A Comparative Evaluation of CNN and Vision Transformer Architectures for Brain Tumor Detection in MRI Scans

Zain Aboobacker

Natick High School

July 14, 2025

Abstract

This study evaluates the performance of convolutional neural networks (CNNs) and Vision Transformers (ViTs) in classifying various brain MRI scans for the detection of tumors. Model series such as EfficientNet, ConvNeXt, ViT, and SwinTransformer were trained on a publicly available multiclass brain tumor dataset. To support experimentation and reproducibility, a custom GUI-based deep learning software was developed, enabling users to train models, configure parameters, apply data augmentation, monitor performance metrics, and generate diagnostic reports. A comparative analysis was conducted using accuracy, precision, recall, F1-score, AUROC, and confusion matrices. Results showed that ViT-B/16, EfficientNetB0, and ViT-B/32 achieved test accuracies of over 98 %. ViT-B/16 performed the best overall, demonstrating that larger model capacity and reduced patch size enhance feature extraction in brain MRI classification. EfficientNetB0 delivered strong performance despite its reduced complexity, demonstrating the strength of CNNs and the potential of transformer-based architectures for detecting brain tumors.

1. Introduction

Brain tumor detection with MRI scans is highly crucial for early treatment and diagnosis in medicine. However, manual evaluations of MRI scans are time-consuming and may be prone to error. This research explores the possibilities of using convolutional neural networks (CNNs), vision transformers (ViTs), and hybrid CNN-Transformer models in the automation of brain tumor detection images with an additional goal of creating a user-friendly deep learning platform for model training and diagnostics.

Despite the promising potential of the automation of tumor detection, many are skeptical due to risks associated with false negatives and positives, and the severe consequences that may come from it. This research aims to analyze the plausibility of these risks by evaluating multiple deep learning models on brain MRI data using accuracy, precision, recall, F1-score, AUROC, and confusion matrices. This study assesses the reliability and diagnostic consistency of AI-based systems in medical applications with a particular focus on brain tumor classification.

In recent years, CNNs have demonstrated strong performance in detecting abnormalities in a plethora of image scans such as X-rays, CT scans, and MRIs. However, most systems require

significant technical expertise to operate. These limitations prevent accessibility to researchers and clinicians without the required background. Vision Transformers (ViTs) are a relatively newer class of models originally designed for natural classification; however, they have shown competitive results in medical images, but remain underexplored in clinical windows. This research contributes to the medical field by directly comparing several CNN and ViT architectures and integrating them into an intuitive interactive platform. By providing performance benchmarks and a practical interface, this study aims to advance the deployment of AI tools in radiological diagnostics.

Additionally, a user-friendly deep learning application was developed, allowing users to load datasets, configure training parameters, monitor real-time performance, and export evaluation results and diagnostic reports. The application allows users to select multiple model architectures and requires no coding knowledge. The interface enables streamlined experimentation and diagnostics for users without extensive background knowledge.

2. Related Works

Deep learning has received significant traction in medical imaging due to its ability to recognize complex patterns in data, allowing accurate diagnosis and detection of subtle anomalies in medical scans. Convolutional neural networks (CNNs), in particular, have been widely applied to MRI classification tasks in recent studies.

Albalawi et al. (2024) developed a custom 25-layer CNN for multiclass brain tumor classification on a dataset of over 7,000 MRI images and achieved an accuracy of 99 %. This significantly outperformed earlier methods, demonstrating the strong feature extraction and classification of CNNs for medical images.

Within the same domain, Zarenia et al. (2025) suggested a hybrid CNN model with a hierarchical multiscale deformable attention module (MS-DAM) to classify brain tumors across 14 classes. Their model, which achieved an accuracy of more than 96.5%, had a segmentation module built in, an end-to-end solution for detection and localization in brain MRI.

Additionally, Mahmud et al. (2023) made a comparison between lightweight CNNs and typical deep architectures like ResNet and VGG, with the conclusion being that simpler CNNs can also produce high accuracy if well optimized for brain MRI classification.

Although CNNs have exhibited strong performance, vision transformers (ViTs) remain relatively unexplored in medical imaging. Recent studies, however, highlight their potential. Reddy et al. (2024) explored the application of fine-tuned ViT (FTVT) models for four-class brain tumor classification using a similar dataset to that of Albalawi et al. (2024). Their best model, FTVT-l16, achieved a classification accuracy of 98.70%, also demonstrating the effectiveness of transformer-based models in medical diagnosis.

Similarly, Poornam et al. (2024) introduced VITALT, a rotation-invariant Vision Transformer architecture for brain tumor classification. VITALT was trained across four MRI datasets and achieved accuracy above 98.8 %, demonstrating the strength of ViTs when rotational variance is taken into account.

Hybrid CNN-VITs have also demonstrated promising performance in the field of medical imaging. Karagoz et al. (2024) introduced ResViT, a hybrid CNN–ViT model pre-trained using self-supervised contrastive learning. On the BraTS and Kaggle MRI datasets, ResViT achieved over 98.4 % accuracy.

These works demonstrate the effectiveness of CNNs, ViTs, and CNN-ViT hybrid models for brain MRI classification. However, most existing research requires advanced expertise and lacks integrated tools for non-expert users. This project addresses that gap by performing a comparative evaluation on a range of CNN and ViT models while also introducing a user-friendly interface for model training, evaluation, and diagnostic reporting.

3. Dataset and Preprocessing

The dataset used in this study consists of brain MRI scans for brain tumor classification. It includes four classes: glioma, meningioma, pituitary, and no tumor. The images were sourced from a publicly available Kaggle dataset, containing a total of 7,022 images evenly distributed across the classes. The dataset was split into training and testing sets using an 80/20 ratio.

Image Standardization

To ensure consistency in input dimensions across all models, each MRI image was resized to a fixed resolution of 224×224 pixels. This resizing operation can be formalized as a mapping function:

$$R : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{224 \times 224 \times C}$$

where H, W, and C represent the original height, width, and number of channels of the input image, respectively. This transformation enforces spatial uniformity across the dataset and aligns with input size requirements of pretrained CNNs and transformer-based models.

