

Deep Learning-Based Brain Tumor Classification in MRI with External Validation Across Multiple Datasets

Classificação de Tumores Cerebrais em Ressonância Magnética usando Deep Learning e Validação Externa em Múltiplos Conjuntos de Dados

Filipe Ferreira Viana¹, Leandro Henrique Furtado Pinto Silva¹, João Fernando Mari¹

Abstract: The automatic classification of brain tumors in magnetic resonance imaging (MRI) remains a challenging task due to the variability of tumor morphology and the heterogeneity of available datasets. This work presents a comparative evaluation of three deep learning models (ResNet-50, DenseNet-121, and ViT-B16) using the recently curated BRISC 2025 dataset and the BCD-MRI dataset. A systematic evaluation protocol was adopted, including internal validation, external evaluation restricted to test splits, and external evaluation over the entire external datasets. Four levels of data augmentation, ranging from none to strong, were applied to assess their impact on generalization. Internal evaluations achieved high and stable results across all models, while external test set evaluations showed only minor drops. More pronounced differences emerged when models were tested on entire external datasets: knowledge transferred from BCD-MRI generalized better to BRISC than the reverse. Data augmentation reduced part of the performance loss, particularly in external evaluations, yielding significant gains in robustness. Overall, DenseNet-121 proved to be the most reliable architecture, ResNet-50 showed resilience when combined with augmentation, and ViT-B16 achieved strong results in restricted cases but was less stable in whole-dataset external evaluations. The main contribution of this study is to demonstrate that progressive data augmentation enhances the external dataset generalization, highlighting the importance of evaluating models beyond internal splits or limited external test sets to better approximate real-world deployment conditions.

keywords: brain tumor classification — magnetic resonance imaging (MRI) — deep learning — data augmentation — external validation

Resumo: A classificação automática de tumores cerebrais em imagens de ressonância magnética (MRI) continua sendo uma tarefa desafiadora devido à variabilidade da morfologia tumoral e à heterogeneidade dos conjuntos de dados disponíveis. Este trabalho apresenta uma avaliação comparativa de três modelos de deep learning (ResNet-50, DenseNet-121 e ViT-B16) utilizando o conjunto de dados recentemente curado BRISC 2025 e o conjunto BCD-MRI. Foi adotado um protocolo sistemático de avaliação, incluindo validação interna, avaliação externa restrita aos conjuntos de teste e avaliação externa considerando os conjuntos completos. Quatro níveis de aumento de dados, variando de nenhum a forte, foram aplicados para avaliar seu impacto na capacidade de generalização. As avaliações internas apresentaram resultados elevados e estáveis em todos os modelos, enquanto as avaliações externas restritas ao conjunto de teste mostraram apenas pequenas quedas de desempenho. Diferenças mais pronunciadas surgiram quando os modelos foram testados em conjuntos externos completos: o conhecimento transferido do BCD-MRI generalizou melhor para o BRISC do que o inverso. O aumento de dados reduziu parte da perda de desempenho, especialmente nas avaliações externas, proporcionando ganhos significativos em robustez. De forma geral, o DenseNet-121 mostrou-se a arquitetura mais confiável, o ResNet-50 apresentou resiliência quando combinado com aumento de dados, e o ViT-B16 alcançou bons resultados em casos restritos, mas foi menos estável nas avaliações externas de conjuntos completos. A principal contribuição deste estudo é demonstrar que o aumento progressivo de dados melhora a generalização em conjuntos externos, ressaltando a importância de avaliar modelos para além das divisões internas ou de conjuntos de teste externos limitados, de modo a aproximar melhor as condições reais de aplicação.

Palavras-Chave: classificação de tumores cerebrais — ressonância magnética (MRI) — deep learning — aumento de dados — validação externa

¹ Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Viçosa (UFV), Brazil

*Corresponding author: joaofmari@gmail.com

DOI: <http://dx.doi.org/10.22456/2175-2745.150908> • Received: 14/10/2025 • Accepted: 27/10/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

The automatic classification of brain tumors using magnetic resonance imaging (MRI) presents a challenge at the intersection of medicine and technology, as brain tumors represent one of the most complex neurological conditions, with increasing incidence and symptoms that often overlap with other neurological pathologies [1]. Accurate identification of these lesions is complex due to the variety of morphological characteristics and the need for analysis in multiple anatomical planes [2].

The use of contrast-enhanced T1-weighted MRI images has become a standard procedure to aid in the diagnosis of brain tumors. However, this method requires subjective visual analyses that are enhanced by the use of expert teams with computer-assisted diagnostic systems, especially those that use robust machine learning and pattern recognition techniques in medical images [3, 4].

