Emerging Trends in Data Technology

Brain Tumour MRI Classification Using Deep Learning

Final Case Study - *Fundamentals of Medical Image Processing (on Brain Tumour MRI Dataset)*

By Shaheer Khan

I. **Introduction**

This project presents an end-to-end deep learning solution for the classification of brain MRI images into four clinically relevant categories: glioma, meningioma, pituitary tumor, and no tumor. The workflow is grounded in the Kaggle “Brain Tumor MRI Dataset” by Masoud Nickparvar and implements a series of professional image processing, augmentation, and explainability techniques as outlined in the notebook “Fundamentals of Medical Image Processing” by Neeraj Sharma. The objective is to demonstrate a reproducible, robust, and interpretable computer vision pipeline suitable for medical image analysis.

II. **Dataset Exploration**

The dataset consists of 7,313 grayscale MRI images, structured in separate folders for training and testing, and further organized by class.

*Training set: 5,712 images*

Glioma: 1,321  **|** Meningioma: 1,339  **|** Pituitary: 1,457 **|** No tumor: 1,595

*Testing set: 1,311 images*

Glioma: 300  **|** Meningioma: 306  **|** Pituitary: 300 **|** No tumor: 405

A preliminary visual inspection of random samples from each class ensured dataset integrity and diversity. The class distribution was slightly imbalanced, particularly for glioma and meningioma, which was later addressed with computed class weights during training.

A close-up of a brain scan

AI-generated content may be incorrect.

Figure 1 - Class visuals

III. **Preprocessing and Augmentation**

Robust preprocessing was crucial to prepare the images for neural network training:

Grayscale normalization and windowing: Each image’s pixel values were normalized using percentile windowing (1st and 99th percentiles), improving contrast and ensuring consistency.

Resizing and letterboxing: All images were resized and zero-padded to a fixed size of 224x224 pixels, preserving aspect ratio and facilitating efficient batch processing.

Data augmentation: A tf.keras.Sequential augmentation layer was defined, applying random flips, rotations, zooms, and contrast shifts to the training data. This increased model robustness and mitigated overfitting.

A close-up of a brain scan

AI-generated content may be incorrect.

Figure 2 - Augmented Data

**A graph of a number of lines

AI-generated content may be incorrect.**

Figure 3 – Colour Density Curves

**A graph of a graph showing different colored lines

AI-generated content may be incorrect.**

Figure 4 – Colour Density Curves (Zoomed In)

IV. **Data Splitting and Loader Design**

Although the original dataset was divided into training and testing sets, a portion of the training data was held out as a validation set for hyperparameter tuning and model selection. Data pipelines were implemented using TensorFlow’s tf.data API, with batched, shuffled, and pre-fetched datasets enabling fast training.

V. **Model Architecture**

A custom Convolutional Neural Network (CNN) was constructed using TensorFlow/Keras. The architecture consisted of several convolutional blocks, each followed by batch normalization and ReLU activation, with max pooling and dropout for regularization. The final dense layers mapped the feature representations to four output classes. The model was compiled with the Adam optimizer and categorical cross-entropy loss, with computed class weights used to address label imbalance.

VI. **Training and Validation**

The model was trained for 25 epochs, with early stopping and learning rate reduction on plateau implemented for optimal performance. Training and validation loss and accuracy were monitored each epoch. The best-performing model weights (based on validation loss) were saved and restored before final evaluation. The plotted training curves illustrated convergence, with validation metrics closely tracking training performance, indicating good generalization.

VII. **Evaluation and Results**

The trained model was evaluated on the held-out test set. Key performance metrics included:

Test accuracy: *Approximately 90.5%*

Test loss: *Approximately 0.28*

A detailed classification report was generated, listing precision, recall, F1-score, and support for each class. A confusion matrix visualized the model’s prediction strengths and weaknesses across categories, with most misclassifications occurring between glioma and meningioma, which share similar visual features in MRIs.

A diagram of a diagram

AI-generated content may be incorrect.