Following resizing, all images were normalized to zero mean and unit variance using the following channel-wise transformation:

$$x' = \frac{x - \mu}{\sigma}$$

where x denotes the original pixel intensity, μ is the mean, and σ is the standard deviation for each channel. In this study, $\mu = \sigma = 0.5$ were used for all RGB channels, scaling pixel values from the $[0, 1]$ range to $[-1, 1]$:

$$x' = \frac{x - 0.5}{0.5}$$

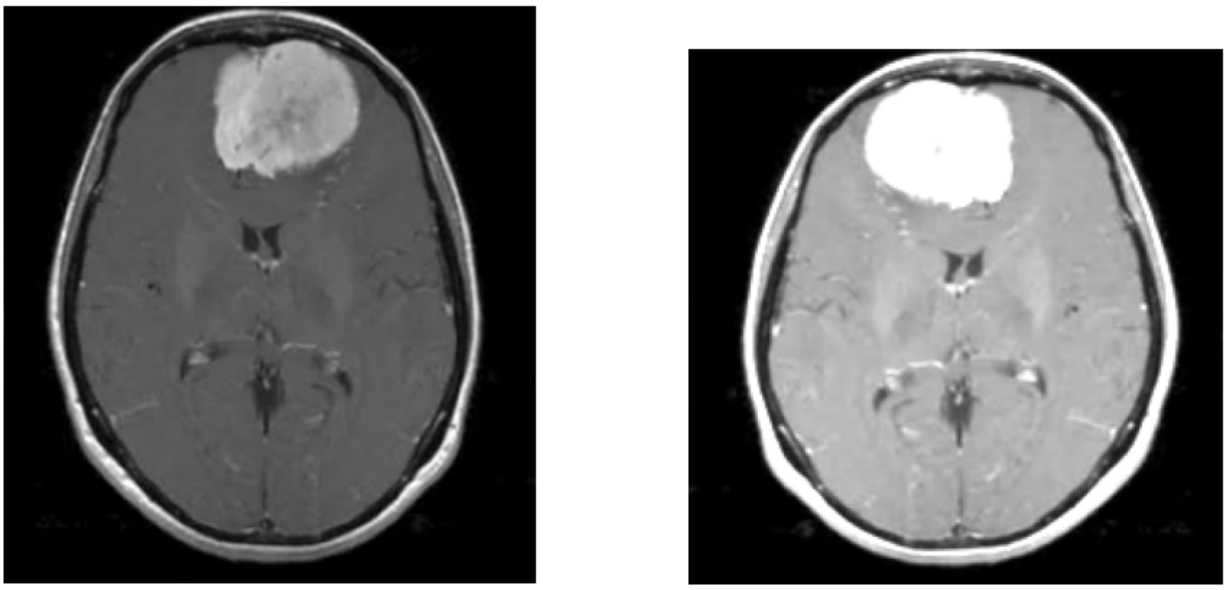
This normalization improves numerical stability and accelerates model convergence during training.

Data Augmentation

To enhance generalization and reduce overfitting, data augmentation was applied to the training images using stochastic transformations. These augmentations simulate natural variations in MRI acquisition, such as orientation, position, and scanner-induced artifacts, while preserving semantic content. The following techniques were used:

- **Random Horizontal Flip**: Applied with probability $p = 0.5$, flipping the image along the vertical axis:
$$x' = \text{Flip}_H(x) \quad \text{if } u < 0.5, \quad u \sim \mathcal{U}(0, 1)$$
- **Random Affine Transformation**: Applied with probability $p = 0.5$, A composite transformation involving small random rotations, translations, and scaling:
$$T(x) = A_{\theta,s,t}(x)$$
where $\theta \in [-15°, 15°]$ is the rotation angle, $s \in [0.95, 1.05]$ is the scaling factor, and t is a translation vector with magnitude up to 5% of the image dimensions.
- **Gaussian Blur**: A convolutional smoothing operation applied with probability $p = 0.2$:
$$x' = x * G_\sigma$$
where G_σ is a 2D Gaussian kernel with standard deviation $\sigma \in [0.1, 0.5]$ and fixed kernel size of 3.

These augmentations were applied dynamically during training, introducing slight variations at each epoch to improve the model's robustness to real-world imaging noise and spatial distortions.



(a) Sample MRI without Data Augmentation and Normalization

(b) Sample MRI with Data Augmentation and Normalization

Figure 1: Example MRI images with and without Data Augmentation / Normalization

4. Methodology

This section outlines the experimental framework used to accurately compare the performance of Convolutional Neural Networks and Vision Transformers in the task of brain tumor classification.

The methodology includes a standardized classification head, model selection, training strategy, and evaluation metrics. All models were trained using the same dataset, preprocessing pipeline, and performance metrics to ensure a fair and consistent comparison.

To isolate the effect of feature extraction capabilities, each model architecture was modified to use a uniform classifier head. The training was conducted in two stages: an initial frozen phase, where only the classifier head was trained, followed by a fine-tuning phase where the entire network was unfrozen and updated.

Classifier Head Architecture

To ensure a consistent evaluation across models, each architecture's original classification layer was replaced with a custom, uniform fully connected (FC) head. This standardized head allows meaningful comparison of backbone feature extraction performance without architectural bias.

The custom classifier head includes:

- A fully connected layer reducing to 512 units
- ReLU activation
- Dropout with probability $p = 0.5$
- A fully connected layer reducing to 256 units
- ReLU activation
- Dropout with probability $p = 0.3$
- An output layer with 4 units (corresponding to the number of tumor classes)

Model Architectures

Three model families were selected due to their architectural diversity and suitability in image classification research. EfficientNet models (Tan & Le, 2019) were introduced to compare the performance of convolutional networks at varying complexities and parameter scale sizes. Vision Transformers (ViT) (**dosovitskiy2020vit**) were used in the study to analyze the effect of patch size and self-attention mechanisms in medical images. Swin Transformers (Liu et al., 2021) and ConvNeXt (**liu2022convnext**) were selected to compare hierarchical transformer architectures against modernized convolutional networks inspired by transformer design principles.

- **CNNs**: EfficientNetB0, EfficientNetB5, EfficientNetB7
- **Vision Transformers**: ViT-B/16, ViT-B/32
- **Hybrid CNN-Transformer Models**: SwinV2Base, ConvNeXtBase

All models were initialized with pretrained weights from ImageNet and fine-tuned on the brain MRI dataset using the custom classification head.

Training Configuration

All models were trained on ideal loss functions and optimizers. However, parameters such as learning rate, batch size, warmup time varied for each model based on their unique architectures. Additionally, model backends were frozen for 5 -10 training epochs, depending on model architecture.

Loss Function

The models were trained using the categorical cross-entropy loss, defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where C is the number of classes, y_i is the ground truth label (one-hot encoded), and $\hat{y}_i$ is the predicted probability for class i after softmax. For all experiments, label smoothing was applied to reduce overconfidence in predictions.

Optimizer

All models were optimized using the AdamW optimizer, a variant of Adam that decouples weight decay from gradient updates. The update rule is:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right)$$

Where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, η is the learning rate, and λ is the weight decay coefficient.

Learning Rate Scheduling

To improve convergence, a cosine annealing learning rate schedule with linear warmup was used. The learning rate at epoch t is defined as:

$$\eta_t = \begin{cases} \eta_0 \cdot \frac{t}{T_{\text{warmup}}}, & \text{if } t < T_{\text{warmup}} \\ \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min}) \left(1 + \cos\left(\frac{\pi(t - T_{\text{warmup}})}{T - T_{\text{warmup}}}\right)\right), & \text{otherwise} \end{cases}$$

where η_0 is the initial learning rate, $\eta_{\min}$ is the minimum learning rate, T is the total number of epochs, and T_{warmup} is the number of warmup epochs.

