figure 2 Overall methodology: from data preparation to model selection and test evaluation

TABLE III. TRAINING CONFIGURATION

Setting	Value
Input size	224 × 224 × 3
Batch size	32
Optimizer	AdamW
Learning rate, Phase 1	1×10^{-3} (frozen backbone)
Learning rate, Phase 2	1×10^{-4} (fine-tuning)
LR Scheduler	CosineAnnealingLR
Loss function	Cross Entropy Loss
Random seed	42
Classifier head	Linear(backbone_features, 4)
Framework	PyTorch + timm

V. RESULTS

A. Architecture Comparison on the Validation Set

EfficientNetV2-S obtained the highest validation macro F1 (0.9786). Swin-Tiny followed with 0.9556 and MobileNetV2 with 0.9095. EfficientNetV2-S was therefore selected for final testing. This outcome applies to the dataset, preprocessing and training conditions used here. It does not show that EfficientNetV2-S would outperform the other models on every MRI dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151198

TABLE IV. ARCHITECTURE COMPARISON ON VALIDATION SET

Model	Model Type	Val. Accuracy (approx.)	Val. Macro F1
MobileNetV2	Lightweight CNN	~91%	0.9095
Swin-Tiny	Vision Transformer	~95%	0.9556
EfficientNetV2-S	Compound-scaled CNN	~98%	0.9786

B. Final Test Performance

On the reserved test set of 1,600 images, EfficientNetV2-S achieved 93.56% accuracy and 93.43% macro F1. With 1,600 images, the 95% confidence interval for an accuracy of 93.56% is roughly ±1.2 percentage points. Small differences between models should therefore be interpreted with caution.

A preliminary MobileNetV2 model built in Keras and TensorFlow (two training phases, 2.42 M parameters) reached 87.06% accuracy and 0.8668 macro F1 on the original Testing folder. Its weakest classes were glioma (recall 0.790) and meningioma (recall 0.7225), while no tumor and pituitary both reached 0.985 recall. That run also used the Testing folder to monitor training. Its scores are therefore optimistic and cannot be compared with the main experiments, whose protocol prevents leakage.

TABLE V. TEST PERFORMANCE COMPARISON

Metric	MobileNetV2 Baseline	EfficientNetV2-S (Final)	Difference
Test accuracy	87.06%	93.56%	+6.50 pp
Test macro F1	86.68%	93.43%	+6.75 pp
Validation macro F1	-	97.86%	—

C. Analysis by Class

No tumor and pituitary images were recognized with perfect or almost perfect recall, and meningioma reached 96.00%. Glioma was the most difficult class, with recall of 79.75% (319 of 400), so roughly one in five glioma images (81) was misclassified.

TABLE VI. TEST RESULTS FOR EACH CLASS: EFFICIENTNETV2-S

Class	Test Recall	Interpretation
Glioma	79.75%	Lowest recall; errors include confusion with meningioma
Meningioma	96.00%	High recall in this test set
No Tumor	100.00%	All actual no-tumor test images correctly identified
Pituitary	98.50%	High recall in this test set

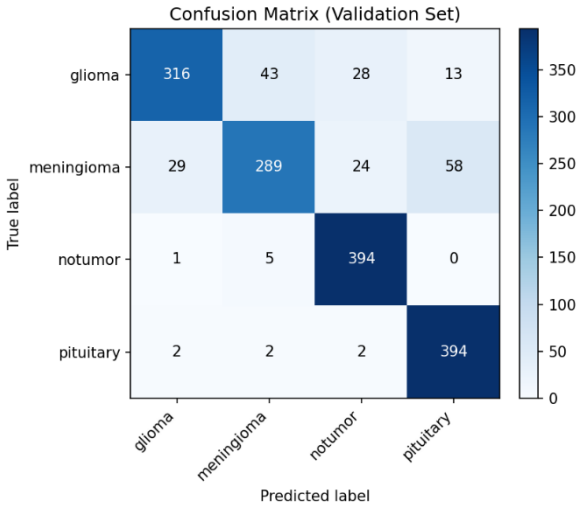


Figure 3 Confusion matrix of EfficientNetV2-S on the 1,600-image test set (G: Glioma, M: Meningioma, N: No Tumor, P: Pituitary).



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151198

The confusion matrix of EfficientNetV2-S (Fig. 4) records 1,497 correct predictions out of 1,600 (93.56%) and 103 errors. Glioma accounts for 81 of them (recall 0.7975), which is about 79% of all mistakes, while meningioma contributes 16 (recall 0.9600) and pituitary tumor contributes 6 (recall 0.9850). No tumor images were all identified correctly (400 of 400). The most frequent confusion was glioma predicted as **meningioma** (**43** images), followed by glioma predicted as **notumor** (**28**). The remaining **13** glioma errors went to **pituitary**. Pituitary precision was **0.8473** (or **84.73%**) and glioma precision was **0.9080** (or **90.80%**). Clinically, **54** of the 1,200 tumor images (**4.5%**) were predicted as no tumor, and **28** of them were glioma. Compared with the baseline, EfficientNetV2-S removed most meningioma errors, but glioma remains the clear weak point.

