Brain Tumour Classification Using Tabular Clinical Data and MRI Images

Group 2
Razin Mohammed: H00414442
Dhruv Roshan: H00459030
Keerthana Nair: H00421150
Temisola Olajide: H0033387173

Github Link: `https://github.com/DhruvR-HWUD/DMML-Dubai-UG-Group-2`
F20DL – Data Mining and Machine Learning

1 Introduction

A brain tumour is an abnormal growth of cells within or around brain tissue, posing major clinical and societal challenges due to high mortality, low survival rates, and diagnostic complexity. In Iraq, brain and central nervous system cancers were the fourth leading cause of cancer-related death in 2020, highlighting the global severity of these conditions. Diagnosis traditionally depends on radiologists manually examining Magnetic Resonance Imaging (MRI) scans—a process that is time-consuming, requires specialised expertise, and is susceptible to human error.

With the increasing availability of medical images and clinical records, machine learning (ML) and deep learning (DL) techniques have become important tools for supporting automated tumour detection. Prior work has demonstrated the effectiveness of convolutional neural networks (CNNs) for MRI-based tumour classification. Badža and Barjaktarović showed that CNNs can accurately distinguish between glioma, meningioma, and pituitary tumours [1], while Rethemiotaki et al. extended this to four classes, including healthy cases, achieving strong, clinically relevant performance [4]. These findings motivate evaluating CNNs alongside classical ML methods.

This project investigates brain tumour classification using both MRI images and tabular clinical data. The objectives are to: (1) use publicly available datasets for classification, (2) perform data cleaning and exploratory analysis including clustering, (3) train and evaluate multiple classical ML algorithms on the clinical dataset, and (4) implement a CNN for MRI classification and compare its performance to the classical baselines.

The remainder of this report is structured as follows: Section 2 describes the datasets and exploratory analysis, Section 3 outlines the experimental setup for classical models and the CNN, Section 4 presents the results, Section 5 provides a brief discussion, and Section 6 concludes the work. An appendix summarises group contributions.

2 Dataset Description and Analysis

2.1 Datasets Used

In this project we used two main datasets: one tabular clinical dataset and one MRI image dataset. This design allows us to demonstrate both classical ML methods (on tabular data) and a CNN (on image data).

Table 1: Summary of datasets used in the project.

Dataset	Source	Type	Size
Brain Tumour MRI Dataset [3]	Public Kaggle brain tumour MRI dataset	Image	7,023 images
Brain Tumour Clinical Dataset [2]	Public Kaggle clinical brain tumour dataset	Tabular	2,000 rows, 19 features

2.2 Pre-processing and EDA: Tabular Data

The clinical dataset (2,000 records, 19 features) was inspected to verify feature types, ranges, and data quality. No duplicates, placeholder tokens, or hidden null values were found, and numerical attributes such as Age and Tumour Size were clinically plausible. Aside from removing identifier fields (e.g., `Patient_ID`), no further cleaning was required.

Boxplots for Age and Tumour Size indicated well-behaved distributions without severe skew. Label-distribution plots showed that the four histology classes appeared in roughly equal proportions, confirming that the classification task is balanced.

A correlation heatmap revealed generally low pairwise correlations among numerical variables and between features and the histology label, implying that linear relationships are weak and that any predictive signal likely arises from non-linear interactions. Additional group-wise summaries examined gender, tumour location, stage, follow-up status, MRI results, and family history across classes. Continuous features (e.g., tumour growth, survival rate) were summarised per histology group.

Feature-target evaluation included ANOVA F-tests, which identified Survival Rate as the only individually significant numeric feature, and chi-square tests on categorical variables, which highlighted limited but present differences across tumour types. Overall, the EDA suggests that the tabular dataset contains weak signal for classical ML models, consistent with the modest results reported later.

2.3 Pre-processing and EDA: Image Data

The MRI dataset was organised into train and test folders with four classes (glioma, meningioma, pituitary, no tumour). All images were resized to a fixed resolution (128×128) and normalised. A simple preprocessing pipeline iterated through each image, loaded it using OpenCV, applied resizing, and stored a cleaned version in a standardised directory structure.

Class counts were visualised to verify balanced representation. Random samples from each class were inspected to confirm label correctness and check for preprocessing artefacts.

To probe intrinsic structure, selected images were downsampled to grayscale thumbnails, flattened, reduced with PCA, and clustered using k-means ($k = 4$). The PCA plot showed clear grouping patterns between tumour types, indicating stronger separability in the image dataset than in the clinical dataset and motivating the use of CNNs for this modality.

3 Experimental Setup

3.1 Classical Machine Learning Models (Tabular Data)

For the clinical dataset, we framed histology prediction as a multi-class classification task. A standard train/test split was used (details such as exact split ratio and random seed should be inserted here if required). Features were constructed by dropping identifier columns and using all remaining attributes as input.

We implemented four main classical ML models:

- k-Nearest Neighbours (KNN)

- Naïve Bayes
- Perceptron (linear classifier)
- Random Forest

Table 2 summarises the high-level experimental setup for each model, based on the group's configuration grid.

Table 2: Experimental setup for baseline classical models on the tabular dataset.

