Automated Brain Tumor Classification from MRI: A Deep Learning Pipeline for Integrative Neuroimaging Research

By Lincoln Dibler

Project Summary:

This project delivers a **reproducible, ethics-aware deep learning pipeline** for brain tumor classification using publicly available MRI scans. Leveraging **transfer learning on the Brain Tumor MRI Dataset** (Kaggle), it identifies glioma, meningioma, pituitary, and non-tumorous cases with high accuracy. Designed with **interpretability, data consistency, and compliance** in mind, the pipeline is modular, version-controlled, and aligned with FAIR and GDPR principles—making it **well-suited for collaboration** across cognitive neuroscience and data science teams.

Background and Motivation:

Brain tumors present a significant clinical and cognitive challenge, with early and accurate diagnosis playing a crucial role in patient outcomes. Magnetic resonance imaging (MRI) remains a cornerstone in neurodiagnostics, yet manual interpretation is time-intensive and subject to variability. With the rise of **machine learning in medical imaging**, there is increasing potential for automated classification tools that support precision and efficiency.

This project was inspired by the intersection of **neuroscience, data science, and ethical research coordination**—core pillars in neurobiosocial research. By developing a **reproducible classification pipeline** for brain tumors using open-access MRI data, this initiative not only addresses a real-world need but also models the collaborative and standards-driven approach necessary for responsible neuroimaging research. It reflects a commitment to scientific openness, technical rigor, and interdisciplinary integration.

Data Description:

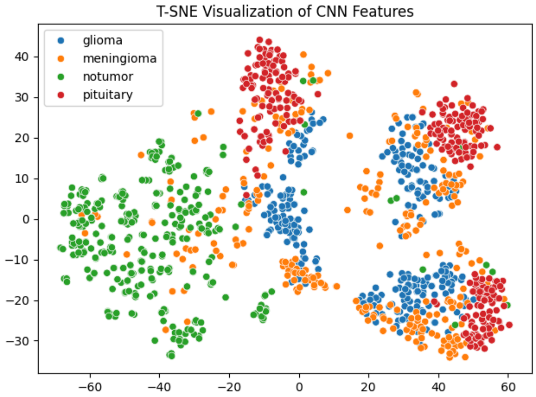
This dataset comprises **7023 brain MRI images** spanning four classes: *glioma, meningioma, pituitary tumor,* and *no tumor*. It is a curated blend of three public datasets—**Figshare, SARTAJ, and Br35H**. The "no tumor" samples originate from Br35H, while glioma images were selectively taken from Figshare due to mislabeling issues identified in the SARTAJ dataset. Brain tumors, whether benign or malignant, pose serious health risks; their early and accurate detection through MRI is vital. This study explores deep learning techniques—specifically CNN-based models—for multi-class classification and tumor localization to support reliable, automated diagnosis.

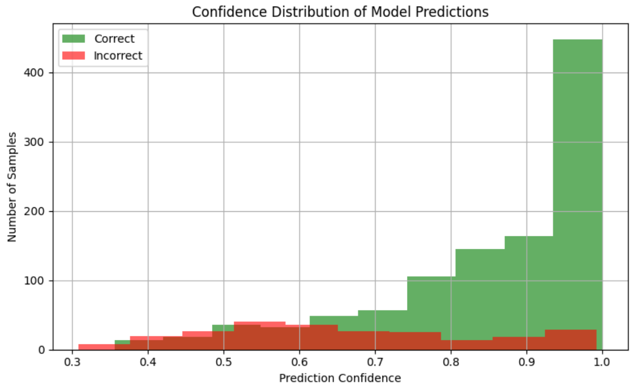
Modeling Approach:

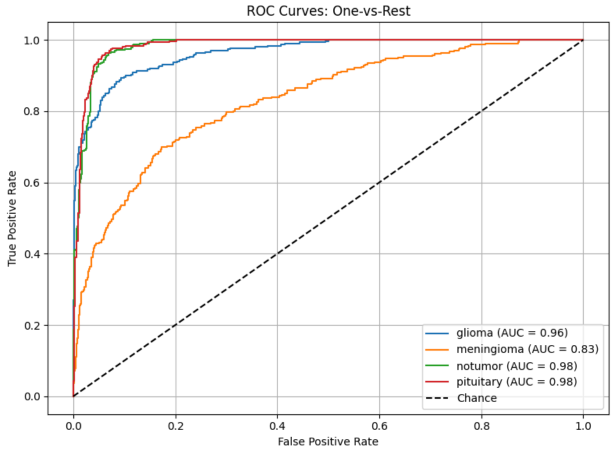
I employed a comparative deep learning approach by training and evaluating three popular convolutional neural network architectures—**MobileNetV2, ResNet50, and EfficientNetB0**—on a multiclass brain MRI classification task. Each model was fine-tuned using transfer learning with tailored input sizes, pretrained weights, and consistent augmentation strategies. I monitored performance across training, validation, and a held-out test set, using **early stopping and learning rate scheduling** to prevent overfitting. To enhance interpretability, I used **Grad-CAM visualizations** alongside class-specific metrics, offering clearer insights into how each model made its predictions.

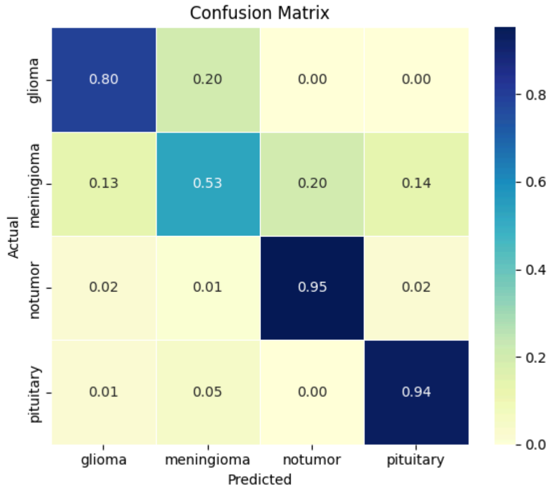
Results and Evaluation:

MobileNetV2: The MobileNetV2 model achieved a solid overall accuracy of **81%**, demonstrating particularly high recall for **no tumor** (0.95) and **pituitary tumor** (0.94) classifications. Precision and F1-scores were consistently strong across those classes, indicating both confidence and reliability. However, performance dropped significantly on **meningioma** cases, with a recall of just **0.53**—highlighting a tendency to misclassify them, possibly due to overlapping features with gliomas. This challenge was further supported by the t-SNE visualization, where meningioma samples formed diffuse, less separable clusters. Despite that, ROC curves revealed high AUC values **(>0.95)** for three out of four classes, reaffirming MobileNetV2's value as a fast, lightweight model with strong discriminative ability across most tumor types.