Advances in deep learning techniques have offered promising solutions for the automatic analysis of brain tumors. However, the effectiveness of these techniques is significantly affected by the quality of image preprocessing, which, when performed inadequately, can lead to the loss of critical information, compromising diagnostic accuracy [5]. Techniques such as contrast and brightness adjustment, resizing, and filtering are essential for preparing images before training deep learning models. However, the decision to apply or omit specific preprocessing steps is not trivial and depends on the application context and the characteristics of the available images [6].

Another challenge in applying deep learning is that, although there are established and robust techniques for problems related to computer vision, it is necessary to evaluate which of these techniques is most suitable, especially considering aspects such as computational performance and suitability for the problem at hand. In this sense, different factors need to be considered, in addition to the architecture itself, such as loss function combinations, hyperparameter optimizations, and data augmentation strategies [7, 8].

In this work, we investigate the impact of various factors on the classification accuracy of brain tumors in MRI images using deep learning-based models. Furthermore, we investigate how the inclusion of advanced deep learning techniques, leveraging specific features of pathology-related datasets, can significantly improve the classification accuracy of medical images.

The remainder of this paper is structured as follows. Section 2 reviews the related literature. Section 3 describes the proposed approach for model training and external evaluation in brain tumor image classification. Section 6 presents and discusses the results. Finally, Section 7 concludes the paper and outlines directions for future research.

2. Related works

Zhang et al. demonstrated the applicability of models such as Faster R-CNN, YOLOv3, and SSD for the automatic detection of brain lesions, achieving an accuracy of up to 89.77% in classification tasks [9]. This work also emphasized the importance of analyzing lesion shape and location as crucial factors for differentiated diagnoses.

Rahman and Islam [10] proposed a parallel convolutional network (PDCNN) architecture for brain tumor detection and classification, achieving an accuracy of 99.29% with reduced inference time. The authors emphasize the importance of data standardization and the use of parallel networks to capture multiple spatial representations simultaneously, which improves model robustness. In this regard, Anantharajan et al. conducted a comprehensive comparison between deep learning and machine learning methods for brain tumor detection in MRI. The research highlights that, although CNN-based models outperform traditional algorithms such as Random Forest and SVM in accuracy, hybrid approaches can still be helpful when considering interpretability and computational cost [11].

Ali et al. [12] presented an approach to improve brain tumor classification using magnetic resonance imaging. The authors evaluated various pre-trained models based on CNN and Transformer architectures on the Brain Tumor MRI Dataset (BT-MRI) and the BCD-MRI dataset. EfficientNet B0 V2 achieved the best results across all metrics, including 100% accuracy. Using this most promising model, the authors evaluated it on data containing noise, blur, and synthetic artifacts, reporting significant performance declines in this scenario. Finally, the study also explores optimization algorithms and data augmentation techniques to improve model generalization and robustness. The Nadam optimizer and the Gaussian Blur and Gaussian Noise (GB-GN) data augmentation method were the most effective in overcoming the inherent challenges of synthetic data.

Recently, Fateh et al. [13] presented the Brain Tumor Segmentation and Classification Dataset (BRISC) along with the Swin-HAFUNet architecture, specifically designed for brain tumor segmentation and classification. BRISC combines 6,000 contrast-enhanced T1 images with specialized annotations across multiple planes (axial, coronal, and sagittal) and detailed segmentation masks. Swin-HAFUNet incorporates hierarchical attention mechanisms inspired by transformers, achieving a weighted average Intersection over Union (IoU) of 82.3%.

Khan et al [14] present a ShallowMRI approach comprising a lightweight architecture with four blocks, each composed of convolution, pooling, and activation operations. In addition to these operations, the approach integrates an attention block to better extract relevant features and pre-processing with contour extraction and local binary patterns. The ShallowMRI with attention operation achieves 98.24%, 98.16%, and 99.70% on the Kaggle Multiclass, BR35H Binary, and Histogram Binary datasets, respectively. These results out-

perform pre-trained models, such as MobileNet, DenseNet, and VGG, on the same image sets. The authors carried out ablation studies that confirm the benefits of the attention block and pre-processing techniques. The ShallowMRI has a low number of FLOPs and trainable parameters, making it suitable for implementation in clinics and real-time applications.

Safwan et al. [15] explored the use of self-supervised learning techniques to improve the performance of three architectures: SimCLR, MoCo, and BYOL, when combined with the ResNet-50 and EfficientNet backbones, for brain image classification. Furthermore, the authors presented T3SSLNet, an architecture composed of four components: (i) the image spectrum enhancement block for data augmentation; (ii) the Frozen Feature Extractor block for hierarchical feature extraction; (iii) the Neural Representation Projection Learning block for learning contrastive-positive pairs; and (iv) the Thawed Classification block for tumor classification. The results indicate that, without fine-tuning, MoCo achieved the best accuracy (95.76%), followed by SimCLR (92.25%) and BYOL (81.80%), all with ResNet as the backbone. With fine-tuning, BYOL showed a significant improvement (approximately 15%), reaching 96.42%, while MoCo and SimCLR reached 96.87% and 97.02%, respectively.