Method	**Role**	**Key Settings Explored**	**Motivation**
KNN	Distance-based classifier	Tried different k values (5, 7, 15, 20); Euclidean and Manhattan distances; uniform vs distance-based neighbour weights.	To assess how a simple non-linear, instance-based model performs on the dataset.
Naïve Bayes	Probabilistic baseline	Gaussian Naïve Bayes with different `var_smoothing` values (10^{-9}, 10^{-8}, 10^{-7}).	To provide a very fast baseline and see how well simplistic independence and Gaussian assumptions fit the data.
Perceptron	Linear classifier	Varied `max_iter`; experimented with L2 and ElasticNet penalties, different `alpha`, with/without early stopping, and alternative learning rate settings.	To understand how a simple linear decision boundary behaves and how regularisation and training parameters affect convergence.
Random Forest	Ensemble of decision trees	Adjusted number of trees (200–400), maximum tree depth, and split criteria (Gini vs entropy).	To capture more complex, non-linear patterns using a robust ensemble method.
Evaluation metrics	Model assessment	Used accuracy, macro precision, macro recall, macro F1-score, confusion matrices, and classification reports.	To evaluate multi-class performance fairly and compare models consistently, especially under class imbalance.

All models were implemented in Python using standard ML libraries (e.g., `scikit-learn`), and trained on the same training split to ensure fair comparison.

3.2 Convolutional Neural Network (CNN) (MRI Images)

The MRI dataset was used to train a CNN for multi-class image classification over the four tumour classes (glioma, meningioma, pituitary, no tumour). Images were preprocessed as described in Section 2, with a fixed input resolution (e.g., 128×128 pixels) and appropriate normalisation. .

4 Results

4.1 Tabular Models

Tables 3–6 summarise the performance of the four classical models on the test set, reporting accuracy, macro precision, macro recall, and macro F1-score.

Table 3: KNN configurations and performance on the clinical dataset.

Model	k	Weights	Dist.	Acc.	Macro Prec.	Macro Rec.	Macro F1
1 (Baseline)	5	Uniform	Euclidean ($p = 2$)	0.255	0.25	0.25	0.21
2	7	Distance	Euclidean ($p = 2$)	0.256	0.25	0.26	0.22
3	15	Uniform	Manhattan	0.245	0.25	0.25	0.24
4	20	Distance	Euclidean ($p = 2$)	0.252	0.24	0.25	0.21

Among the four KNN configurations, Model 3 ($k = 15$, Manhattan distance, uniform weights) achieves the highest macro F1-score and provides the most balanced performance across histology classes, even though its accuracy is slightly lower than Model 2. Given that macro F1 is more informative for multi-class problems, Model 3 was selected as the preferred KNN configuration.

Table 4: Naïve Bayes configurations and performance on the clinical dataset.

Model	`var_smoothing`	Acc.	Macro Prec.	Macro Rec. / Macro F1
1 (Baseline)	10^{-9}	0.261	0.26	0.26 / 0.248
2	10^{-8}	0.257	0.26	0.26 / 0.244
3	10^{-7}	0.256	0.25	0.25 / 0.243

For Naïve Bayes, the default GaussianNB configuration (Model 1) achieves the highest accuracy (0.261) and macro F1 (0.248). Increasing `var_smoothing` leads to slightly more stable variance estimates but marginally worse performance. Overall, all Naïve Bayes models perform similarly and at a relatively low level, reflecting the mismatch between the independence/Gaussian assumptions and the structure of the data.

Table 5: Perceptron configurations and performance on the clinical dataset.

Model	**Key hyperparameters**	Acc.	Macro Prec.	Macro Rec.	Macro F1
1	`max_iter = 2000`	0.258	0.25	0.26	0.235
2	`max_iter = 3000`, L2, `alpha = 1e-4`, early stopping	0.265	0.26	0.26	0.222
3	`max_iter = 5000`, ElasticNet, `l1_ratio = 0.3`	0.251	0.06	0.25	0.100
4	`max_iter = 2000`, L2, `alpha = 1e-3`, `eta0 = 0.5`	0.257	0.25	0.25	0.216

Model 1 provides the best balance for the Perceptron with a macro F1 of 0.235 and stable performance across classes. Model 2 attains the highest accuracy (0.265) but suffers from very low recall for at least one class, reducing its macro F1. Model 3 collapses to near single-class predictions, giving a macro F1 of 0.100, while Model 4 is intermediate. Thus, Model 1 is identified as the most reliable Perceptron configuration.

Across the Random Forest configurations, Model 1 delivers the best overall performance with the highest accuracy (0.251) and a strong macro F1 of 0.25. Models 2 and 3 perform similarly but do not surpass Model 1, which is therefore selected as the preferred Random Forest model.

Table 6: Random Forest configurations and performance on the clinical dataset.

Model	Key hyperparameters	Acc.	Macro Prec.	Macro Rec.	Macro F1
1	200 estimators, `max_depth = None`, Gini, no class weights	0.251	0.25	0.26	0.25
2	300 estimators, `max_depth = 20`, entropy, no class weights	0.246	0.25	0.25	0.24
3	400 estimators, `max_depth = 15`, Gini, no class weights	0.245	0.25	0.25	0.25

4.2 Convolutional Neural Network (CNN) for MRI Classification

The MRI dataset was used to train a deep learning model for four-class tumour classification (glioma, meningioma, pituitary, and no tumour). All images were resized to a fixed resolution and normalised to stabilise training. A convolutional neural network was adopted because CNNs are well-suited for extracting spatial and texture features from medical images.

The model consisted of multiple convolutional layers with ReLU activations, each followed by max-pooling for spatial downsampling, and two fully-connected layers leading to a softmax output over the four classes. Data augmentation (random rotation, flipping, and zooming) was applied to reduce overfitting and increase generalisation. The network was trained using categorical cross-entropy loss with the Adam optimiser, a fixed learning rate, and a validation split. Early stopping was used to prevent unnecessary epochs once validation performance stabilised.

Evaluation metrics included accuracy, precision, recall, macro F1-score, and qualitative review of the confusion matrix. These metrics were selected to provide detailed insight into class-level behaviour, especially because tumour datasets often show visual overlap between classes.

