Brain Tumor Multiclass Classification from MRI using CNNs and Model Interpretability with Grad-CAM

Shammi Singh
Technische Hochschule Ulm (THU)
M.Sc. Intelligent Systems
Advanced Machine Learning Project
singsh01@thu.de
https://github.com/Singhsh01/brainTumorCNN-with-grad-Cam

***Abstract*—Magnetic resonance imaging (MRI) is widely used for brain tumor assessment, but manual interpretation is time-consuming and subject to variability. This project implements an end-to-end deep learning workflow for *four-class* brain MRI classification (glioma, meningioma, pituitary tumor, and no tumor). A preprocessing pipeline removes large non-informative black regions via contour-based cropping, standardizes image size, and normalizes pixel values. A transfer-learning model based on ResNet-50 is fine-tuned for multiclass classification and evaluated with and without data augmentation (augmentation applied to the training split only). Beyond accuracy, model interpretability is studied using Grad-CAM and Guided Grad-CAM to visualize image regions that contribute most to predictions. Both pipelines achieve high test accuracy (approximately 98.8%), and Grad-CAM examples indicate that the model often focuses on anatomically plausible regions for tumor classes while highlighting normal-structure evidence for the no-tumor class.**

***Index Terms*—Brain MRI, tumor classification, convolutional neural networks, transfer learning, ResNet-50, data augmentation, Grad-CAM, interpretability.**

I. Introduction

Brain tumors are abnormal growths of cells in the brain or surrounding tissues and pose a serious threat to neurological function and patient survival. Common tumor categories include glioma (often aggressive), meningioma (frequently benign and arising from the meninges), pituitary tumors (affecting hormonal regulation), and the absence of tumor (healthy tissue) [2], [3].

Magnetic Resonance Imaging (MRI) is widely used for brain tumor diagnosis due to its high soft-tissue contrast and non-invasive nature [2]. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have demonstrated strong performance in image analysis by automatically learning discriminative features from raw pixel data [8]. CNN-based approaches are increasingly applied to medical imaging tasks such as MRI classification [1].

However, medical imaging datasets are often limited in size and may contain acquisition artifacts, intensity variations, and large non-informative background regions. These factors can harm generalization and can lead to models relying on spurious correlations [10]. This work addresses these issues through a reproducible preprocessing pipeline (contour-based cropping + standardization) and evaluates generalization via two comparable training pipelines: (i) *no augmentation* and (ii) *augmentation applied to the training data only* [12].

Finally, interpretability is incorporated using Gradient-weighted Class Activation Mapping (Grad-CAM), enabling visual explanations of the model's decision process and increasing transparency in a medical context [6], [11].

II. Dataset

The dataset contains T1-weighted brain MRI images with four labels (glioma, meningioma, pituitary, and no tumor) [3]. After preprocessing and cropping, the total dataset size is 7023 images. A fixed test split is used, and the remaining training data is split into training and validation subsets. Using a held-out test split is important for unbiased evaluation of generalization performance [4].

TABLE I
DATASET SPLIT SIZES (AFTER PREPROCESSING/CROPPING).

Split	Images	Shape	Classes
Train	4569	$200 \times 200 \times 3$	4
Validation	1143	$200 \times 200 \times 3$	4
Test	1311	$200 \times 200 \times 3$	4

TABLE II
CLASS DISTRIBUTION IN THE TEST SET (COUNTS).

Class	Test Count
Glioma	300
Meningioma	306
No tumor	405
Pituitary	300

Note on imbalance: The dataset shows moderate class imbalance (e.g., the no-tumor class has the highest count). The same splits are kept across both pipelines to ensure a fair comparison [9].

III. PREPROCESSING PIPELINE

Many MRI images contain large black borders and non-informative background that do not contribute to learning and may mislead CNN feature extraction by encouraging reliance on irrelevant pixels [4]. Therefore, a cropping routine is applied before model training.

A. Contour-based Cropping

Cropping is performed by: (1) converting to grayscale, (2) applying Gaussian blur, (3) thresholding and morphological cleanup (erode/dilate), (4) selecting the largest external contour, and (5) cropping the image using the contour extreme points (left/right/top/bottom). This reduces irrelevant areas and standardizes the region-of-interest (ROI) around the head anatomy, which can improve robustness by focusing representation learning on informative regions [1].

B. Resizing and Normalization

All cropped images are resized to 200×200 pixels to match the ResNet-50 input size. Images are normalized to the range $[0, 1]$ by dividing by 255.0, which stabilizes optimization by keeping input magnitudes bounded [4]. Labels are encoded and converted to one-hot vectors for multiclass classification with categorical cross-entropy [4].