3. Material and methods

3.1 Datasets

For our experiments, we used the recently introduced BRISC 2025 dataset¹ (Brain tumor Image Segmentation and Classification), which comprises 6,000 MRI images categorized into four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. BRISC 2025 is derived from the widely used “Brain Tumor MRI Dataset”², itself compiled from three sources: “Brain Tumor Classification (MRI)”³, the “Brain Tumor Dataset”⁴ (Figshare), and the “Br35H Dataset”⁵.

Compared to its predecessor, BRISC 2025 offers several advantages thanks to an enhanced curation process [13]. Specifically, only T1-weighted images were retained, mislabeled samples were reviewed and removed by expert radiologists, and all images underwent standardized preprocessing. In addition, BRISC provides precise tumor delineations, paving the way for future research on tumor detection, segmentation, and explainable AI methods.

For external evaluation, we utilized the dataset “Brain Cancer Detection - MRI Images Dataset” (BCD-MRI)⁶ [16],

which comprises MRI scans categorized into the same four classes as BRISC 2025.

Table 1 summarizes the class distribution across both datasets.

Class	BRISC 2025	BCD-MRI	Total
Glioma	1,401	1,321	2,722
Meningioma	1,635	1,333	2,968
Pituitary	1,757	1,445	3,202
No Tumor	1,207	1,514	2,721
Total Images	6,000	5,613	11,613

Table 1. Comparison of class distribution between the BRISC 2025 and BCD-MRI datasets.

Figure 1 presents randomly selected MRI samples from each class of the BRISC 2025 and BCD-MRI datasets.

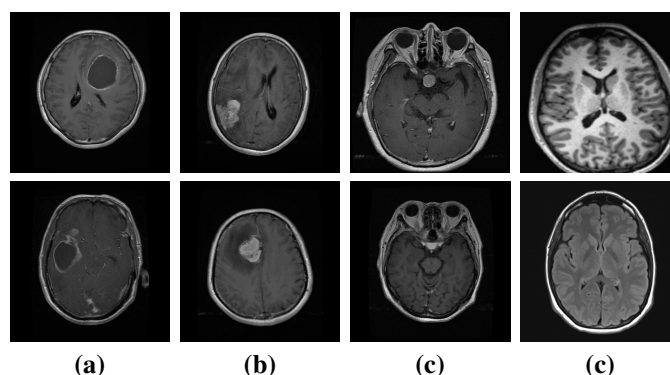


Figure 1. Representative MRI images from each class in the BRISC 2025 (top row) and BCD-MRI (bottom row) datasets. (a) Glioma, (b) Meningioma, (c) Pituitary tumor, and (d) No Tumor.

3.2 Deep Learning Architectures

We selected three deep learning-based architectures to conduct the experiments: ResNet-50, DenseNet-121, and Vision Transformer (ViT-B16).

ResNet-50, proposed by He et al. [17], stands out for using residual connections that allow the flow to incorporate input features through shortcut connections combined with a main path so that the input residual is always considered and avoids vanishing gradients. These operations enhance learning capacity, particularly for complex patterns that necessitate extended training times.

DenseNet-121 [18] is an approach composed primarily of dense blocks and transition layers. Unlike a traditional CNN, in DenseNet, each convolutional layer concatenates new features with previous ones, which cumulatively increases the size of the feature map. Therefore, to control network size, this approach uses Transition Layers to reduce dimensionality and the number of channels, typically through Batch Normalization and Pooling operations. In this sense, DenseNet enables feature reuse, a more stable gradient for deep training, and a smaller number of parameters compared to ResNet,

¹<<https://www.kaggle.com/datasets/briscdataset/brisc2025>>

²<<https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>>

³<<https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri>>

⁴<https://figshare.com/articles/dataset/brain_tumor_dataset/1512427>

⁵<<https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection>>

⁶<<https://universe.roboflow.com/kaplan/brain-cancer-detection-mri-images>>

which consequently also guarantees a lower number of FLOPs without performance loss.

Finally, the Vision Transformer (ViT-B/16) [19] uses global attention mechanisms instead of convolutions [20]. This model divides the image into small patches, transforms them into embeddings, and processes them by transformer layers, efficiently capturing long-range spatial relationships.