Evaluation Metrics

Model performance was assessed using the following evaluation metrics:

- **Accuracy**:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision and Recall (per class)**:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score** (harmonic mean of precision and recall):

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

  Macro and micro averaging were applied to summarize scores across all classes.

- **AUROC (Area Under ROC Curve)**: Measures class separability using the true positive rate (TPR) and false positive rate (FPR):

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}$$

  The ROC curve plots TPR vs. FPR at various threshold levels; the area under this curve quantifies model discrimination capability.

5. Results

This section presents the outcomes of model evaluations across the multiple CNN and ViT model architectures trained. The results are organized into quantitative metrics and visual diagnostics. Model performance was assessed using accuracy, precision, recall, F1-score (macro and micro), AUROC, confusion matrices, and ROC curves, which results highlight the relative strengths and weaknesses of each architecture and provide insight into their classification behavior. Additionally, this section delves into the reliability of data augmentation in brain tumor classification in MRI scans.

Quantitative Results

To ensure accurate comparison of the models, quantitative metrics such as accuracy, precision, recall, F1-score (macro and micro), and AUROC were computed for each model in order to establish a plausible comparative analysis.

Unless otherwise stated, all evaluation metrics are macro-averaged across all classes. Additionally, all models reported are trained without data augmentation unless otherwise stated.

Table 1 below summarizes the performance of four deep learning models, ViT-B/16, SwinTransformerV2Base, ConvNeXtBase, and EfficientNetB0. Metrics reported include accuracy, precision, recall, F1 score, and AUROC (Area Under the Receiver Operating Characteristic Curve)

Table 1: Performance metrics (%) for ViT-B/16, ConvNeXt, SwinTransformerV2, and EfficientNetB0 on the brain tumor MRI dataset.

Model	Accuracy	Precision	Recall	F1 Score	AUROC
ViT-B/16	98.17	98.29	98.17	98.20	99.87
ConvNeXtBase	96.51	96.81	96.51	96.58	99.85
SwinTransformerBase	93.66	94.20	93.66	93.77	99.45
EfficientNetB0	98.01	98.10	98.01	98.04	99.94

Table 2 shows the performance of the ViT-B/16 model with and without data augmentation on the brain tumor MRI classification task. Percent change is calculated as the relative improvement from the no-augmentation baseline.

Table 2: Performance of ViT-B/16 with and without data augmentation. All values are percentages. Percent change is relative to the no-augmentation baseline.

Metric	No Augmentation	With Augmentation	Change
Accuracy	98.17 %	98.25 %	+0.08 %
Precision	98.29 %	98.37 %	+0.08 %
Recall	98.17 %	98.25 %	+0.08 %
F1 Score	98.20 %	98.28 %	+0.08 %
AUROC	99.87 %	99.91 %	+0.04 %

Table 3 shows a correlation between model complexity and performance by summarizing the performance of the EfficientNet model family. Variants such as EfficientNetB0, EfficientNetB5, and EfficientNetB7 were the models trained in the study.

Table 3: Comparison of EfficientNet variants on test set metrics. All performance values are shown as percentages.

Model	Parameters (M)	Accuracy	Precision	Recall	F1-Score	AUROC
EfficientNetB0	5.3	98.01 %	98.10 %	98.01 %	98.04 %	99.94 %
EfficientNetB5	28.9	96.69 %	96.83 %	96.69 %	96.74 %	99.78 %
EfficientNetB7	66.3	96.20 %	96.46 %	96.20 %	96.25 %	99.80 %

For class-wise and dataset-wise evaluation, both macro and micro metrics were evaluated for each model trained. Table 4 shows both macro and micro metrics of a trained ViT-B/16 model.

Table 4: Macro and Micro-averaged performance metrics (%) for the ViT-B/16 model on the brain tumor MRI dataset. Similar values suggest class balance in the dataset.

Averaging	Accuracy	Precision	Recall	F1 Score
Macro	98.17 %	98.29 %	98.17 %	98.20 %
Micro	98.32 %	98.32 %	98.32 %	98.32 %

Table 5 reports the performance of ViT-B/16 and ViT-B/32 across all evaluation metrics.

Table 5: Comparison of ViT-B/16 and ViT-B/32 on test set performance metrics (%).

Model	Accuracy	Precision	Recall	F1-Score	AUROC
ViT-B/16	98.17 %	98.29 %	98.17 %	98.20 %	99.87 %
ViT-B/32	98.00 %	98.15 %	98.00 %	98.03 %	99.90 %

Visual Diagnostics

In order to delve further into class-wise performance of models, visual diagnostics such as ROC (Receiving Operating Characteristic) graphs and confusion matrices were used. These visual graphs offered insight into overall model performance on separate classes while also offering crucial insight into the presence of false negatives and positives in the trained model.

- **Confusion Matrix**: A tabular summary of prediction results showing the distribution of true vs. predicted labels across all classes. Figure 2 displays a confusion matrix of a ViT-B/16 model trained without data augmentation.

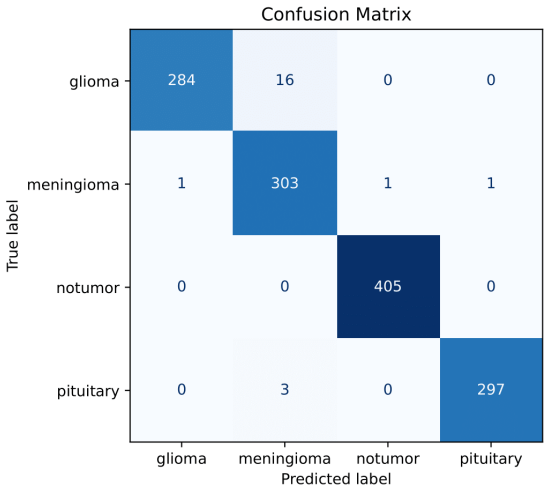


Figure 2: Example Confusion Matrix for ViT-B/16

- **ROC Curve**: A visual representation of model performance across all thresholds. Figure 3 displays a ROC Curve of a ViT-B/16 model trained without data augmentation.

