Brain Tumour Classification: A Comparative Analysis

Tabular Clinical Data vs. MRI Imaging

Team: Group 2

Razin Mohammed

Dhruv Roshan

Keerthana Nair

Temisola Olajide

F20DL Data Mining and Machine Learning

The Problem:

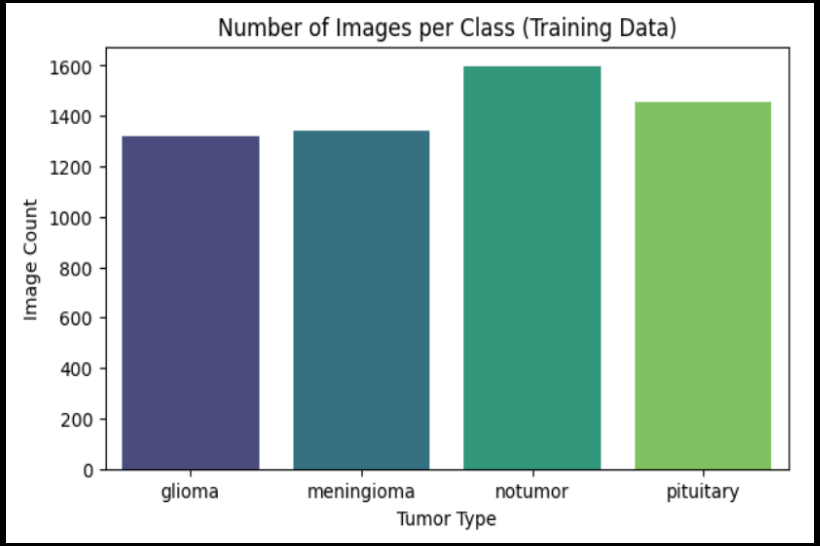
- Brain tumours are a major cause of cancer-related deaths.
- Manual MRI diagnosis is time-consuming and prone to human error.

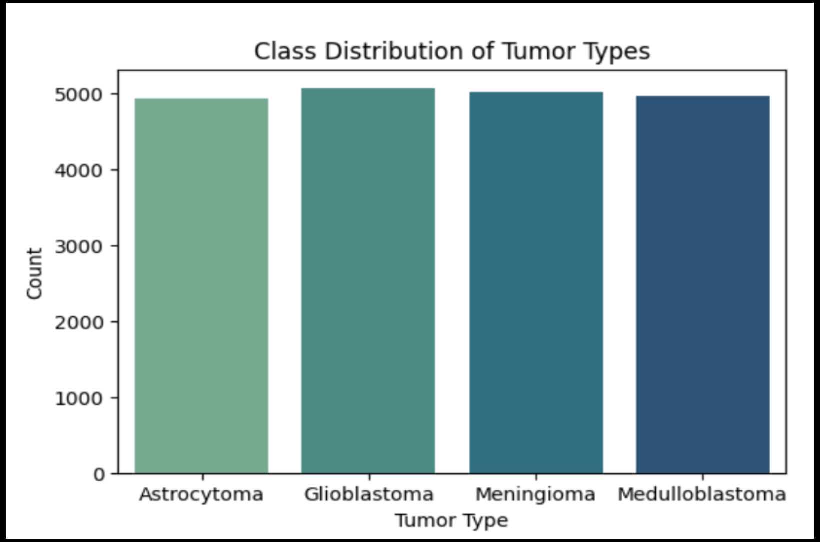
Our Objectives:

- Compare performance between **Classical ML** (Tabular Data) and **Deep Learning** (MRI Images).
- Perform rigorous EDA and Data Cleaning.
- Implement a CNN to automate tumour detection.

Data Overview & Preprocessing

- **Dataset 1: Clinical Tabular Data**
    - **Size:** 2,000 records, 19 features (Age, Tumour Size, etc.).
    - **Findings:** Low correlation between features; no single feature strongly predicted the tumour type.
- **Dataset 2: MRI Image Data**
    - **Size:** 7,023 images across 4 classes (Glioma, Meningioma, Pituitary, No Tumour).
    - **Preprocessing:** Resized to 128×128, normalized, and augmented (rotation/flipping).
- **Analysis:** PCA & K-Means clustering (k=4) on images showed clear grouping, suggesting images have better separability than tabular data.





Classical ML Approach (Tabular)

- **Models Implemented:**
  - **k-Nearest Neighbours (KNN):** Best k=15 (Manhattan distance).
  - **Naïve Bayes:** Gaussian assumption (Baseline).
  - **Perceptron:** Linear classifier with L2 regularization.
  - **Random Forest:** Ensemble method (200 estimators).
- **Regression Analysis**
  - **Goal:** Predict continuous variables (Survival Rate).
  - **Models:** Linear Regression, Ridge, Lasso, SVR, and Random Forest Regressor
- **The Challenge:**
  - Framed as a multi-class classification task.
  - Features showed weak linear relationships.