3.2.1 Data Augmentation

To assess the impact of data augmentation on model performance, we designed four progressive augmentation pipelines, all implemented with the torchvision library. In every case, images were resized to 224×224 pixels, converted to tensors, and normalized using the ImageNet mean and standard deviation values.

None (Baseline): Images were only resized, converted, and normalized, without any augmentation. This configuration served both as the baseline for comparison and as the transformation used during validation and test.

Mild: In addition to the baseline transformations, random horizontal flipping (with probability 0.5) and random rotation up to 15° were applied.

Moderate: Building on the mild pipeline, color perturbations were added using ColorJitter, varying brightness, contrast, saturation, and hue by up to 0.1.

Strong: The strongest augmentation pipeline combined all the moderate transformations with an additional Gaussian blur (kernel size = 3).

These four augmentation levels, ranging from none to strong, enabled a systematic investigation of the trade-off between robustness to visual variability and the potential loss of discriminative detail.

3.2.2 External Evaluation

The BRISC dataset is originally divided into training and test sets. From its training split, we reserved 20% of the samples for validation. The BCD-MRI dataset, in turn, is already provided with predefined training, validation, and test splits, which we preserved as is. Table 1, complemented by Table 2, reports the number of images in each subset for both datasets.

Our external evaluation protocol consisted of three steps: Step 1, baseline internal evaluation; Step 2, cross-dataset test evaluation; and Step 3, cross-dataset full evaluation.

Step 1: Each dataset was used independently, with models trained on its training set, monitored on the validation split for early stopping and learning rate scheduling, and finally evaluated on the corresponding test set. These results establish the baseline for comparison.

Step 2: A model trained on dataset A (training and validation) was evaluated on the test set of dataset B. This setup does not leverage the entire external dataset but enables a fair comparison with the baseline in Step 1, since the evaluation is restricted to the same test partitions.

Step 3: The same model trained on dataset A (training and validation) was evaluated on the entire dataset B (training, validation, and test). This strategy allows us to assess the

robustness of the trained model on a larger and more diverse external dataset, reducing potential bias from relying solely on the test split.

3.3 Training Procedure

All models were fine-tuned using a batch size of 32 across all architectures (ResNet-50, DenseNet-121, and ViT-B/16), starting from ImageNet-pretrained weights available via torchvision library. Optimization was performed using the AdamW optimizer with an initial learning rate of 1×10^{-4} , and the cross-entropy loss was used as the objective function.

To enhance convergence and generalization, we employed a learning rate reduction on plateaus scheduler, which reduces the learning rate when the validation loss ceases to improve (10 epochs).

Early stopping was also applied with a patience of 21 epochs to prevent overfitting. The model checkpoint corresponding to the lowest validation loss was retained for final evaluation.

4. Model Evaluation

All models were evaluated on their respective test sets using two standard metrics: overall accuracy and the macro-averaged F1-score. The metrics were computed from the confusion matrix terms: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Overall accuracy measures the proportion of correctly classified samples and is defined as:

$$\text{Accuracy} = \frac{\sum_{c=1}^C \text{TP}_c}{\sum_{c=1}^C (\text{TP}_c + \text{FP}_c + \text{FN}_c)}, \quad (1)$$

where C is the total number of classes.

The F1-score evaluates the trade-off between false positives and false negatives. For each class c , it is expressed as:

$$\text{F1-score}_c = \frac{2 \cdot \text{TP}_c}{2 \cdot \text{TP}_c + \text{FP}_c + \text{FN}_c}, \quad (2)$$

and the final macro-averaged F1-score is computed as the macro-average across all classes:

$$\text{F1-score}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{F1-score}_c. \quad (3)$$

5. Computational Environment

The experimental setup relied on a workstation built with an Intel Core i5 processor (3.0 GHz), 32 GB memory, and an NVIDIA GeForce GTX 1080 Ti GPU (11 GB). Development was carried out in Python 3.12, employing PyTorch 2.7.1 with CUDA 12.6. Model implementation and augmentation steps were handled with torchvision 0.22.1. For preprocessing and evaluation, we used scikit-learn (1.7), scikit-image (0.25.2), Pillow (11.0), and Matplotlib (3.10.3).

Datasets	BRISC 2025				BCD-MRI			
Class	Train	Val.	Test	Total	Train	Val.	Test	Total
Glioma	917	230	254	1,401	956	251	114	1,321
Meningioma	1,063	266	306	1,401	948	252	133	1,333
Pituitary	1,165	292	300	1,401	995	317	133	1,445
No Tumor	853	214	140	1,401	1,035	300	179	1,514
Total Images	3,998	1,002	1,000	6,000	3,934	1,120	559	5,613

Table 2. Distribution of images across training, validation, and test splits for the BRISC 2025 and BCD-MRI datasets.