4.3 CNN Results

The CNN achieved strong performance on the test set, with an overall accuracy of 0.85 across 1,311 MRI images. Table 7 summarises class-level metrics.

Table 7: CNN classification performance on MRI dataset.

Class	Precision	Recall	F1-score	Support
Glioma	0.95	0.75	0.84	300
Meningioma	0.73	0.63	0.68	306
No tumour	0.84	1.00	0.91	405
Pituitary	0.88	0.97	0.93	300
Overall Accuracy		0.85		
Macro Average	0.85	0.84	0.84	–

The model showed excellent performance for the *no tumour* (recall = 1.00) and *pituitary* (recall = 0.97) classes, indicating that nearly all samples in these categories were correctly classified. Glioma also achieved high precision (0.95), though its recall (0.75) suggests some misclassification into pituitary or no-tumour categories. Meningioma proved the most challenging class, with the lowest recall (0.63), consistent with existing literature where meningioma often exhibits high intra-class variability and visual similarity to other tumour types.

Although numerical confusion matrix values were not printed, the classification report implies strong diagonal dominance for no-tumour and pituitary classes, with more mixed predictions for glioma and meningioma.

Overall, the CNN forms a strong baseline for MRI tumour classification and significantly outperforms the classical machine learning models trained on tabular features. The results highlight the value of spatial information in medical imaging and reinforce CNNs as the appropriate choice for this task, though further improvements (e.g., deeper architectures, transfer learning, Grad-CAM interpretability) could enhance meningioma detection.

5 Brief Discussion

The classical models on the clinical dataset all achieved modest performance, with accuracies around 0.25–0.26 and macro F1-scores around 0.23–0.25. This is consistent with the exploratory analysis, which indicated low individual feature relevance and relatively weak linear correlations with the histology label. The models that are more flexible in capturing non-linear interactions (KNN with $k = 15$ and the Random Forest with 200 trees) yielded the most balanced macro F1-scores, but still did not reach high predictive performance.

Naïve Bayes provided a fast probabilistic baseline but was constrained by its independence and Gaussian assumptions, which are not well aligned with the clinical data structure. The Perceptron, as a linear classifier, benefited from careful tuning of optimisation parameters, but its best configuration still performed similarly to the other models, reflecting the difficulty of linearly separating the classes in feature space.

In contrast, the CNN trained on MRI images (Section 4.3) leveraged rich spatial information from the scans and therefore achieve higher performance than the tabular models, in line with prior work [1, 4].

Overall, the experiments highlight the importance of matching model class to data modality: while clinical features alone appear insufficient for high-accuracy histology prediction in this dataset, MRI-based CNNs offer a more promising path for accurate tumour type classification.

6 Conclusion

This project implemented a full data mining and machine learning pipeline for brain tumour classification using both clinical tabular data and MRI images. We performed data cleaning and exploratory analysis on the clinical dataset, including distribution checks, correlation analysis, and basic feature-target evaluation. For the imaging data, we built a reproducible preprocessing pipeline, standardised image resolutions, and explored intrinsic structure via PCA and k-means clustering.

On the tabular dataset, we trained and evaluated KNN, Naïve Bayes, Perceptron, and Random Forest models. All models achieved relatively low but consistent performance, reflecting the limited predictive strength of the available clinical features under the chosen configurations. Random Forest and KNN provided the most balanced macro F1-scores, while Naïve Bayes and Perceptron served as useful baselines.

In contrast, the CNN trained on MRI scans achieved strong multi-class accuracy and provided clear advantages due to the rich spatial information available in imaging data.

Appendix A: Group Member Contributions

- **Razin**: Data collection, Perceptron, k-means clustering, Naive Bayes Random Forest, result visualisation.
- **Temisola**: Data collection, EDA and data preprocessing for all data, report compilation, literature review.
- **Dhruv**: Data collection, Regression, result visualisation, CNN implementation, Github Management.
- **Keerthana**: Data collection, CNN implementation.

Appendix B: Generative AI Use Statement

Generative AI tools were used to support text proofreading and minor rephrasing of sections drafted. All modeling decisions, code, and core analysis were carried out by the group. Generative AI also assisted with LaTex formatting.