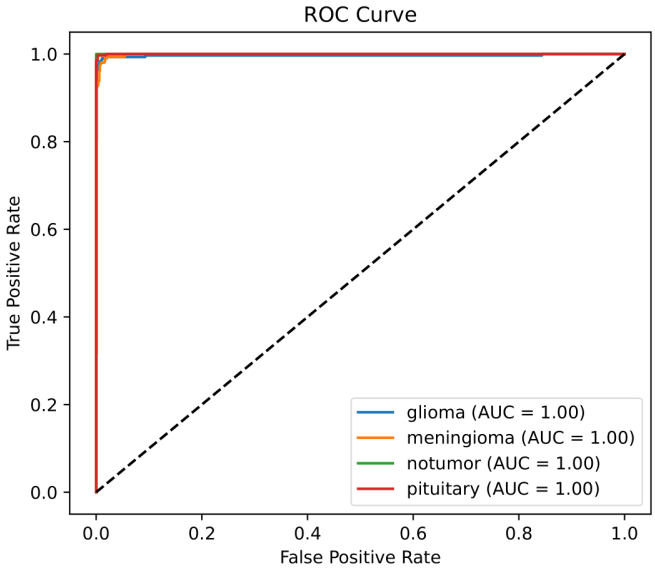


Figure 3: Example ROC Curve for ViT-B/16

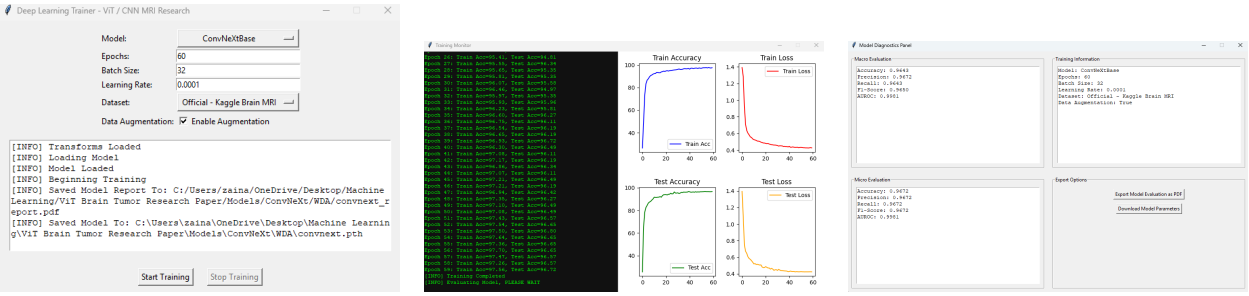
6. Software Interface

Overview

A custom GUI-based deep learning application was developed to support experimentation, reproducibility, and diagnostic evaluation of brain tumor classification models. Built using Python's built-in Tkinter framework, the software provides an intuitive interface that allows users to configure hyperparameters for model training while viewing results in real-time. This software was developed to improve user accessibility and reproducibility, especially for users without prior programming experience.

Key Features

- **Training Configuration** User-editable fields for model architecture, learning rate, batch size, epochs, and data augmentation.
- **Real-Time Console** Dynamic logging of training and testing accuracies with real-time graphical plots.
- **Post-Training Diagnostics** After training is completed, a diagnostics panel opens displaying testing results.
- **Export Options** Users can export models and export model reports, including metrics, confusion matrices, and ROC curves.



(a) Model and Dataset Selection (b) Real-Time Training Console (c) Post-Training Diagnostics Panel

Figure 4: Screenshots of the software interface. Users begin by selecting model parameters and dataset (left), monitor training live(center), and view detailed evaluation metrics after training (right).

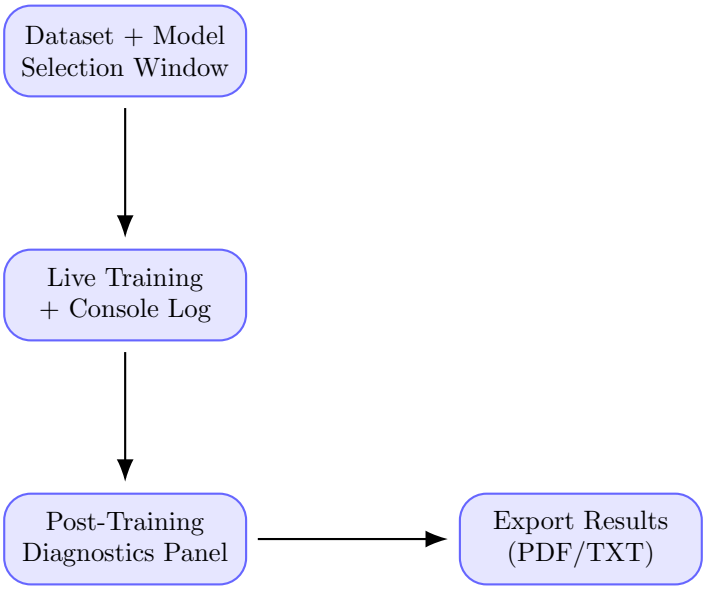


Figure 5: Software interface workflow: users progress from dataset/model selection to training, then to diagnostics and export.

By combining real-time training visualization, easy configuration, and diagnostic output in one interactive and intuitive interface. This effectively bridges the gap between raw machine learning code and practical usability. It serves as an effective research aid and an educational platform for AI-based medical image analysis.

7. Discussion

This study revealed that ViT-B/16, EfficientNetB0, and ViT-B/32 achieved the strongest classification performance with test accuracies exceeding 98 %. Among these models, ViT-B/16 performed best overall, suggesting that small patch sizes and high model capacity enable superior feature extraction in brain MRI classification tasks. Additionally, EfficientNetB0 achieved high performance

capabilities despite its relatively lower model complexity compared to other models trained. This highlights the power of efficient CNN architectures when combined with strong inductive biases.

ViT-B/32 demonstrated nearly equivalent performance to ViT-B/16, which indicates that despite larger patch sizes, effective feature extraction can still be conducted. Additionally, this suggests that ViT-B/32 may be a viable alternative for scenarios where computational capacity is limited.

EfficientNetB5, EfficientNetB7, and ConvNextBase produced solid results with accuracies varying between 96 - 97 %. However, this highlights that higher model complexity does not translate to improved performance compared to simpler model architectures. Notably, EfficientNetB7 did not yield performance improvements over EfficientNetB0 despite its higher complexity.

SwinV2Base displayed the lowest performance among the evaluated models, with an accuracy of 94 %. Despite its success in natural vision tasks, its model structure may not have aligned well with the properties of brain MRI data. Additionally, the model's reliance on large-scale pretraining and longer training schedules limited its effectiveness in constrained conditions.

Data augmentation improved performance across nearly all models. While the magnitude of improvement varied by model architecture, the consistent upward trend reinforces the value of augmentation in enhancing generalization and robustness.

ROC Curves for all high-performing models demonstrated strong class separability, with AUROC values consistently above 0.98. The ROC Curves were steep near the origin, confirming low false positive rates (FPR) at high true positive rates (TPR), which is essential in a clinical context where unnecessary alarm must be minimized without compromising sensitivity.

Confusion matrices further revealed that most classification errors occurred between glioma and meningioma classes, as they exhibit overlapping radiological features on MRI scans. Importantly, the number of false negatives was low, but not zero. Despite overall success in preventing false negatives, the presence of false negatives is crucial in medical diagnostics, as failure to detect tumors can delay treatment and impact patient outcomes.

Limitations

This study was limited to a single, relatively small Brain MRI dataset and did not evaluate cross-instituational generalization or external validation. Additionally, computational resource constraints limited the range of hyperparameter tuning, training duration, and the use of extra data augmentation pipelines. Certain models, such as SwinV2Base, may have underperformed due to insufficient training time or a lack of larger-scale datasets.

Conclusion

This study conducted a comparative evaluation of Convolutional Neural Network (CNN) and Vision Transformer (ViT) architectures for multiclass brain tumor classification through brain MRI scans. Among the evaluated models, ViT-B/16, EfficientNetB0, and ViT-B/32 achieved the strongest classification performance with test accuracies exceeding 98 %. These results suggest that both efficient CNN architectures and transformer-based models can achieve robust diagnostic accuracy in the field of medical imaging diagnostics.