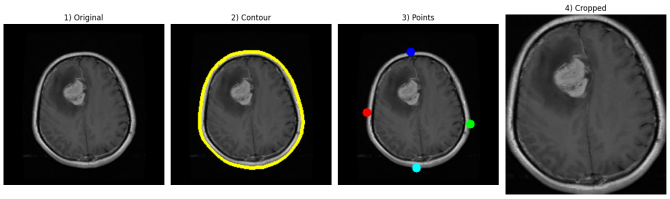


Fig. 1. Contour-based cropping pipeline. From left to right: (1) original MRI, (2) detected largest contour, (3) extreme points used to define the ROI, (4) final cropped image.

C. Why an Image-Resolution Scatter Check Matters

A resolution scatter (width vs. height) is used as a sanity check to detect extreme or inconsistent image sizes. Large variations can indicate corrupted files or inconsistent acquisition/export settings and may introduce unintended scale bias [10]. Confirming consistent resolutions after cropping improves reproducibility and stabilizes training.

IV. DATA AUGMENTATION STRATEGY (TRAIN ONLY)

MRI datasets are typically limited, which increases overfitting risk. Data augmentation is therefore applied *only to the training split* (never to validation/test) to improve generalization [12].

In medical imaging, augmentation must remain anatomically plausible to avoid introducing label-breaking transformations [12]. For this project, conservative augmentation is used: small rotations, slight translations (width/height shifts), and moderate zoom. Horizontal flipping is disabled because left-right inversion can be anatomically unsafe for brain MRIs and may confuse hemispheric cues [12].

V. MODEL ARCHITECTURE AND TRAINING SETUP

A. ResNet-50 Transfer Learning

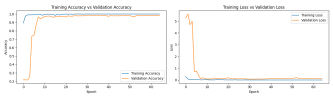
ResNet-50 is used as a convolutional backbone pretrained on ImageNet [5]. Residual connections mitigate optimization difficulties in deep networks by enabling more stable gradient flow [5]. Transfer learning is commonly used when labeled medical imaging data is limited, since pretrained features can be adapted to the target domain [4].

The final classification head is replaced by: GlobalAveragePooling2D, Dropout($p = 0.4$), and a Dense(4, softmax) layer. Dropout acts as a regularizer and can reduce overfitting by discouraging co-adaptation of features [14]. All layers are set trainable (fine-tuning) to adapt the feature hierarchy to MRI domain patterns.

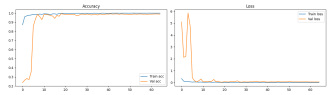
The model summary indicates approximately 23.6M parameters (total parameters: 23,595,908; trainable parameters: 23,542,788).

TABLE III
SUMMARY OF MODEL PERFORMANCE ACROSS PIPELINES.

Pipeline	Metric	Value
No augmentation	Final Train Accuracy	1.0000
	Final Validation Accuracy	0.9816
	Best Validation Accuracy	0.9851
	Best Validation Loss	0.0767
Train-only augmentation	Test Accuracy	0.9878
	Test Loss	0.0357



(a) No augmentation: training vs. validation.



(b) Train-only augmentation: training vs. validation.

Fig. 2. Training and validation learning curves for both pipelines.

B. Optimization and Callbacks

Training uses Adam optimizer with learning rate 1×10^{-4} and categorical cross-entropy loss [4], [7]. The validation split is 20% of the training data.

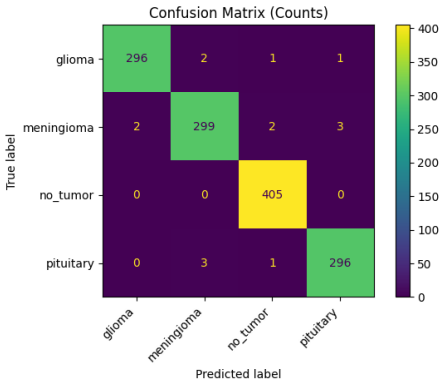
Callbacks include: (i) EarlyStopping on validation loss with patience 40 and restoring best weights, (ii) ModelCheckpoint saving the best validation-loss model, and (iii) ReduceLROnPlateau to lower learning rate when validation loss plateaus. A large early-stopping patience allows convergence even if improvements are small; best weights are preserved by checkpointing, which is a common practical safeguard when validation loss improvements are incremental [4].

VI. RESULTS

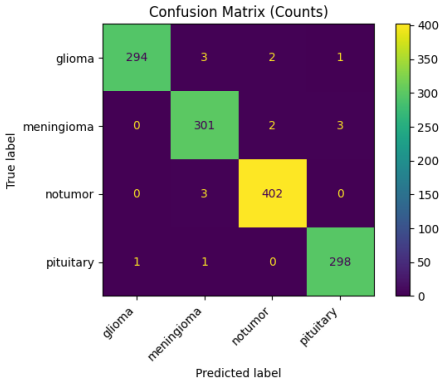
Two pipelines are evaluated: **(A) No-Augmentation** and **(B) Train-only Augmentation**. Both achieve high accuracy on the held-out test set.