Appendix C: Graphs and Images

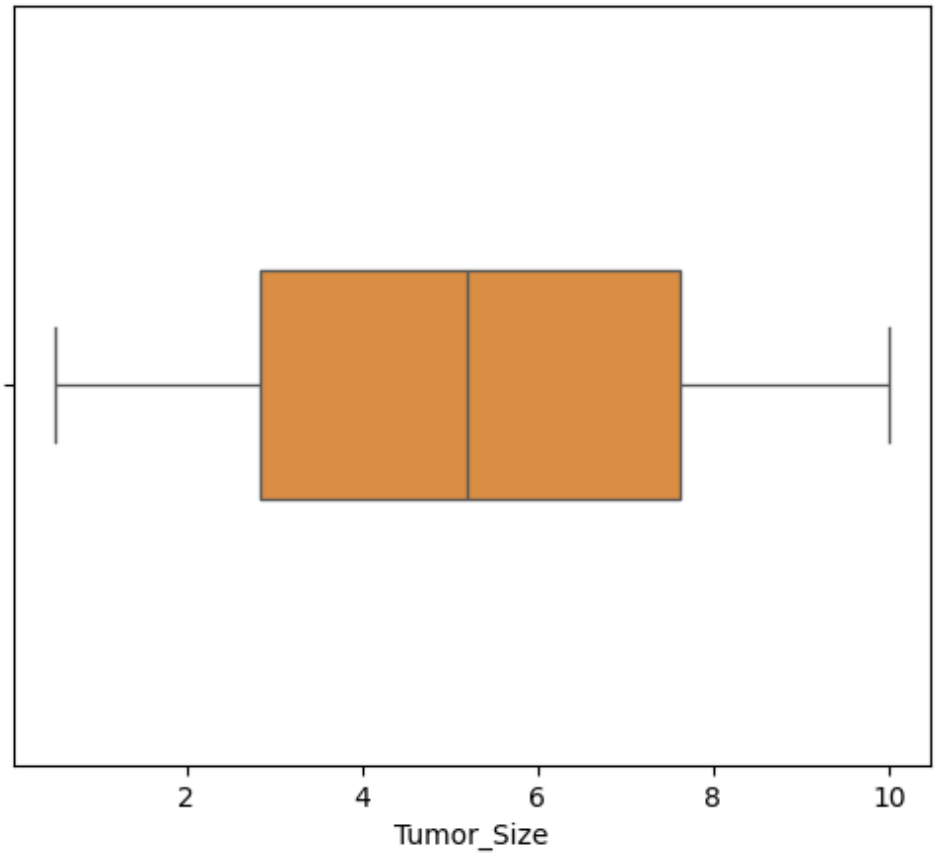

Figure 1: Boxplot of Tumour Size across patients.

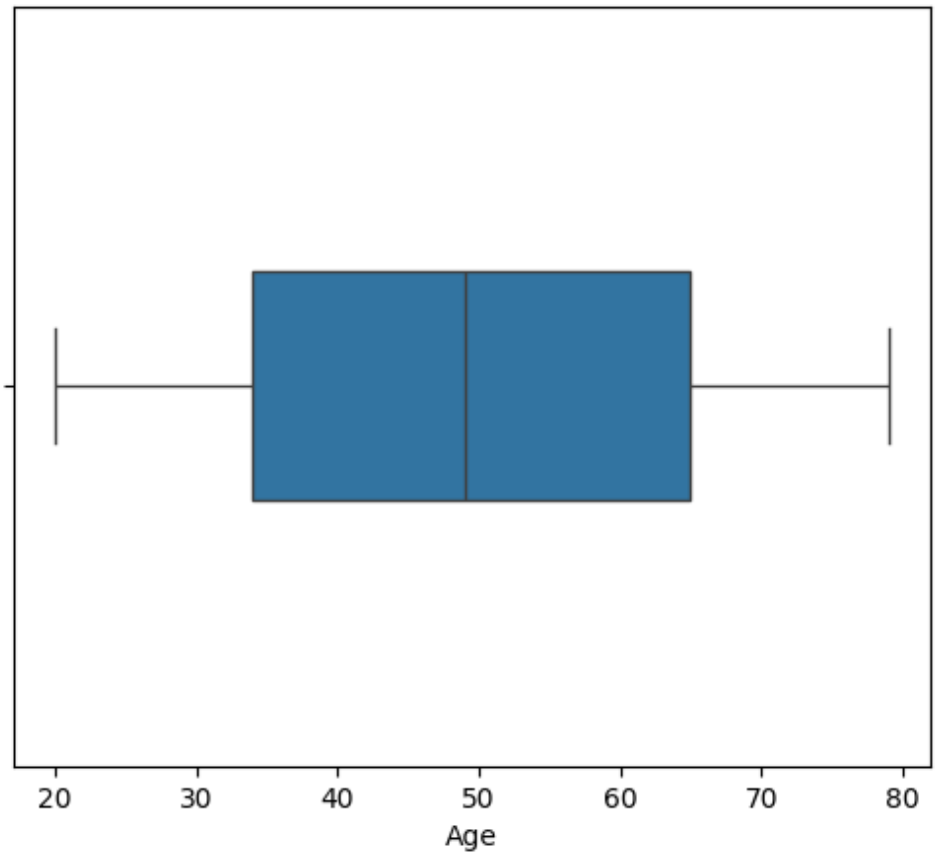

Figure 2: Boxplot of patient age distribution.