**VIII. Visualizations and Interpretability**

Prediction Samples

Randomly selected test images were displayed with both predicted and ground-truth labels, enabling qualitative assessment of the model’s predictions.

Training Curves

Accuracy and loss curves for both training and validation sets were plotted, providing a clear view of learning dynamics and helping diagnose any issues of overfitting or underfitting.

Confusion Matrix

A confusion matrix heatmap was generated for the test set, visually summarizing the model’s classification performance across all four tumor classes. The confusion matrix provided an intuitive breakdown of true positives, false positives, and misclassifications for each category. This allowed us to easily identify which tumor types were most often confused with one another.

A collage of images of a brain

AI-generated content may be incorrect.

IX. **Discussion and Reflection**

This project successfully implemented a complete workflow for the classification of brain MRI images using deep learning. The strong test accuracy (over 90%) highlights the effectiveness of careful preprocessing, data augmentation, and class balancing in building robust models even with moderately sized medical datasets. The use of a custom CNN, instead of large pre-trained networks, proved sufficient due to the high quality of preprocessing and augmentation, making the pipeline more lightweight and accessible.

The **confusion matrix** revealed valuable insights into the nature of classification errors. Misclassifications most frequently occurred between tumor types with visually overlapping characteristics, such as glioma and meningioma. This is expected in medical imaging, where subtle variations may not be easily captured by the model, especially in the absence of additional clinical context or multi-modal data. While the model performed well on “no tumor” and “pituitary” classes, improvement is possible for the other categories by integrating more advanced techniques or multi-stage models.

Reflecting on the process, several key lessons emerged:

**i. Data Quality Matters:** Ensuring consistency through windowing and normalization was crucial. Early visual sanity checks helped identify outliers or corrupted images.

**ii. Augmentation and Regularization:** Augmentation not only increased the effective size of the training set but also made the model more resilient to variations seen in real-world scans.

**ii. Explainability:** The confusion matrix was instrumental in understanding model behaviour and identifying systematic weaknesses. Although we did not implement Grad-CAM in this workflow, such methods are recommended for future work to provide additional clinical interpretability.

**iv. Iterative Debugging:** The project faced technical challenges (e.g., shape mismatches and data pipeline errors). Resolving these required step-by-step debugging, reinforcing the value of modular code and incremental validation in deep learning projects.

**v. Hardware Considerations:** Moving to a GPU runtime in Colab significantly accelerated model training, underlining the importance of appropriate computational resources for deep learning.

**vi. Limitations:** While the results are promising, the model’s performance is limited by the quality and diversity of the dataset. True clinical deployment would require external validation on independent datasets and possibly the integration of additional patient data (age, symptoms, scan modality). The model may also be sensitive to MRI acquisition differences (scanner brand, protocol), which were not explicitly addressed.

A graph of a graph showing a variety of data

AI-generated content may be incorrect.

A graph of a graph showing a variety of data

AI-generated content may be incorrect.

X. **Future Directions**

**i. Model Improvements:** Experimentation with transfer learning, ensemble methods, or deeper architectures could further boost performance, especially for challenging cases.

**ii. Explainability:** Implementing Grad-CAM or other saliency map methods would make the model’s predictions more transparent, aiding clinician acceptance.

**iii. Deployment:** Creating a web-based inference tool or integrating the pipeline into a clinical decision support system are potential next steps.

XI. **Personal Reflection**

Working through the full pipeline—from raw data to actionable results—provided a comprehensive understanding of the challenges and best practices in medical AI. The iterative, cell-by-cell debugging process, inspired by the referenced notebook, proved highly effective for both learning and reproducibility. The project demonstrates not only the technical feasibility of automated brain tumor classification but also the importance of transparency, collaboration, and ongoing refinement in medical AI development.

XII. **References**

Masoud Nickparvar, Brain Tumor MRI Dataset, Kaggle (2020)

Neeraj Sharma, Fundamentals of Medical Image Processing, Kaggle Notebook (2021)

Chollet, F. et al., Deep Learning with Python (2021)