Additionally, the study introduced a custom-built graphical user interface used to streamline model

selection, data augmentation, configure model hyperparameters, monitor performance in real-time, and export model reports, including confusion matrices and ROC curves, all without requiring programming expertise. This combination of strong model performance and accessible, easy-to-use software highlights the potential of deep learning to support clinical decision making and education applications in radiology.

Future Work

These future research studies can incorporate validation of trained models on more varied datasets, including external institutions and other imaging modalities such as CT and PET. Enhancements can involve ensembling of models, use of tools such as Grad-CAM to localize tumors and visualization, and optimizing hyperparameter search with automated techniques.

References

Albalawi, E., Thakur, A., Dorai, D. R., Bhatia Khan, S., Mahesh, T. R., Almusharraf, A., Aurangzeb, K., & Anwar, M. S. (2024). Enhancing brain tumor classification in mri scans with a multi-layer customized convolutional neural network approach [Published June 12, 2024]. *Frontiers in Computational Neuroscience*, *18*, 1418546. https://doi.org/10.3389/fncom.2024.1418546

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2010.11929

Karagoz, A., Dogan, F., Topcu, M., & Gungor, T. (2024). Resvit: Residual vision transformer for brain tumor classification with self-supervised learning. *arXiv preprint arXiv:2411.12874*. https://arxiv.org/abs/2411.12874

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. https://arxiv.org/abs/2103.14030

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986. https://arxiv.org/abs/2201.03545

Mahmud, M., Alam, M. A., Ahmed, I., & Rahman, A. H. M. A. (2023). A comparative study of lightweight cnns for brain tumor classification. *IEEE Access*, *11*, 27693–27703. https://doi.org/10.1109/ACCESS.2023.3256411

Poornam, B., Anushya, R., & Kavitha, G. (2024). Rotation-invariant vision transformer for brain tumor classification. *Frontiers in Neuroinformatics*, *18*, 1414925. https://doi.org/10.3389/fninf.2024.1414925

Reddy, K., Sharma, A., & Mehta, R. (2024). Fine-tuned vision transformers for brain tumor classification using mri. *Frontiers in Oncology*, *14*, 1400341. https://doi.org/10.3389/fonc.2024.1400341

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, *97*, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

Zarenia, E., Akhlaghi Far, A., & Rezaee, K. (2025). Automated multi-class mri brain tumor classification and segmentation using deformable attention and saliency mapping [Published March 8, 2025]. *Scientific Reports*, *15*(1), 8114. https://doi.org/10.1038/s41598-025-92776-1

Appendix A. Training Recipes

Table 6: Training Recipes for All Models

Model	LR	Batch Size	Epochs	Freeze Epochs	Warmup Epochs
SwinV2Base	1e-4	32	60	5	6
ConvNeXtBase	1e-4	32	60	5	6
ViT-B16	1e-4	32	50	5	6
ViT-B32	1e-4	32	50	5	6
EfficientNetB0	5e-4	32	30	5	3
EfficientNetB5	3e-4	32	50	5	5
EfficientNetB7	1e-4	32	60	5	6

Appendix B. Model Architecture Summaries

Model	Type	Params (M)	Input Size	Patch/Kernel	Pretrained
EfficientNetB0	CNN	5.3	224×224	3×3	ImageNet-1k
EfficientNetB5	CNN	30.0	456×456	3×3	ImageNet-1k
EfficientNetB7	CNN	66.0	600×600	3×3	ImageNet-1k
ConvNeXt-Base	CNN	88.6	224×224	7×7	ImageNet-21k
Swin-Base	Transformer	87.8	224×224	Shifted 4×4	ImageNet-21k
ViT-B/16	Transformer	86.6	224×224	16×16	ImageNet-21k
ViT-B/32	Transformer	88.6	224×224	32×32	ImageNet-21k

Table 7: Summary of architectures evaluated in this study.

Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix C. Training Results for All Models

Table 8: Macro-averaged performance metrics (%) for all trained models, with and without data augmentation.

Model	Aug	Accuracy	Precision	Recall	F1 Score	AUROC
ViT-B/16	Yes	98.25	98.37	98.25	98.28	99.91
ViT-B/16	No	98.17	98.29	98.17	98.20	99.87
ViT-B/32	Yes	97.92	98.11	97.92	97.97	99.94
ViT-B/32	No	98.00	98.15	98.00	98.03	99.90
EfficientNetB0	Yes	98.17	98.32	98.17	98.21	99.90
EfficientNetB0	No	98.01	98.10	98.01	98.04	99.94
EfficientNetB5	Yes	96.95	97.22	96.95	97.03	99.76
EfficientNetB5	No	96.69	96.83	96.69	96.74	99.78
EfficientNetB7	Yes	96.20	96.46	96.20	96.25	99.80
EfficientNetB7	No	96.10	96.35	96.10	96.17	99.71
ConvNeXtBase	Yes	96.43	96.72	96.43	96.50	99.81
ConvNeXtBase	No	96.51	96.81	96.51	96.58	99.85
SwinV2Base	Yes	93.73	94.23	93.73	93.85	99.46
SwinV2Base	No	93.66	94.20	93.66	93.77	99.45

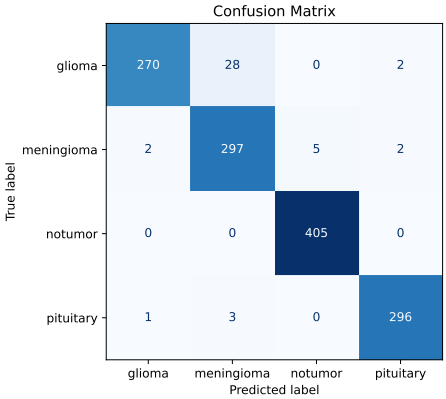
Appendix D. Appendix: Model Performance Visualizations

This appendix presents both ROC Curves and Confusion Matrices for each model trained in the study **with data augmentation** and **without data augmentation**.