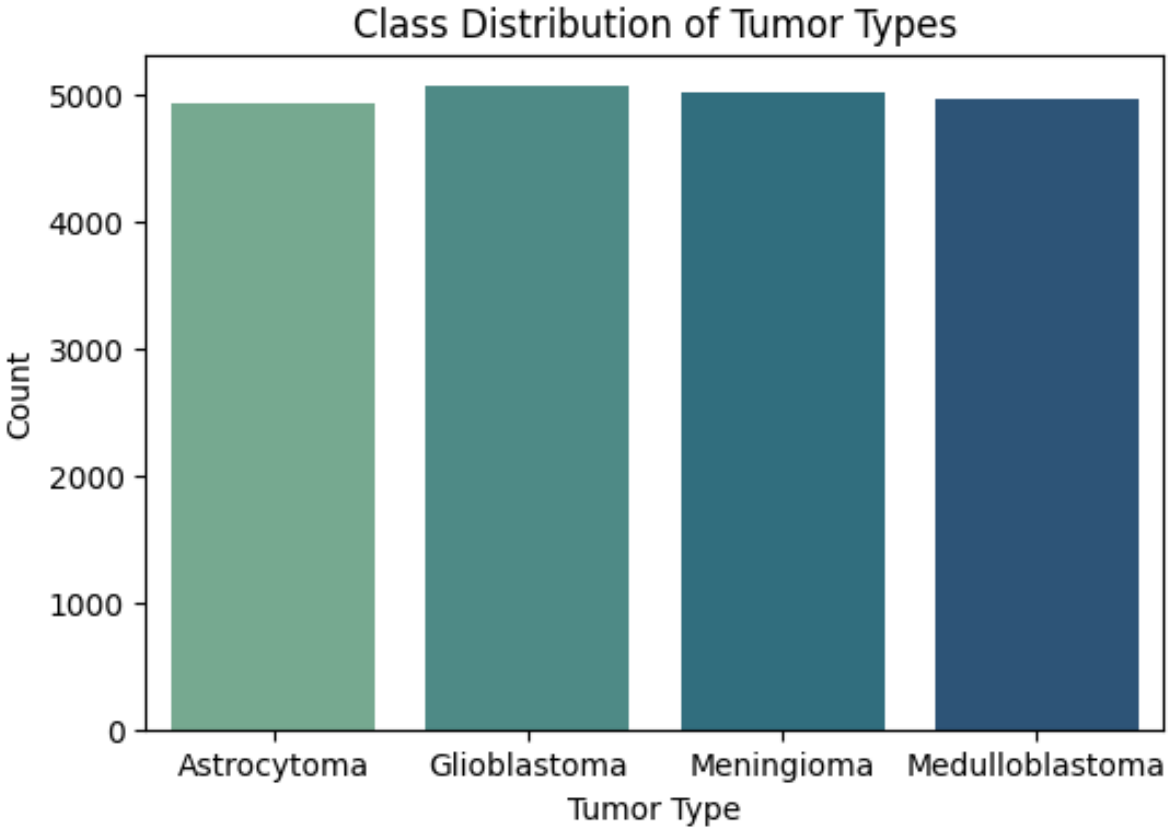

Figure 3: Bar chart showing distribution of histology classes.

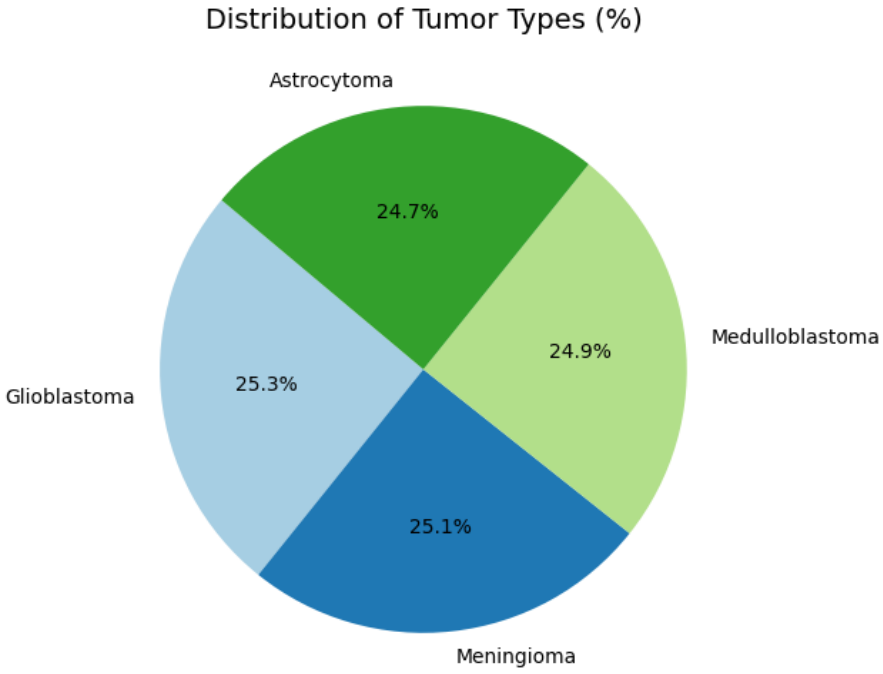

Figure 4: Pie chart showing percentage share of histology classes.