VI. DISCUSSION

A. Main Findings

The most important finding is the distance between glioma recall (79.75%) and recall for the other classes (96.00% to 100.00%). High overall accuracy can sit alongside a real weakness on one clinically relevant class, which is why results must be reported for each class. Possible causes include visual overlap with other tumor types, differences in how tumors look, image quality and inconsistent labels.

B. Gap Between Validation and Test Performance

EfficientNetV2-S reached a validation macro F1 of 0.9786 but a test macro F1 of 0.9343, a gap of about 4.4 points. Two explanations are plausible, and they do not exclude each other. First, the validation set is a random subset of the Training folder, so almost identical slices from one patient may be present in both training and validation, which would make validation scores optimistic. Second, the Testing folder may differ from the training data in distribution or difficulty. No patient identifiers exist, so the present data cannot separate the two explanations. The test score should therefore be regarded as the more dependable estimate, and validation should serve only for model selection.

C. Comparison with Existing Work

This study differs from the reviewed publications in how the task is defined and how experiments are designed. Chattopadhyay and Maitra [2] report 99.74% accuracy for a CNN based method, and Abdusalomov et al. [3] report 99.5% for an enhanced YOLOv7 detection system. These values only summarize the cited papers. They do not claim that those models are better or worse than the present classifier, because the datasets, tasks, preprocessing and evaluation protocols are different.

VII. LIMITATIONS

This study has several limitations that need to be acknowledged:

- **Evaluation data.** All results come from one curated Kaggle dataset, and no external or multi institution testing was performed. Other versions of the dataset contain different numbers of images.
- **Possible patient level leakage.** Patient identifiers are not available, so slices from one patient may appear in the training, validation and test sets, which may inflate scores.
- **Possible source specific shortcuts.** The dataset combines several sources, so the model may partly rely on image traits of a source instead of tumor anatomy.
- **Acquisition variability.** Scans from several hospitals or scanner models were not included. Differences in field strength, imaging protocol, slice thickness or patient mix could therefore lower accuracy.
- **Statistical evidence.** The confidence interval on the test set is roughly ±1.2 points.
- **Glioma sensitivity.** A glioma recall of 79.75% needs further work before any deployment can be considered.
- **Scope.** Only four image level classes on 2D slices are covered. Grading, segmentation, 3D context, multi sequence MRI and coexisting pathologies are outside the scope.
- **Clinical use.** This is a research prototype that has not been tested prospectively. It must not guide clinical decisions or stand in for trained medical staff.

VIII. CONCLUSION

This paper compared MobileNetV2, Swin-Tiny and EfficientNetV2-S for four class brain MRI classification. All models followed one training, validation and test protocol that avoids data leakage, together with a two phase transfer learning procedure. EfficientNetV2-S gave the best validation macro F1 (0.9786) and reached 93.56% accuracy and 93.43% macro F1 on the reserved test set of 1,600 images. Keeping training, validation and test data clearly separated



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151198

gives a more defensible evaluation than choosing models on the test set. The main unresolved issue is the glioma recall of 79.75%.

Future work includes: (i) evaluation on independent datasets from multiple institutions, using splits made by patient; (ii) focal loss, class weighted training or targeted augmentation for glioma; (iii) larger architectures and higher input resolutions, such as bigger EfficientNetV2 variants or DeiT; (iv) three dimensional and multi sequence MRI; and (v) expert radiologist review and rigorous external validation before any clinical use.

REFERENCES

[1] S. Solanki, U. P. Singh, S. S. Chouhan, and S. Jain, "Brain tumor detection and classification using intelligence techniques: An overview," *IEEE Access*, vol. 11, pp. 12870–12886, 2023, doi: 10.1109/ACCESS.2023.3242666.

[2] A. Chattopadhyay and M. Maitra, "MRI-based brain tumour image detection using CNN based deep learning method," *Neuroscience Informatics*, vol. 2, no. 4, Art. 100060, 2022, doi: 10.1016/j.neuri.2022.100060.

[3] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain tumor detection based on deep learning approaches and magnetic resonance imaging," *Cancers*, vol. 15, no. 16, Art. 4172, 2023, doi: 10.3390/cancers15164172.

[4] M. Nickparvar, "Brain Tumor MRI Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF CVPR*, 2018, pp. 4510–4520.

[6] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. ICML*, 2021, pp. 10096–10106.

[7] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF ICCV*, 2021, pp. 10012–10022.

[8] R. Wightman, "PyTorch Image Models (timm)," GitHub, 2019. [Online]. Available: https://github.com/huggingface/pytorch-image-models.

[9] J. Cheng *et al.*, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PLoS ONE*, vol. 10, no. 10, Art. e0140381, 2015.

[10] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[11] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2017.

 This work is licensed under a Creative Commons Attribution 4.0 International License