6. Results and discussion

Table 3 summarizes the results of all experiments conducted with the three deep learning models, ResNet-50, DenseNet-121, and ViT-b16, considering four progressively stronger data augmentation strategies: none, mild, moderate, and strong.

For each model, the first two rows correspond to the baseline experiments, in which the model is trained on the training set (with its respective validation set) and evaluated on the test set of the same dataset. The remaining rows present the results of external evaluations, where the model is trained on the training set of one dataset (using the corresponding validation set) and tested on a different dataset.

Rows labeled with “(test set)” indicate that the evaluation was performed exclusively on the test set of the external dataset, enabling a fair comparison with the baseline results. Rows labeled with “(whole ds)” correspond to evaluations performed on the entire external dataset, providing more robust estimates due to the larger amount of test data available.

Figures 2 and 3 illustrate the accuracy and F1-score, respectively, obtained with the three models under different data augmentation strategies. Each column corresponds to a specific evaluation setup: (a) internal evaluation, where training and testing are performed on the same dataset; (b) external evaluation restricted to the test set of the target dataset; and (c) external evaluation using the entire target dataset. Bars represent the performance of models, allowing a visual comparison of model behavior across augmentation strategies and evaluation protocols.

When analyzing the overall scenario through Figures 2 and 3, it becomes clear that both BRISC and BCD-MRI yield strong results under internal evaluation and in external evaluation restricted to the test set. As expected, internal evaluations are more stable and consistently produce the highest values across almost all models. However, the scenario changes when the evaluation is extended to the entire external dataset (whole ds). In this setting, models trained on BCD-MRI experience only a moderate reduction when tested on BRISC, or even a small improvement in some cases. For example, DenseNet-121 achieves 0.9770 accuracy on the BRISC test set and 0.9833 on the entire BRISC (without augmentation), staying close to its internal baseline of 0.9820. In contrast, models trained on BRISC show a substantial decline when tested on the full BCD-MRI dataset. For instance, ResNet-50 drops from 0.9660 accuracy on the BCD-MRI test set to only

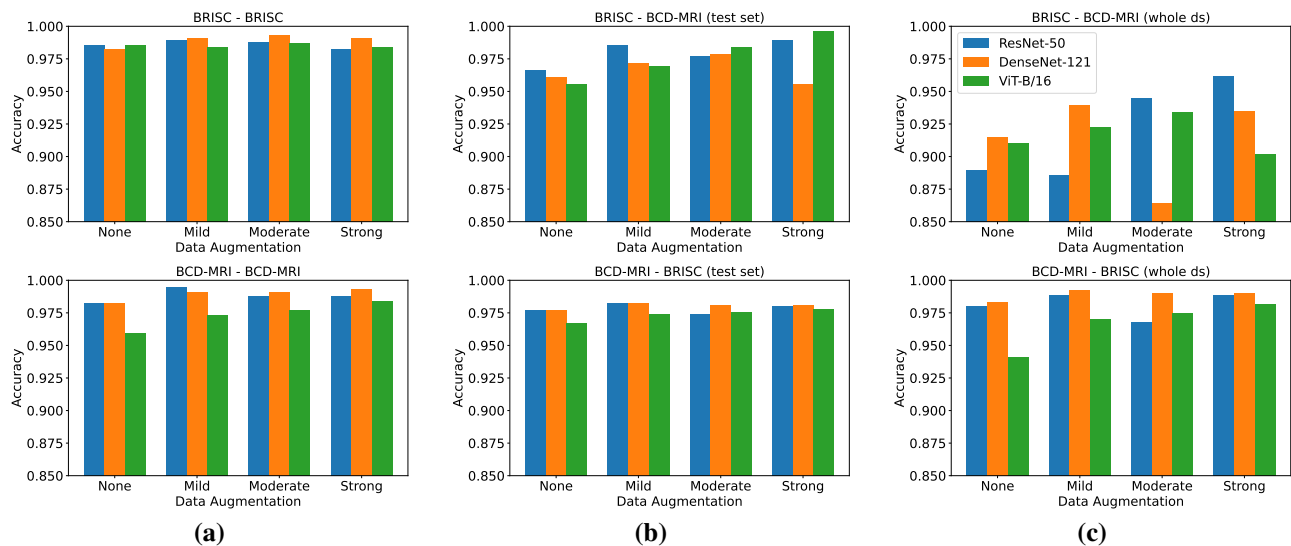
0.8895 on the whole BCD-MRI (without augmentation), partially recovering to 0.9613 with strong augmentation. These results indicate that knowledge learned from BCD-MRI generalizes better to BRISC than the reverse. Since BCD-MRI is considerably smaller and less extensively documented than BRISC, which explicitly lists all image sources, future investigations into the structure and composition of BCD-MRI are needed to clarify this behavior.