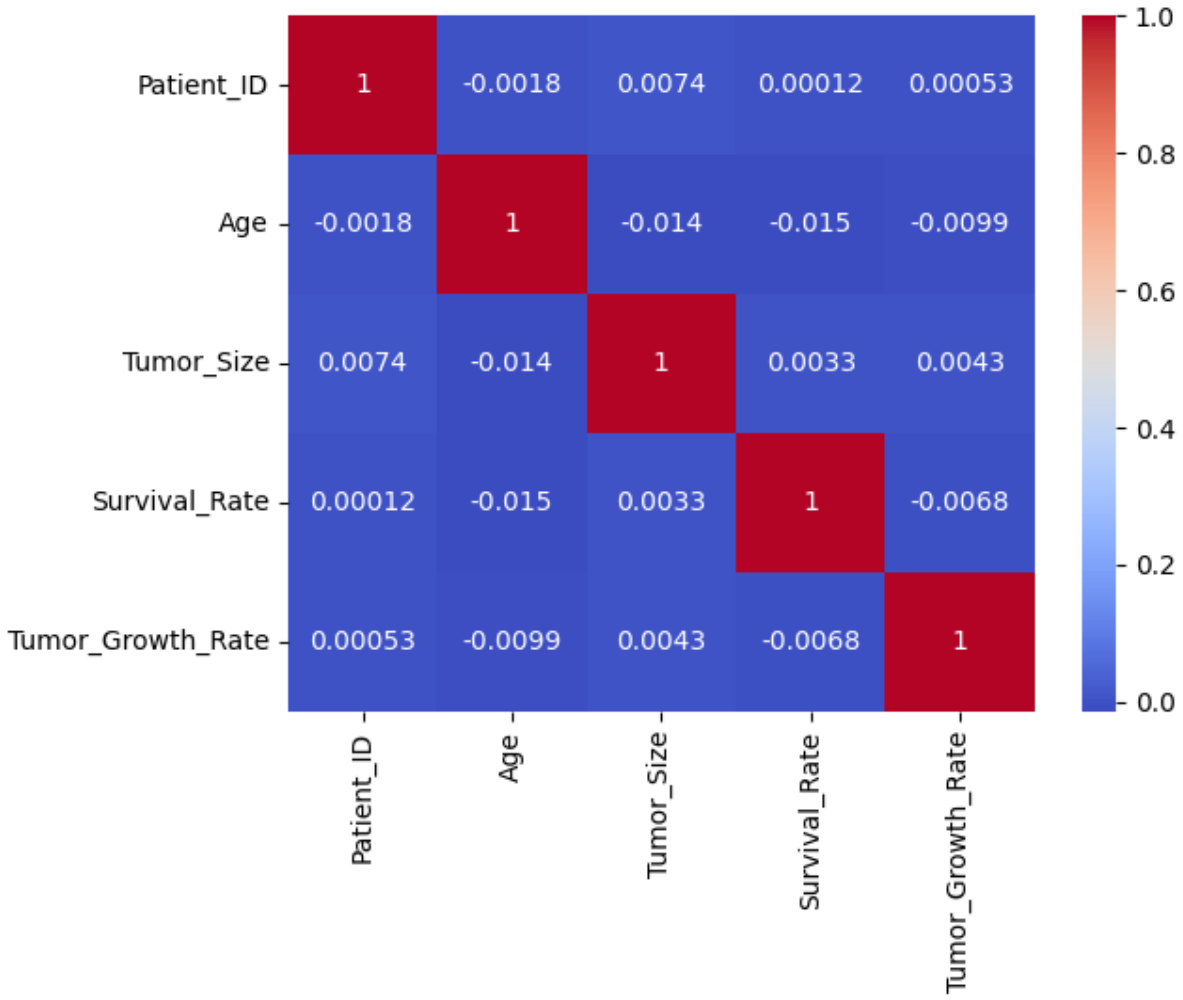

Figure 5: Correlation heatmap of numerical features in the clinical dataset.

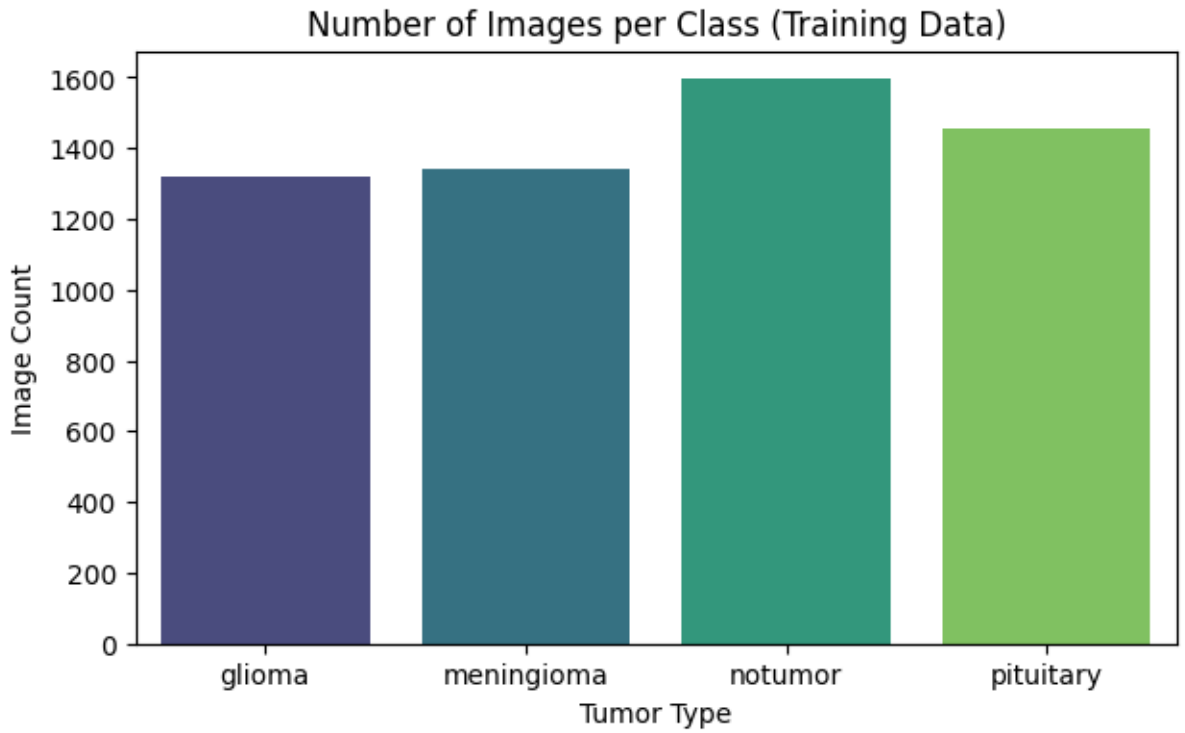

Figure 6: Image count per tumour class in the MRI dataset.

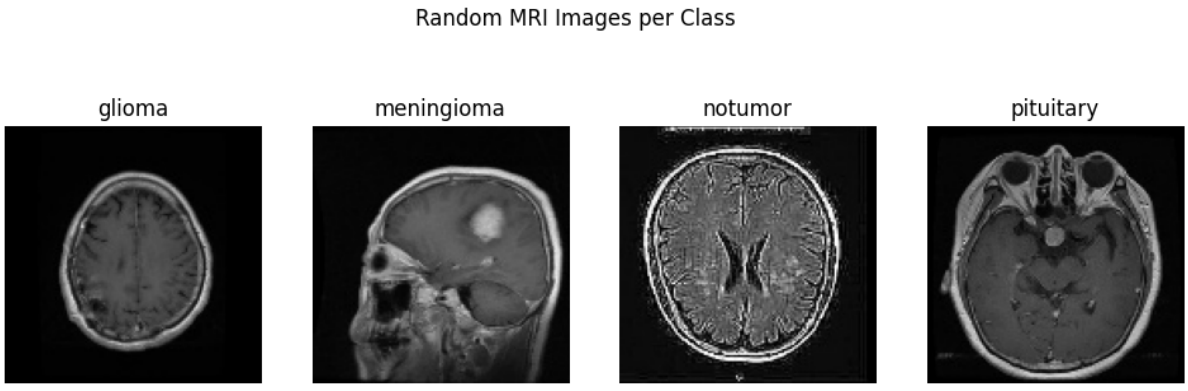

Figure 7: Random MRI samples from each tumour class.