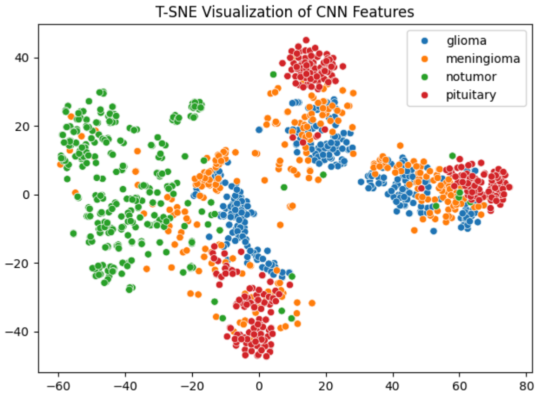


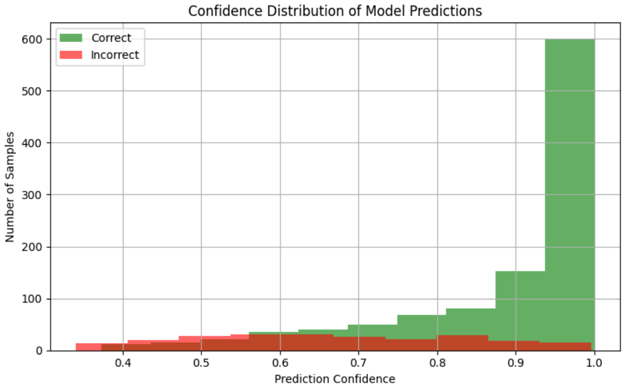

Confusion Matrix

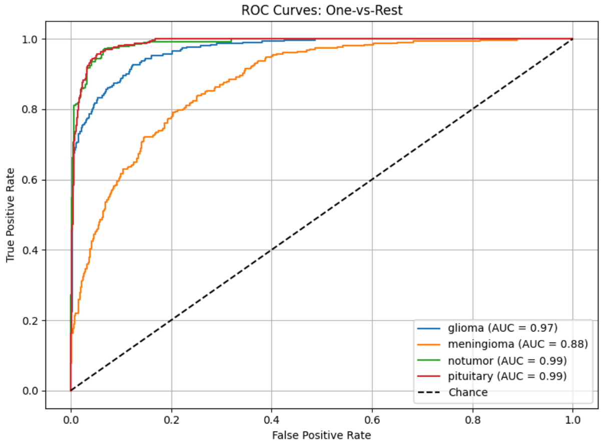
Actual \ Predicted	glioma	meningioma	notumor	pituitary
glioma	0.80	0.20	0.00	0.00
meningioma	0.13	0.53	0.20	0.14
notumor	0.02	0.01	0.95	0.02
pituitary	0.01	0.05	0.00	0.94

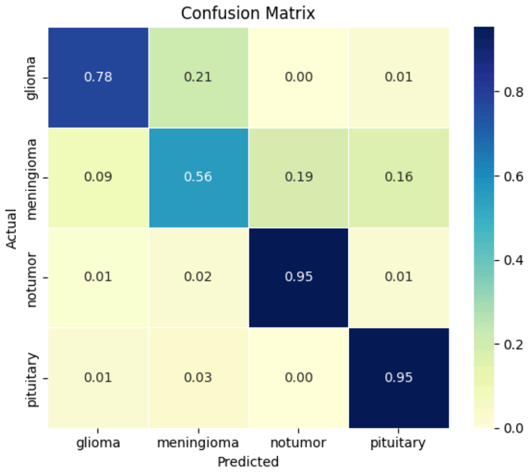
ResNet50: The ResNet50 model delivered an overall test accuracy of **82%**, with especially strong recall on **no tumor** (0.95) and **pituitary tumor** (0.95) samples, confirming its reliability in detecting both non-pathological scans and macroadenomas. It also showed notable improvement in **glioma** detection compared to MobileNetV2, achieving a balanced F1-score of 0.82. However, **meningioma** classification remained the weakest link, with a recall of just 0.56 and a more diffuse presence in the t-SNE projection — suggesting lingering confusion with glioma features. The ROC analysis echoed these trends, with **AUCs above 0.97** for most classes and a slightly lower but still respectable **0.88 for meningioma**. Prediction confidence was generally well calibrated: the majority of correct classifications occurred at higher confidence thresholds. These findings position ResNet50 as a powerful,

deep CNN backbone that captures nuanced tumor features while maintaining generalizability — particularly strong for critical true positive detections.