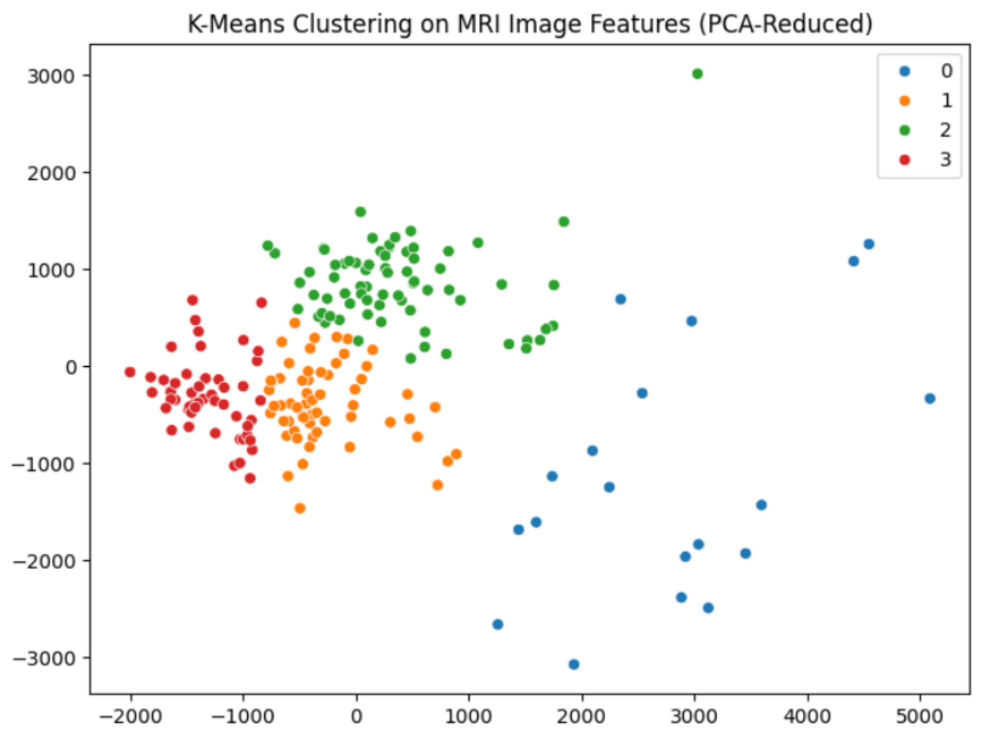


Figure 8: k-means clustering ($k = 4$) on PCA-reduced MRI features.

Deep Learning Approach (CNN)