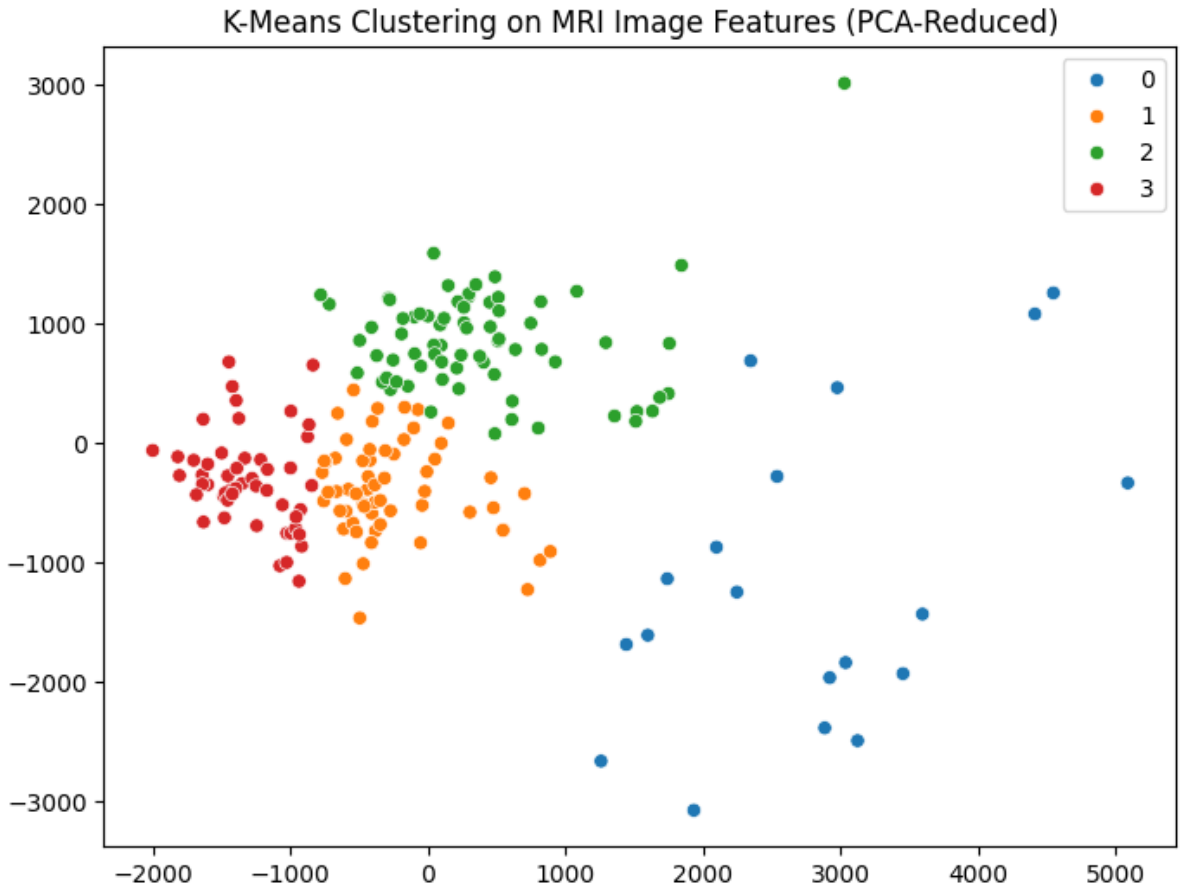

Figure 8: k-means clustering ($k = 4$) on PCA-reduced MRI features.

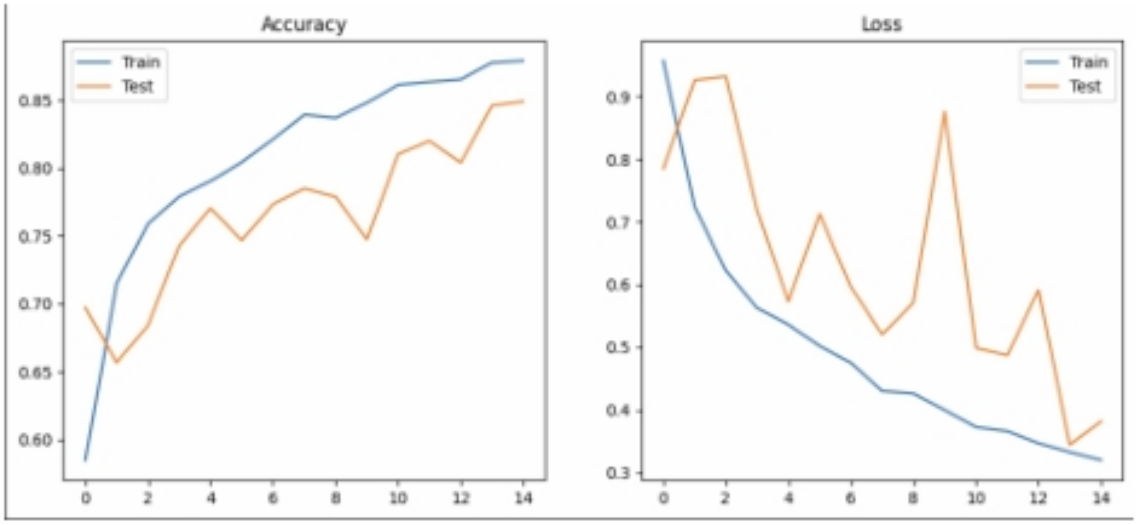

Figure 9: CNN training and validation accuracy/loss curves.