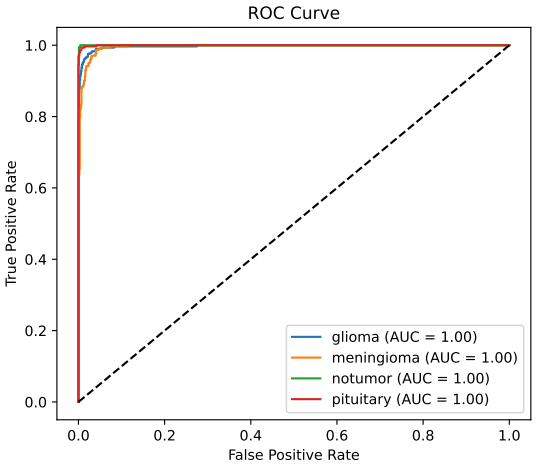
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.1. ConvNeXt-Base

Without Data Augmentation



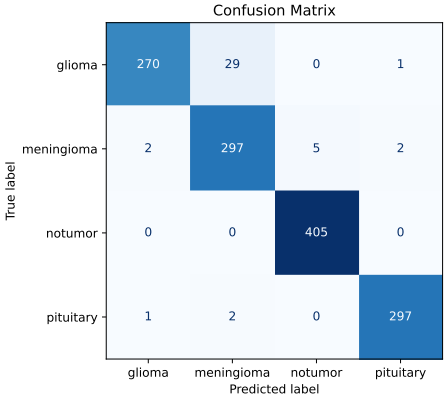
Confusion Matrix



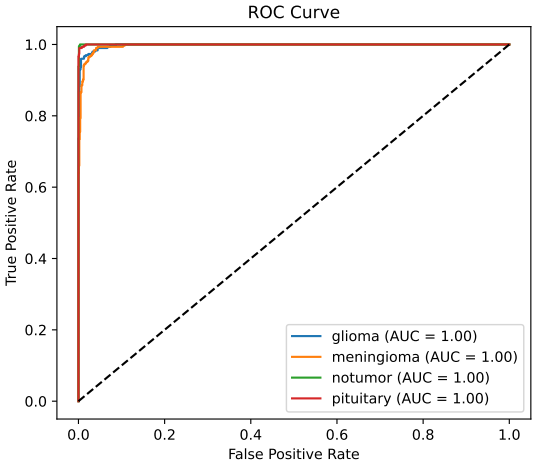
ROC Curve

Figure 6: ConvNeXt-Base – No Augmentation

With Data Augmentation



Confusion Matrix



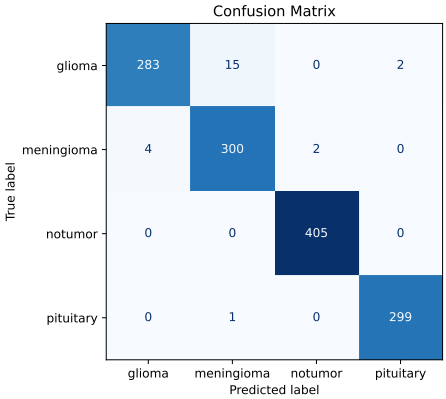
ROC Curve

Figure 7: ConvNeXt-Base – With Augmentation

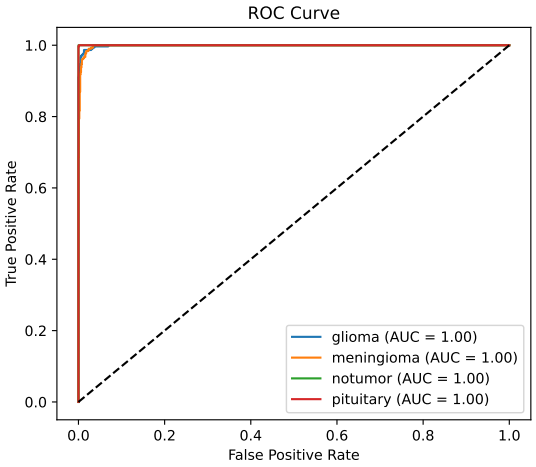
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.2. EfficientNet-B0

Without Data Augmentation



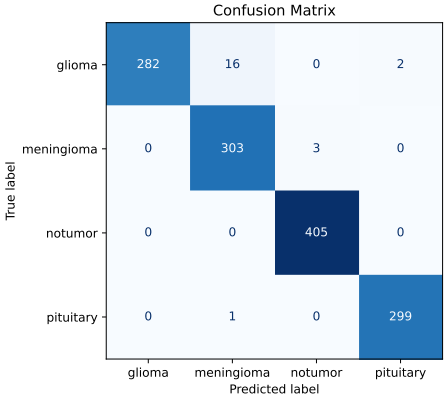
Confusion Matrix



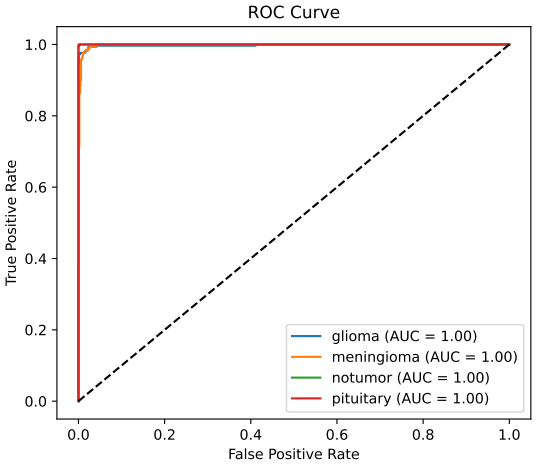
ROC Curve

Figure 8: EfficientNet-B0 – No Augmentation

With Data Augmentation



Confusion Matrix



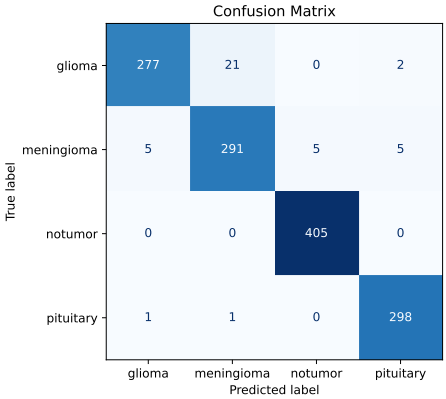
ROC Curve

Figure 9: EfficientNet-B0 – With Augmentation

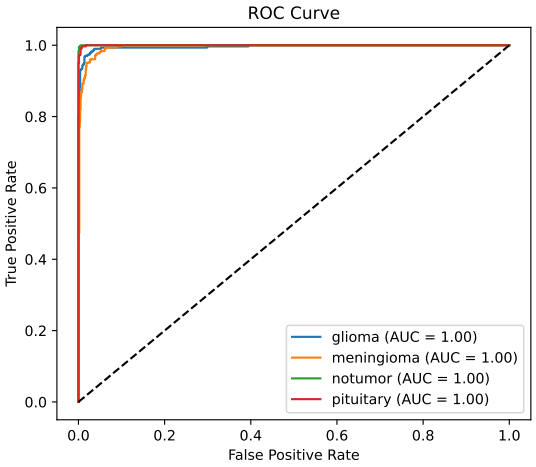
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.3. EfficientNet-B5