Data augmentation mitigates part of the performance gap, particularly in external evaluation scenarios. For ResNet-50 trained on BRISC and tested on the BCD-MRI test set, accuracy increases from 0.9660 without augmentation to 0.9893 with strong augmentation, while F1 improves from 0.9681 to 0.9897. Even in the more challenging whole-dataset evaluations, augmentation helps recover part of the lost performance; for example, ResNet-50 trained on BRISC rises from 0.8895 to 0.9613 in accuracy on the whole BCD-MRI as augmentation strength increases. In contrast, in internal evaluations, the effect is marginal, since results are already close to saturation. Overall, these findings indicate that augmentation is most effective for enhancing cross-dataset generalization and robustness in external evaluations, rather than for boosting internal evaluation metrics.

Considering the best performance in each evaluation setting, internal evaluations achieve their top accuracy with DenseNet-121, with 0.9930 on BRISC with moderate augmentation and 0.9928 on BCD-MRI with strong augmentation. In external evaluation restricted to the test set, the highest results are split between DenseNet-121 and ViT-16: DenseNet-121 reaches 0.9820 when trained on BCD-MRI and tested on BRISC (mild augmentation), while ViT-16 achieves 0.9964 when trained on BRISC and tested on the BCD-MRI test set (strong augmentation), outperforming the CNNs in this particular setting. In the more challenging whole-dataset evaluation, the best scores are slightly lower but still competitive: ResNet-50 trained on BRISC attains 0.9613 on the whole BCD-MRI with strong augmentation, while DenseNet-121 trained on BCD-MRI achieves 0.9923 on the whole BRISC with mild augmentation. Nevertheless, ViT-16 does not reach the top in whole-dataset evaluations and shows the most pronounced instability, with accuracy dropping from 0.9553 on the BCD-MRI test set to 0.9099 on the whole BCD-MRI when trained on BRISC without augmentation. Overall, these results confirm DenseNet-121 as the most reliable architecture

Table 3. Accuracy and F1-score for different training/testing dataset combinations across severity data augmentation levels. Empty cells indicate experiments not performed.

Model	Train/Val. - Test	Accuracy				F1-score			
		None	Mild	Moderate	Strong	None	Mild	Moderate	Strong
ResNet-50	BRISC – BRISC	0.9850	<u>0.9890</u>	0.9880	0.9820	0.9847	<u>0.9894</u>	0.9891	0.9835
	BCD-MRI – BCD-MRI	0.9821	<u>0.9946</u>	0.9875	0.9875	0.9814	<u>0.9946</u>	0.9876	0.9878
	BRISC – BCD-MRI (test set)	0.9660	0.9857	0.9767	<u>0.9893</u>	0.9681	0.9866	0.9780	<u>0.9897</u>
	BCD-MRI – BRISC (test set)	0.9770	<u>0.9820</u>	0.9740	0.9800	0.9774	<u>0.9812</u>	0.9705	0.9786
	BRISC – BCD-MRI (whole ds)	0.8895	0.8853	0.9444	<u>0.9613</u>	0.8914	0.8896	0.9446	<u>0.9611</u>
	BCD-MRI – BRISC (whole ds)	0.9803	<u>0.9887</u>	0.9680	0.9883	0.9804	<u>0.9886</u>	0.9666	0.9882
DenseNet-121	BRISC – BRISC	0.9820	0.9910	<u>0.9930</u>	0.9910	0.9834	0.9922	<u>0.9931</u>	0.9923
	BCD-MRI – BCD-MRI	0.9821	0.9911	0.9911	<u>0.9928</u>	0.9817	0.9909	0.9909	<u>0.9930</u>
	BRISC – BCD-MRI (test set)	0.9606	0.9714	<u>0.9785</u>	0.9553	0.9628	0.9733	<u>0.9805</u>	0.9592
	BCD-MRI – BRISC (test set)	0.9770	<u>0.9820</u>	0.9810	0.9810	0.9776	<u>0.9830</u>	0.9814	0.9805
	BRISC – BCD-MRI (whole ds)	0.9150	<u>0.9392</u>	0.8644	0.9344	0.9152	<u>0.9401</u>	0.8702	0.9349
	BCD-MRI – BRISC (whole ds)	0.9833	<u>0.9923</u>	0.9902	0.9900	0.9834	<u>0.9924</u>	0.9903	0.9900
ViT-16	BRISC – BRISC	0.9850	0.9840	<u>0.9870</u>	0.9840	0.9865	0.9838	<u>0.9883</u>	0.9849
	BCD-MRI – BCD-MRI	0.9589	0.9732	0.9767	<u>0.9839</u>	0.9591	0.9727	0.9756	<u>0.9834</u>
	BRISC – BCD-MRI (test set)	0.9553	0.9696	0.9839	<u>0.9964</u>	0.9565	0.9704	0.9845	<u>0.9967</u>
	BCD-MRI – BRISC (test set)	0.9260	0.9740	0.9750	<u>0.9780</u>	0.9244	0.9732	0.9739	<u>0.9789</u>
	BRISC – BCD-MRI (whole ds)	0.9099	0.9225	<u>0.9339</u>	0.9017	0.9098	0.9227	<u>0.9340</u>	0.9037
	BCD-MRI – BRISC (whole ds)	0.9408	0.9702	0.9745	<u>0.9813</u>	0.9399	0.9693	0.9737	<u>0.9812</u>