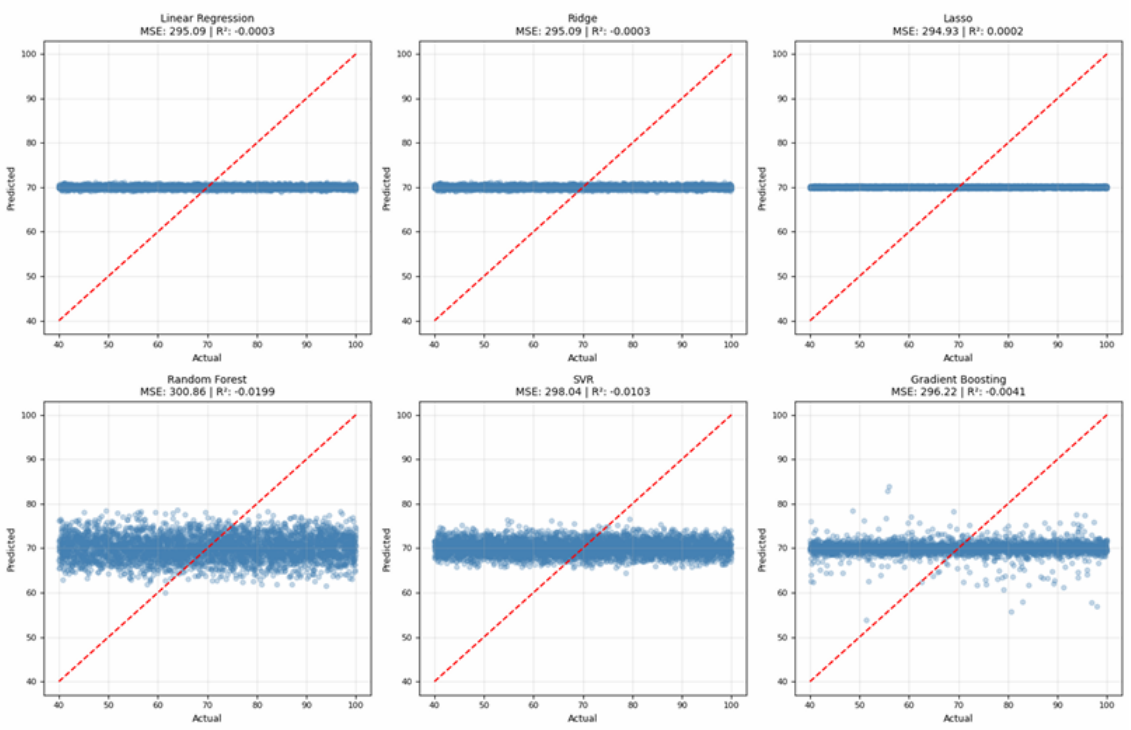

Figure 10: Predicted vs. actual values for regression models.

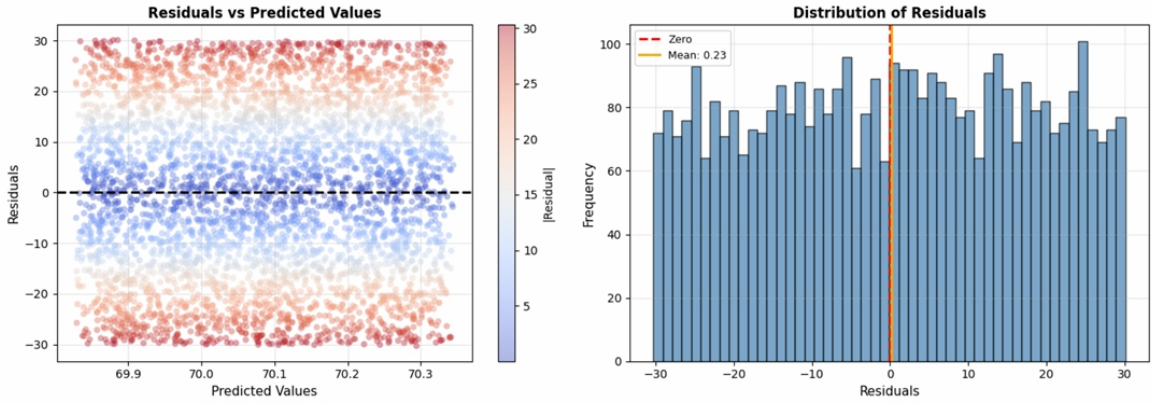

Figure 11: Residual plot showing model errors for regression models.

References

[1] M. M. Badža and M. Č. Barjaktarović. Classification of brain tumors from mri images using a convolutional neural network. *Applied Sciences*, 10(6):1999, 2020.

[2] Miadul. Brain tumor clinical dataset. Kaggle dataset, 2025. `https://www.kaggle.com/datasets/miadul/brain-tumor-dataset`, Accessed 2025.

[3] M. Nickparvar. Brain tumor mri dataset. Kaggle, 2025. Accessed 2025.

[4] I. Rethemiotaki et al. Brain tumour detection from magnetic resonance imaging images using convolutional neural networks. *Scientific Reports*, 14:1850, 2024.