Without Data Augmentation



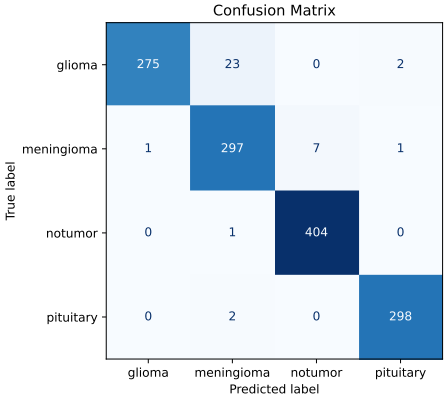
Confusion Matrix



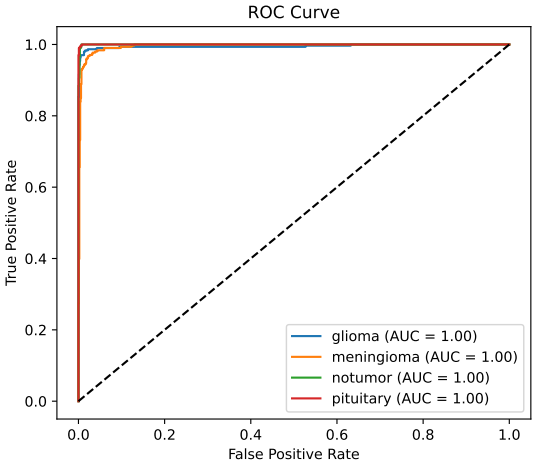
ROC Curve

Figure 10: EfficientNet-B5 – No Augmentation

With Data Augmentation



Confusion Matrix



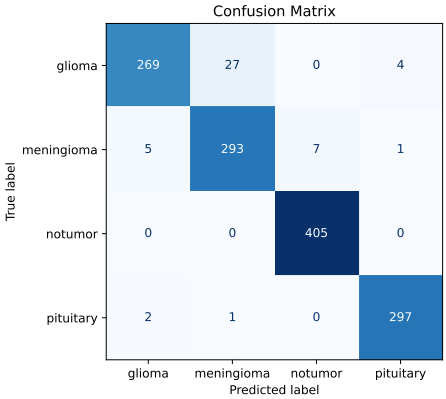
ROC Curve

Figure 11: EfficientNet-B5 – With Augmentation

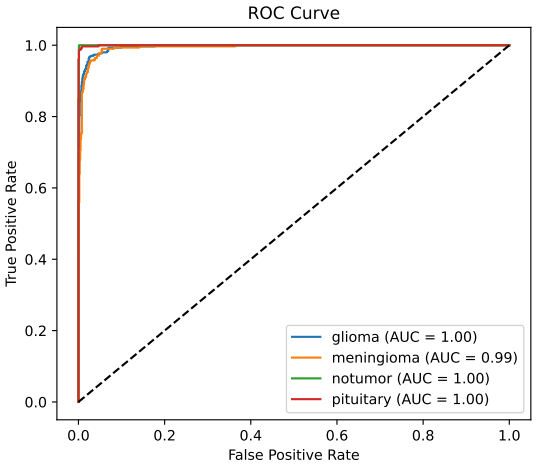
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.4. EfficientNet-B7

Without Data Augmentation



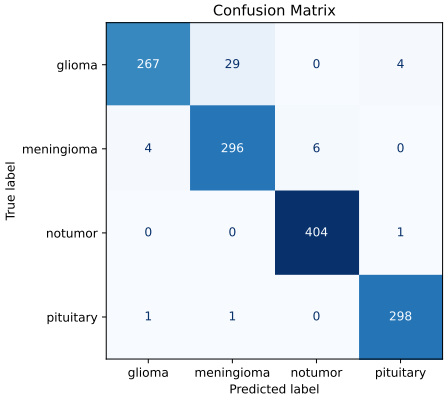
Confusion Matrix



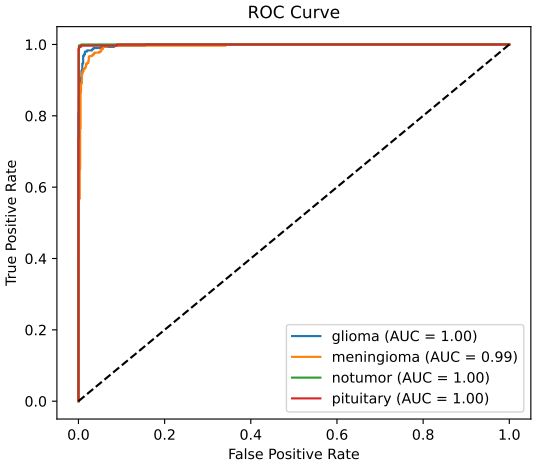
ROC Curve

Figure 12: EfficientNet-B7 – No Augmentation

With Data Augmentation



Confusion Matrix



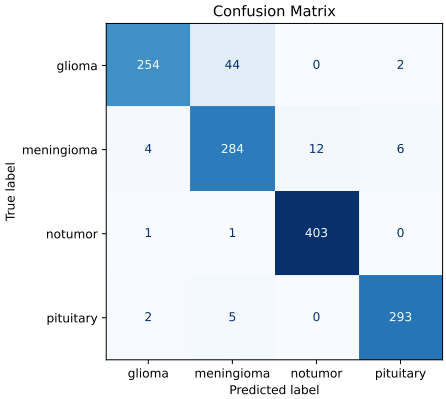
ROC Curve

Figure 13: EfficientNet-B7 – With Augmentation

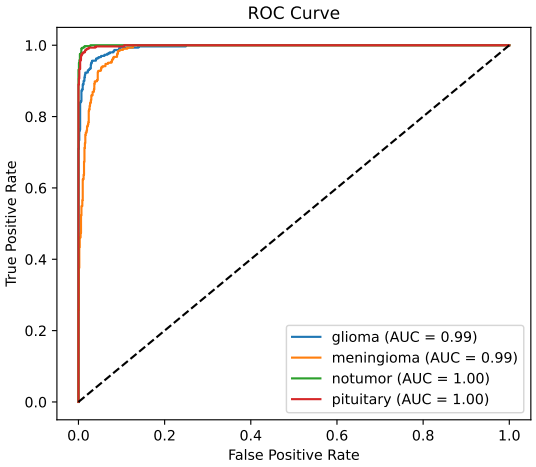
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.5. SwinV2-Base