**Figure 2.** Accuracy results for the three models across four data augmentation strategies. (a) internal evaluation, (b) external evaluation restricted to the test set of the target dataset, and (c) external evaluation using the entire target dataset. Bars highlighting differences in performance across augmentation levels and evaluation setups.

and ResNet-50 as also very resilient in external scenarios.

7. Conclusion

This work evaluated ResNet-50, DenseNet-121, and ViT-B16 for brain tumor classification in MRI using the BRISC 2025 and BCD-MRI datasets. Internal results were consistently high, while external validation on the test set showed only

small performance drops but more unstable results. External evaluation over the whole dataset provided clearer findings: models trained on BRISC and tested on the full BCD-MRI showed significant performance drops. In contrast, models trained on BCD-MRI transferred better to BRISC, in some cases even showing slight improvements. These results indicate that using the whole external dataset is more realistic than relying only on the test split, and it is essential for verifying

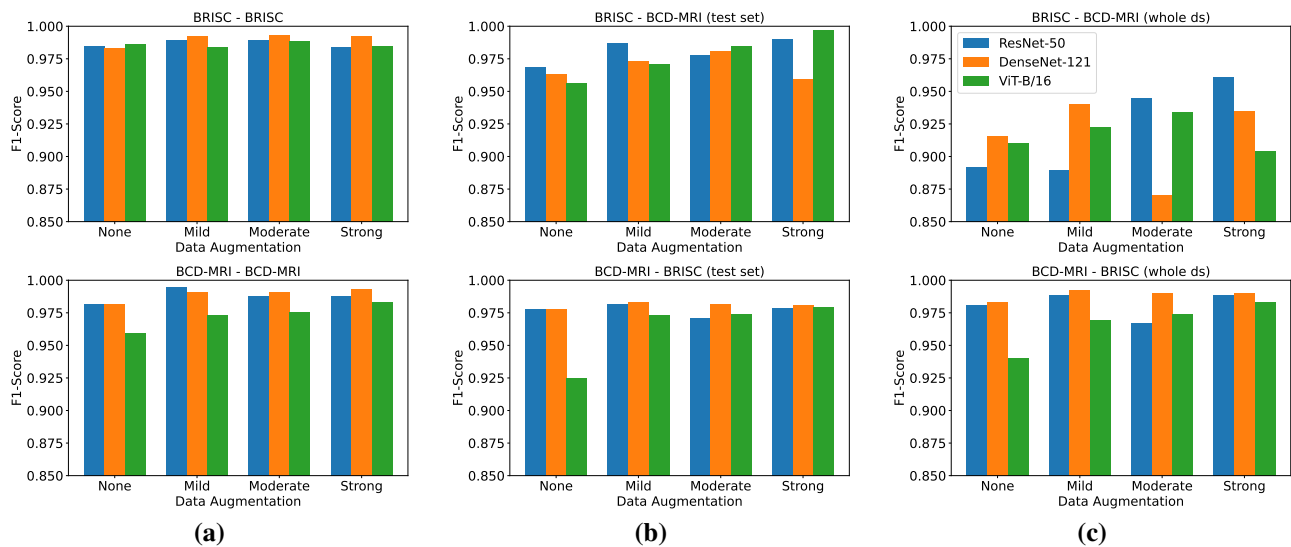


Figure 3. F1-score results for the three models across four data augmentation strategies. (a) internal evaluation, (b) external evaluation restricted to the test set of the target dataset, and (c) external evaluation using the entire target dataset. Bars highlighting differences in performance across augmentation levels and evaluation setups.

how well models may perform in real-world scenarios.