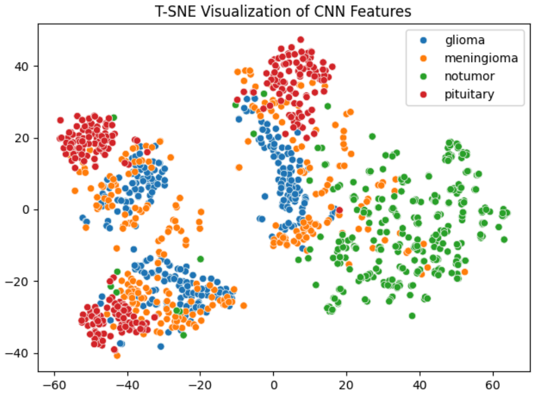


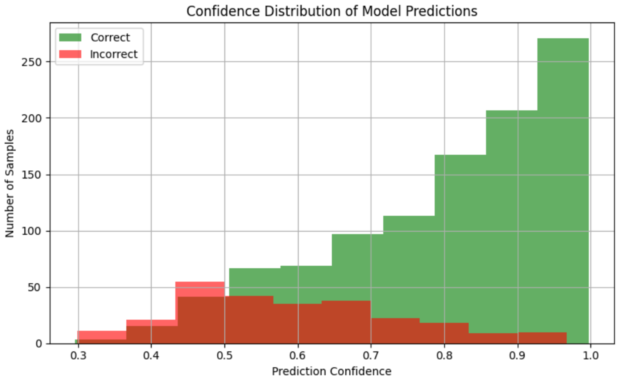


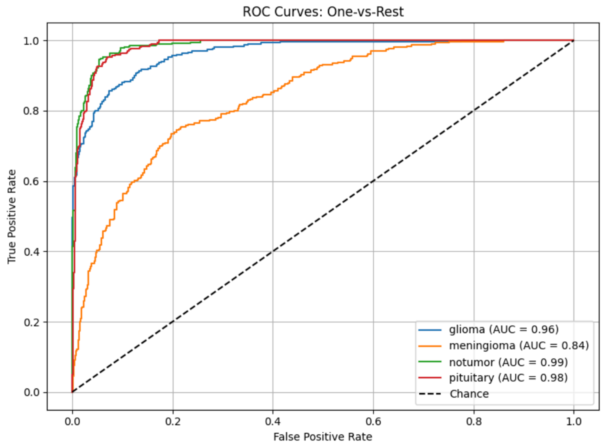


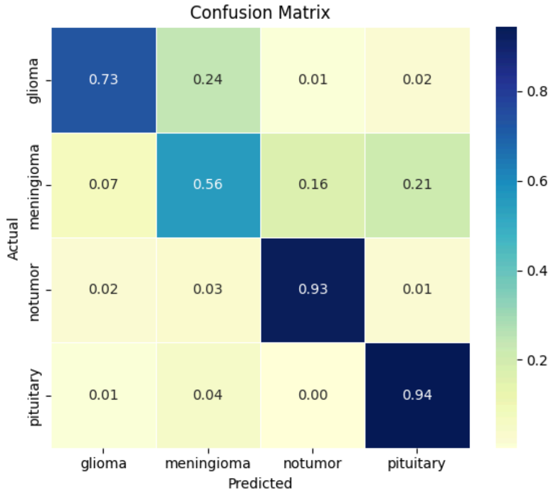


EfficientNetB0: The EfficientNetB0 model achieved an overall accuracy of **80%**, showing strong capability in identifying **no tumor** and **pituitary tumor** cases with recall values of **0.93** and **0.94**, respectively. These results align with the model's near-perfect AUC scores for those classes (**0.99** and **0.98**), confirming its confidence and reliability in high-stakes predictions. **Glioma** detection was moderately strong with an F1-score of **0.79**, but **meningioma** remained the most challenging tumor type, with a recall of only **0.56** and the lowest AUC at **0.84** — indicating that its features were harder for the model to separate. This pattern was echoed in the t-SNE embedding, where overlap between glioma and meningioma points suggested shared feature representations. Despite these challenges, EfficientNetB0 proved to be a compact yet competitive backbone, balancing high sensitivity in key classes with fast inference and efficient feature separation.









MobileNetV2 FineTuning: I fine-tuned the **last 10 layers of MobileNetV2** using a **low learning rate (5e-6)** and **class weighting to address imbalance**. Early stopping kicked in at Epoch 4, restoring weights from Epoch 1, suggesting limited benefit and signs of **overfitting**. Final test accuracy (**81.01%**) was similar to the frozen baseline. Minor augmentation was applied to improve meningioma recall, but had little effect on performance.

Grad-Cam Insights: Grad-CAM (Gradient-weighted Class Activation Mapping) is a widely used technique for interpreting CNN predictions by generating heat maps that highlight the most influential image regions. It does this by computing gradients of a target class with respect to the final convolutional layer, making it a post hoc, model-agnostic method. In this project, Grad-CAM was applied to test images across all three models **to assess whether predictions were grounded in meaningful tumor features**.