Without Data Augmentation



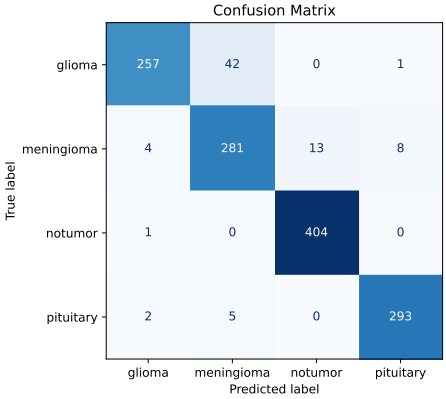
Confusion Matrix



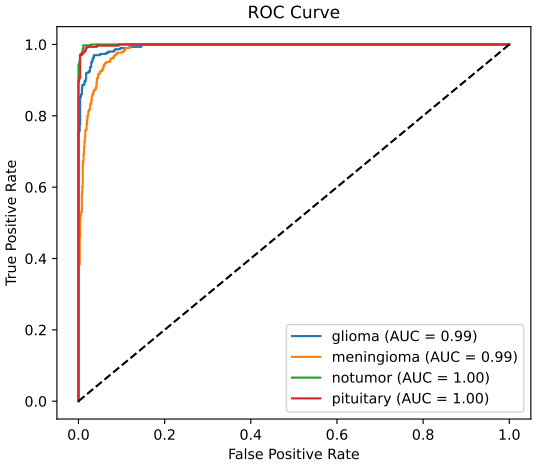
ROC Curve

Figure 14: SwinV2-Base – No Augmentation

With Data Augmentation



Confusion Matrix



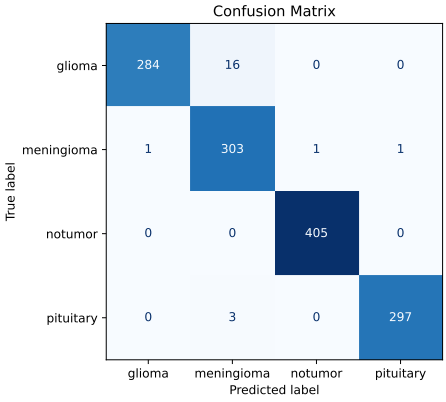
ROC Curve

Figure 15: SwinV2-Base – With Augmentation

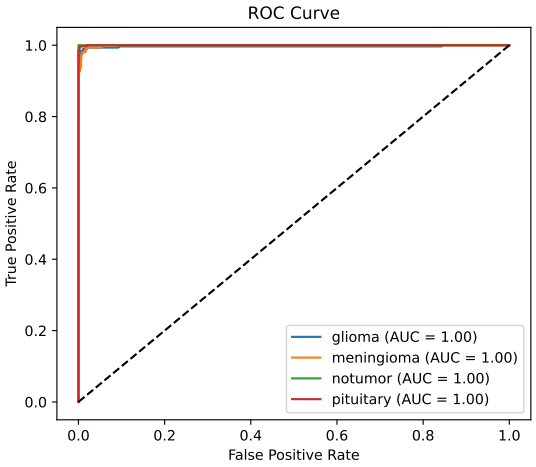
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.6. ViT-B/16

Without Data Augmentation



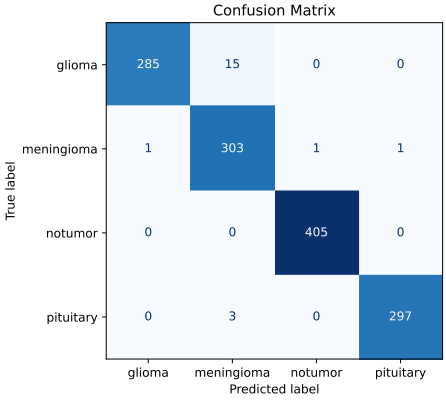
Confusion Matrix



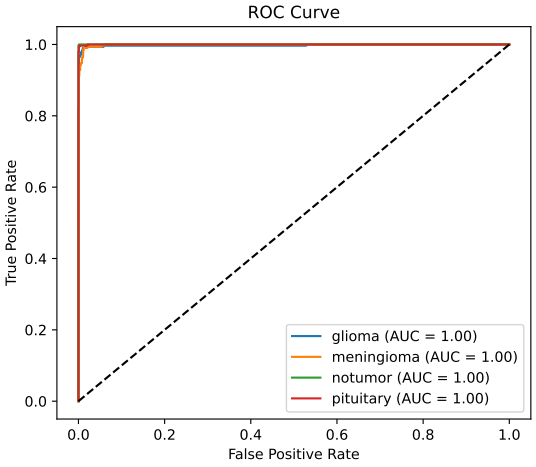
ROC Curve

Figure 16: ViT-B/16 – No Augmentation

With Data Augmentation



Confusion Matrix



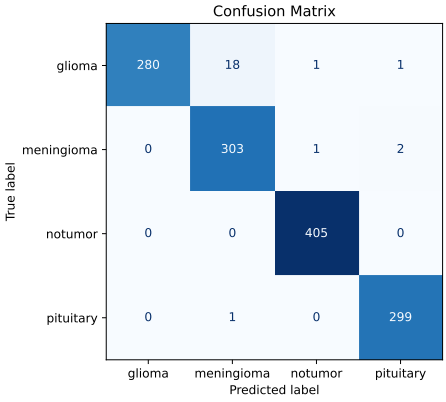
ROC Curve

Figure 17: ViT-B/16 – With Augmentation

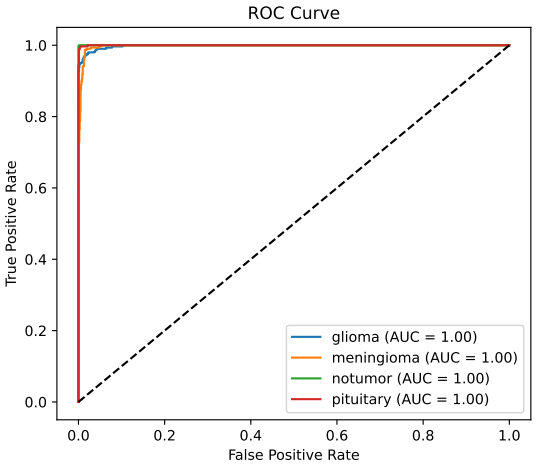
Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`

Appendix D.7. ViT-B/32

Without Data Augmentation



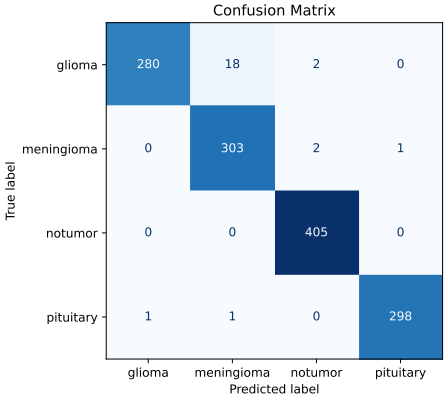
Confusion Matrix



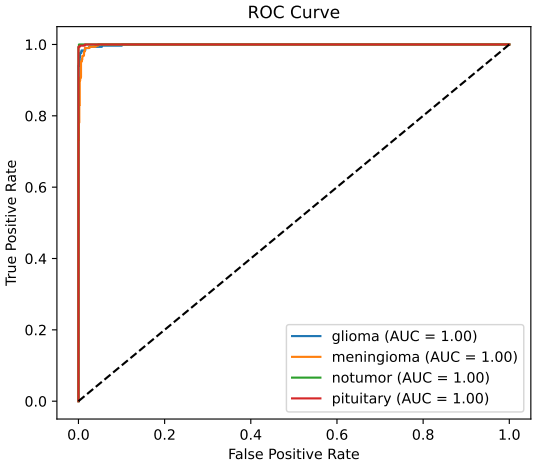
ROC Curve

Figure 18: ViT-B/32 – No Augmentation

With Data Augmentation



Confusion Matrix



ROC Curve

Figure 19: ViT-B/32 – With Augmentation

Supplementary materials and code available at:
`https://github.com/zaina-ml/brain-mri-classifier`