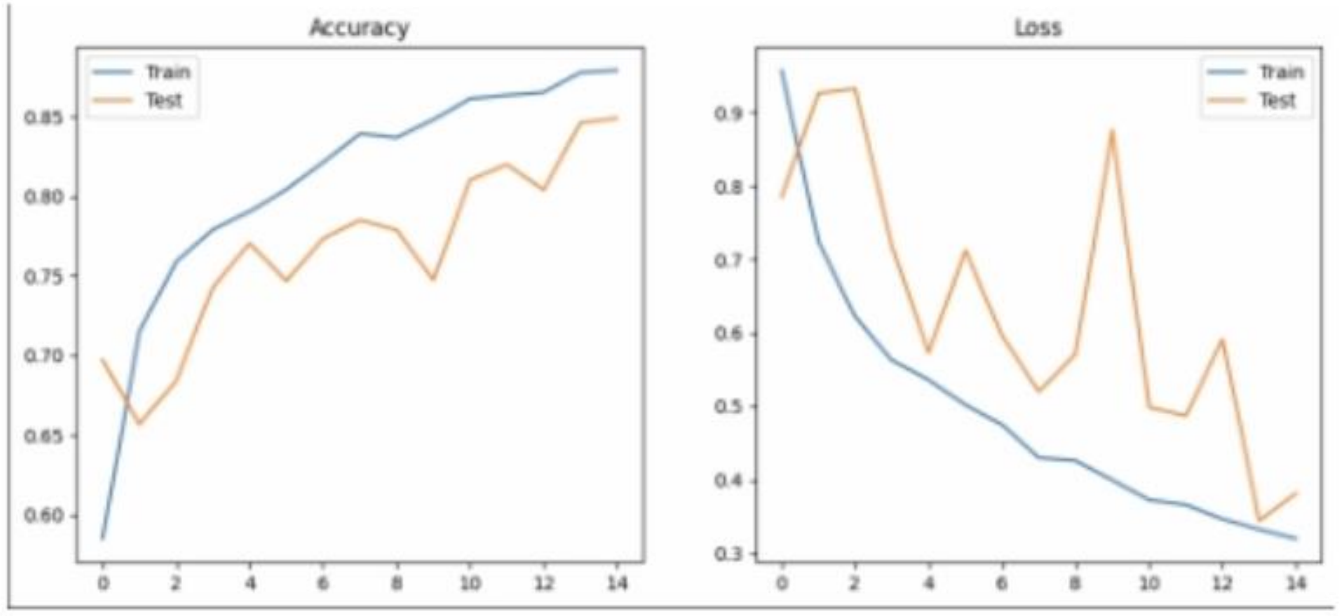
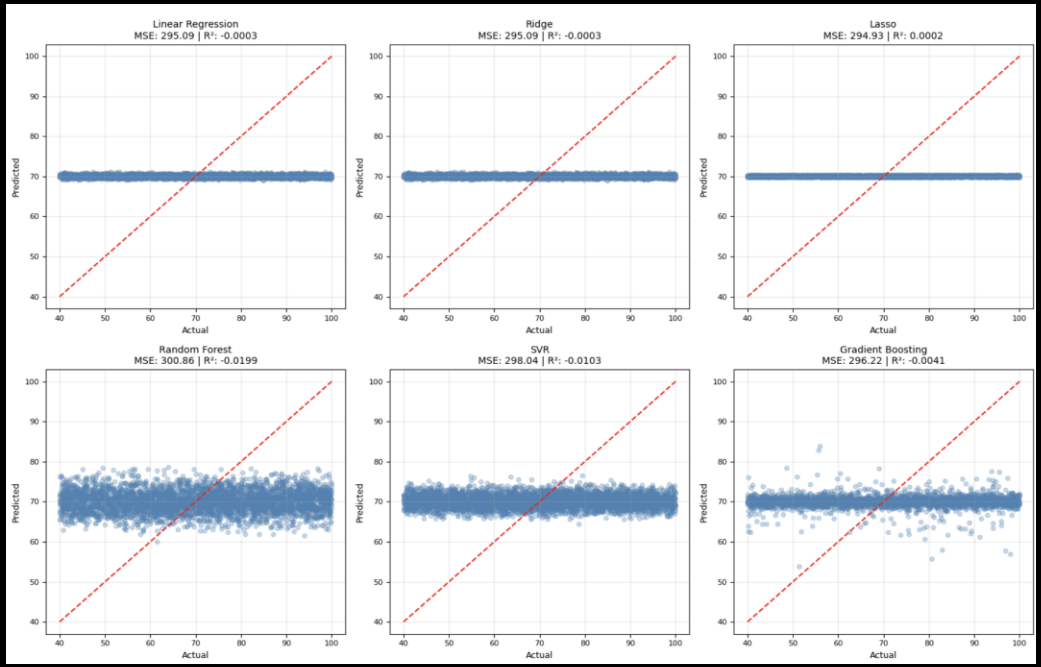


Figure 9: CNN training and validation accuracy/loss curves.

- **Architecture:**
  - **Input:** 128×128 MRI scans.
  - **Layers:** Convolutional layers with ReLU activation + Max Pooling.
  - **Output:** Softmax layer for 4 classes.
- **Training:**
  - Optimizer: Adam.
  - Loss Function: Categorical Cross-Entropy.
  - Technique: Early stopping to prevent overfitting.

Key Results & Comparison



- **Regression Analysis (Tabular Data):**
  - **Objective:** Attempted to predict continuous clinical variables (e.g., Survival Rate).
  - **Outcome:** The models failed to capture data variance.
- **Evidence:** Residual plots show high errors, and "Predicted vs. Actual" graphs show flat-line predictions, indicating the models simply predicted the mean rather than learning relationships.

- **Tabular Models (Classical ML):**
  - **Performance:** ~25-26% Accuracy (Poor).
  - **Insight:** The models performed near random-guessing levels, confirming the tabular data lacks predictive power.

- **CNN (MRI Images):**
  - **Performance: 85% Accuracy**.
  - **Class Breakdown:**
    - "No Tumour" & "Pituitary": Excellent recall (1.00 and 0.97).
    - "Meningioma": Most difficult class (0.63 recall) due to visual overlap.

Conclusion & Future Work

- **Conclusion:**
  - **Modality Matters:** Clinical features alone (age, size) are insufficient for histology prediction. MRI imaging is essential.
  - **CNN Superiority:** Deep learning successfully extracted spatial patterns that classical models missed.
- **Future Work:**
  - **Improvement:** Deeper architectures to improve Meningioma detection.
  - **Feature Engineering:** Attempt to extract new features from the images to enrich the tabular dataset.