A. Quantitative Performance

Observation: Both pipelines generalize strongly. Training accuracy approaching 1.0 while validation accuracy stabilizes slightly lower is consistent with mild overfitting for large fine-tuned CNNs on limited datasets [4]. Validation loss can plateau or oscillate; checkpointing ensures the best validation-loss model is retained.



(a) No augmentation: confusion matrix (test set).



(b) Train-only augmentation: confusion matrix (test set).

Fig. 3. Confusion matrices on the held-out test set for both pipelines. Most errors occur between visually similar tumor classes; the `no_tumor` class remains strongly separable in both settings.

B. Classification Performance

C. Metrics and Error Analysis

To understand class-wise performance beyond accuracy, precision, recall, and F1-score are reported [9]. Confusion matrices provide a direct view of which classes are most frequently confused and help diagnose systematic errors [9].

VII. INTERPRETABILITY WITH GRAD-CAM

High accuracy alone is insufficient in medical contexts; interpretability helps validate whether predictions depend on meaningful regions [6]. Grad-CAM generates a class-discriminative heatmap by weighting feature maps in the last convolutional layer using gradients of the class score [11].

TABLE IV
CLASSIFICATION PERFORMANCE ON THE TEST SET FOR BOTH PIPELINES.

Pipeline	Class	Precision	Recall	F1-score	Support
No augmentation	Glioma	0.9933	0.9867	0.9900	300
	Meningioma	0.9836	0.9771	0.9803	306
	No tumor	0.9902	1.0000	0.9951	405
	Pituitary	0.9867	0.9867	0.9867	300
Train-only augmentation	Glioma	0.9966	0.9800	0.9882	300
	Meningioma	0.9773	0.9837	0.9805	306
	No tumor	0.9901	0.9926	0.9914	405
	Pituitary	0.9868	0.9933	0.9900	300

Overall accuracy (test set): No augmentation = 0.9886 (1311 samples), Train-only augmentation = 0.9878 (1311 samples).

A. How to Read the Heatmaps

Warm colors (red/yellow) indicate stronger contribution to the predicted class score, while cool colors indicate weaker contribution [11]. Grad-CAM does *not* prove causality or provide tumor boundaries; it indicates which spatial regions most influenced the model decision [11].

B. Guided Backpropagation and Guided Grad-CAM

Guided backpropagation produces sharper, pixel-level saliency by modifying gradient flow through ReLU units [13]. Guided Grad-CAM combines Grad-CAM with guided backpropagation, often producing sharper, edge-like patterns that emphasize fine structures contributing to a prediction [11], [13]. Such saliency-style explanations should be interpreted carefully, as they explain the model's sensitivity rather than medical ground truth [6].

C. Insights from Qualitative Examples

For tumor classes, heatmaps frequently highlight localized regions that align with abnormal intensity patterns. For the no-tumor class, heatmaps may still appear because the model uses evidence of normal anatomical texture/symmetry to support the decision "no tumor" (i.e., explaining the predicted score, not locating a tumor) [11].

VIII. LIMITATIONS AND FUTURE WORK

Dataset constraints: The dataset size and class imbalance may bias learning and limit generalization to other scanners or protocols [10]. **Clinical validity:** Grad-CAM provides model explanations, not clinical ground truth; expert review would be required for medical claims [6]. Future work includes cross-dataset evaluation, calibration analysis, additional regularization, and clinically guided augmentation policies.

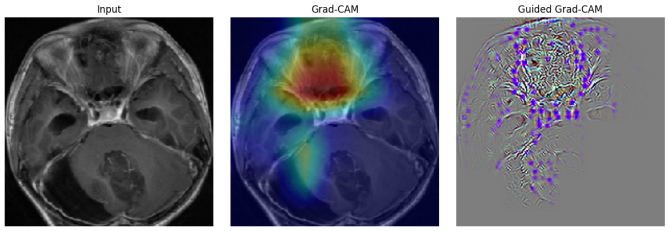


Fig. 4. Example explanation for a single MRI sample. Grad-CAM highlights class-discriminative regions at coarse spatial resolution, while Guided Grad-CAM combines Grad-CAM with guided backpropagation to produce sharper, high-frequency saliency patterns.