Data augmentation helped reduce these drops, particularly in external evaluations. For example, ResNet-50 trained on BRISC improved from 0.8895 to 0.9613 accuracy on the full BCD-MRI when strong augmentation was applied. The effect in internal evaluations was small, since results were already close to saturation. DenseNet-121 was the most stable model, ResNet-50 proved resilient when combined with augmentation, and ViT-B16 achieved the highest accuracy in some restricted cases but was less stable overall.

The main contribution of this work is to show that progressive data augmentation can improve generalization on external data, especially in evaluations over entire datasets, and to highlight the importance of carefully choosing augmentation strategies. As future work, we plan to study the structure of BCD-MRI to better understand its behavior, explore augmentation methods designed specifically for the medical imaging domain, and test additional datasets to confirm the robustness of these findings.

Acknowledgements

We would like to thank CAPES, CNPq, and FAPEMIG for the financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Author Contributions

Conceptualization: F.F.V, L.H.F.P.S., and J.F.M. Methodology: F.F.V, L.H.F.P.S., and J.F.M. Code Development: F.F.V, L.H.F.P.S., and J.F.M. Experiments: F.F.V, L.H.F.P.S., and J.F.M. Discussion of Results: F.F.V, L.H.F.P.S., and J.F.M.

Writing - Original Draft: F.F.V. Writing - Review & Editing: L.H.F.P.S., and J.F.M. Supervision: L.H.F.P.S., and J.F.M.

References

- [1] ESTEVA, A. et al. A guide to deep learning in healthcare. *Nature Medicine*, Nature Publishing Group, v. 25, n. 1, p. 24–29, 2019.
- [2] LITJENS, G. et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, Elsevier, v. 42, p. 60–88, 2017.
- [3] GORDILLO, N.; MONTSENY, E.; SOBREVILLA, P. State of the art survey on mri brain tumor segmentation. *Magnetic resonance imaging*, Elsevier, v. 31, n. 8, p. 1426–1438, 2013.
- [4] RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical image computing and computer-assisted intervention*. [S.l.], 2015. p. 234–241.
- [5] SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, Springer, v. 6, n. 1, p. 1–48, 2019.
- [6] DONG, S.; WANG, P.; ABBAS, K. A survey on deep learning and its applications. *Computer Science Review*, Elsevier, v. 40, p. 100379, 2021.
- [7] SEYREK, E. C.; UYSAL, M. A comparative analysis of various activation functions and optimizers in a convolutional neural network for hyperspectral image classification. *Multimedia Tools and Applications*, Springer, v. 83, n. 18, p. 53785–53816, 2024.

- [8] SILVA, L. H. F. P. et al. Evaluating combinations of optimizers and loss functions for cloud removal using diffusion models. In: *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*. [s.n.], 2025. p. 648–656. ISBN 978-989-758-728-3. ISSN 2184-4321. Disponível em: <<https://doi.org/10.5220/0013252100003912>>.
- [9] ZHANG, S. et al. Stroke lesion detection and analysis in mri images based on deep learning. *Journal of Healthcare Engineering*, Hindawi, v. 2021, p. 5524769, 2021.
- [10] RAHMAN, T.; ISLAM, M. S. Mri brain tumor detection and classification using parallel deep convolutional neural networks. *Measurement: Sensors*, Elsevier, v. 26, p. 100694, 2023.
- [11] ANANTHARAJAN, S. et al. Mri brain tumor detection using deep learning and machine learning approaches. *Measurement: Sensors*, Elsevier, v. 31, p. 101026, 2024.
- [12] ALI, M. A. et al. Enhancing mri brain tumor classification: A comprehensive approach integrating real-life scenario simulation and augmentation techniques. *Physica Medica*, Elsevier, v. 127, p. 104841, 2024.
- [13] FATEH, A. et al. *BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification with Swin-HAFNet*. 2025. Disponível em: <<https://arxiv.org/abs/2506.14318>>.
- [14] KHAN, S. U. R. et al. Shallowmri: a novel lightweight cnn with novel attention mechanism for multi brain tumor classification in mri images. *Biomedical Signal Processing and Control*, Elsevier, v. 111, p. 108425, 2026.
- [15] SAFWAN, M. N. et al. T3sslnet: Triple-method self-supervised learning for enhanced brain tumor classification in mri. *IEEE Access*, IEEE, 2025.
- [16] KAPLAN. Open Source Dataset, *Brain Cancer Detection - MRI Images Dataset*. [S.l.]: Roboflow, 2023. <<https://universe.roboflow.com/kaplan/brain-cancer-detection-mri-images>>. Visited on 2025-09-01.
- [17] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778.
- [18] HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 4700–4708.
- [19] DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2021.
- [20] VASWANI, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems*, v. 30, 2017.