To better understand MobileNetV2's performance, especially on its weakest class — meningioma — Grad-CAM was used on representative test examples. In the correctly classified case (Figure 1), the model's attention was centered on the tumor, supporting a successful prediction. However, in the misclassified example (Figure 2), labeled as "no tumor," the heatmap emphasized peripheral brain areas, neglecting the tumor entirely. This likely led to the incorrect decision and illustrates how subtle visual cues or background bias can misguide predictions.

These examples emphasize the importance of visual interpretability in clinical AI, helping identify not just how well models perform, but *why* they succeed or fail.

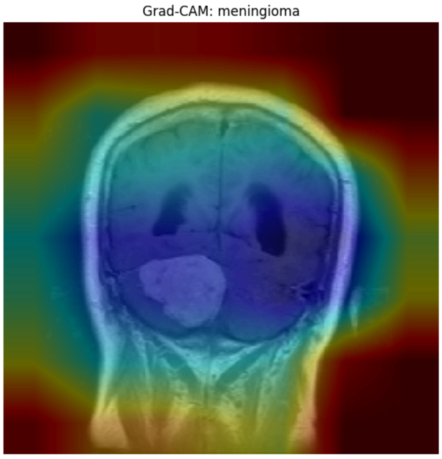


Figure 1

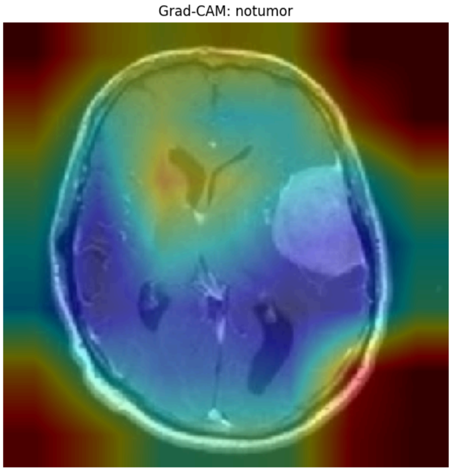


Figure 2

Interpretability and Insights:

Across all three **convolutional neural networks, interpretability analyses** revealed consistent patterns in model behavior, underscoring both their diagnostic promise and their limitations. All models achieved **strong performance on pituitary tumors and no tumor cases**, likely due to distinct visual features and less inter-class overlap — as confirmed by their tight clusters in t-SNE space and high ROC AUCs. In contrast, **meningioma consistently posed a challenge**: lower recall across models, diffuse t-SNE embeddings, and the lowest confidence calibration suggested feature ambiguity or **overlap with glioma presentations**. The ROC curves, while still above baseline, reflected this uncertainty with reduced AUC scores. These patterns hint at intrinsic similarities between certain tumor classes in the dataset, possibly **requiring enhanced feature disentanglement or refined labeling**. Importantly, the models' confidence distributions indicate that incorrect predictions tend to cluster at lower probability thresholds, suggesting that uncertainty-aware strategies — such as deferring low-confidence predictions for secondary review — could enhance clinical safety. Visual interpretability tools like Grad-CAM, which were applied during evaluation, c**onfirmed that models often attended to meaningful tumor regions** — though occasional focus on surrounding anatomy hinted at **possible reliance on spurious features.**

Limitations and Future Work:

This project was developed entirely in a **resource-constrained environment** using Google Colab and a consumer-grade Lenovo Legion laptop **without access to local GPU acceleration**. As a result, training time was limited, hyperparameter tuning was minimal, and larger architectures or ensemble techniques couldn't be fully explored. Despite these constraints, the final models were carefully benchmarked and analyzed, including detailed interpretability workflows using Grad-CAM.

Why This Matters:

Brain tumors are among **the most complex and life-altering conditions** a person can face—and **timely, accurate diagnosis** can mean the difference between life and loss. This project aimed to explore how deep learning could support that process, and the outcomes demonstrate that it's **not only feasible, but promising**. The system successfully learned to distinguish between multiple tumor types, provided meaningful visual explanations through Grad-CAM, and uncovered actionable patterns for improvement. With further development and larger-scale data, this work lays the groundwork for building a model that's both clinically informed and practically impactful.