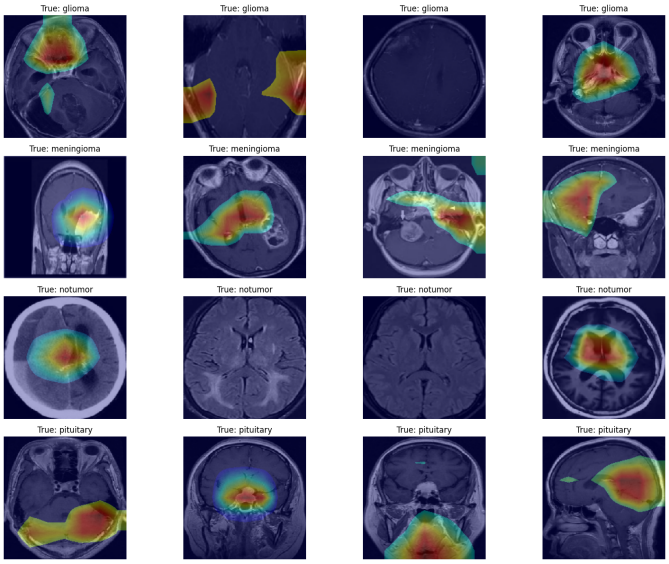


Fig. 5. Grad-CAM overlays for multiple test samples across all four classes. The heatmaps indicate which regions most influenced the predicted class score. Some activations can overlap with non-tumor structures or borders, which motivates careful preprocessing and conservative augmentation.

IX. CONCLUSION

This project implemented a complete brain MRI multiclass classification workflow with robust preprocess-

ing, ResNet-50 transfer learning, and interpretability via Grad-CAM. Both pipelines achieved approximately 98.8% test accuracy. Cropping reduced non-informative regions, and conservative train-only augmentation provided a controlled generalization test. Grad-CAM and Guided Grad-CAM visualizations offered qualitative evidence that the model often focuses on relevant regions for tumor predictions while relying on normal-structure cues for the no-tumor class [11].

REFERENCES

[1] ABIWINANDA, Nyoman ; HANIF, Muhamad u. a.: Brain tumor classification using convolutional neural network. In: *World Neurosurgery* 124 (2019), S. e297–e305

[2] BAUER, Stefan ; WIEST, Roland ; NOLTE, Lutz-Peter ; REYES, Mauricio: A survey of MRI-based medical image analysis for brain tumor studies. In: *Physics in Medicine & Biology* 58 (2013), Nr. 13, S. R97

[3] CHENG, Jun: *Brain MRI Images for Brain Tumor Detection*. 2017. – Kaggle dataset

[4] GOODFELLOW, Ian ; BENGIO, Yoshua ; COURVILLE, Aaron: *Deep Learning*. MIT Press, 2016

[5] HE, Kaiming ; ZHANG, Xiangyu ; REN, Shaoqing ; SUN, Jian: Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016

[6] HOLZINGER, Andreas ; LANGS, Georg ; DENK, Helmut: What do we need to build explainable AI systems for the medical domain? In: *arXiv preprint arXiv:1712.09923* (2017)

[7] KINGMA, Diederik P. ; BA, Jimmy: Adam: A Method for Stochastic Optimization. In: *arXiv preprint arXiv:1412.6980* (2014)

[8] LECUN, Yann ; BENGIO, Yoshua ; HINTON, Geoffrey: Deep learning. In: *Nature* 521 (2015), Nr. 7553, S. 436–444

[9] POWERS, David M. W.: Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. In: *Journal of Machine Learning Technologies* 2 (2011), Nr. 1, S. 37–63

[10] QUINONERO-CANDELA, Joaquin (Hrsg.) ; SUGIYAMA, Masashi (Hrsg.) ; SCHWAIGHOFER, Anton (Hrsg.) ; LAWRENCE, Neil D. (Hrsg.): *Dataset Shift in Machine Learning*. MIT Press, 2009

[11] SELVARAJU, Ramprasaath R. ; COGSWELL, Michael u. a.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, S. 618–626

[12] SHORTEN, Connor ; KHOSHGOFTAAR, Taghi M.: A survey on image data augmentation for deep learning. In: *Journal of Big Data* 6 (2019), Nr. 1, S. 1–48

[13] SPRINGENBERG, Jost T. ; DOSOVITSKIY, Alexey ; BROX, Thomas ; RIEDMILLER, Martin: Striving for Simplicity: The All Convolutional Net. In: *arXiv preprint arXiv:1412.6806* (2014)

[14] SRIVASTAVA, Nitish ; HINTON, Geoffrey ; KRIZHEVSKY, Alex ; SUTSKEVER, Ilya ; SALAKHUTDINOV, Ruslan: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In: *Journal of Machine Learning Research* 15 (2014), S. 1929–1958