ISSN 2959-6157

Comparison of Machine Learning Models for Brain Disease Classification Based on CT and MRI Scans

Ying Ding

School of Computer Science, the University of Birmingham, Birmingham, Britain
YXD352@student.bham.ac.uk

Abstract:

The complexity and severity of brain diseases have led to increased focus on their diagnosis and treatment. Due to the inherent drawbacks of manual medical diagnosis, such as error-prone and costly, and the recent widespread use of Artificial Intelligence (AI) in medicine, it is a worthy topic to explore machine learning in diagnosing brain diseases utilizing Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). In this paper, three brain diseases using CT or MRI datasets from Kaggle were merged and processed, and then three models of decision tree (DT), random forest (RF) and K-Nearest Neighbors (KNN) were used to make classification prediction. Furthermore, their performance was compared through classification reports and confusion matrix to analysis results. The results showed that DT performed worse than the other two models in this task, with an accuracy of 0.91, whereas RF and KNN performed similar overall, each achieving an overall accuracy of 0.96. Notably, RF exhibited less confusion, especially between some similar categories, which indicates the ability of handle complex data more effectively. Additionally, the strengths and weaknesses of the three models are discussed in the paper. These experimental results show that machine learning model, particularly RF and KNN, is a useful tool for diagnosing brain diseases with high accuracy, which could substantially assist clinical practices by providing reliable and efficient diagnostic support.

Keywords: Brain diseases; CT scan; MRI imaging; machine learning.

1. Introduction

The brain serves as the control center of the human body, which regulates important functions of many organs, control body movement, interpret senses and originate cognitive function. As a part of the nervous system, the brain communicates with nerves to process the signals and to respond, ensuring common daily activity. There are a wide range of disorders and diseases affect the brain, which might influence behavior, personality or even life of a human, such as tumors, mental illness, epilepsy etc. [1]. According to brain research reported by the European Commission, approximately 165 million Europeans have trouble with brain disorders and 75% of people will be likely to experience a neurological or mental disorder during their lifetime [2]. It is supported by Lancet Neurology in 2021 that more than 1 in 3 individuals are affected by neurological conditions, making it a leading cause of illness and disability globally [3]. Therefore, it is increasingly necessary to diagnose and prevent brain diseases timely. In the last decades, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and other brain imaging scans are widely employed to detect and classify different brain diseases and disorders [4]. However, the traditional methods of diagnosis are mainly carried out by doctors and manual diagnosis, which can be time-consuming, costly and misdiagnosed. Hence, with the development of Artificial Intelligence (AI), it is considered gradually to utilize the emerging computerized technology, which has strong prediction ability, to assist clinicians in diagnosing brain diseases.

In recent years, AI has made great progress and significant breakthroughs with the application of various algorithms, such as random forests, logistic regression and K-nearest Neighbors (KNN). These algorithms have been extensively applied across various fields, including finance, biochemistry, and particularly healthcare. Many studies have used Machine Learning (ML) approaches in the sector of disease diagnosis. For instance, Dixit and Kala developed a 1D Convolutional Neural Network (CNN) model for early detection of heart disease based on an affordable Electrocardiogram (ECG) sensor, which demonstrated a 93% accuracy in identifying heart disease [5]. Ahammed et al. offered a skin disease classification model achieving up to 97% accuracy on the Human Against Machine with 10000 training images (HAM10000) dataset and 95% on the International Skin Imaging Collaboration (ISIC) 2019 dataset [6]. Similarly, ML has played a crucial role in the detection and forecasting of brain diseases for several years. Woźniak et al., for example, introduced a novel Correlation Learning Mechanism (CLM) that enhanced CNN performance in evaluating CT brain scans, achieving approximately 96% accuracy and 95% precision and recall by improving filter selection and accelerating neural network training [7]. Furthermore, Parallel Deep Convolutional Neural Network (PDCNN) was proposed by Rahman and Islam, which enhanced MRI brain tumor classification by extracting global and local features, achieving up to 98.12% accuracy and outperforming existing methods [8]. However, most of these studies only focused on analyzing either MRI or CT images individually, without considering models that integrate both types of images in the past. Additionally, the previous models have limitations in generalizability and practicality. This paper aims to develop a multifunctional model with the superior generalization ability that can perform predictive classification on both MRI and CT in brain diseases.

In order to address the issues, this paper combined two datasets about brain CT and one MRI datasets from Kaggle, including aneurysm, tumor, cancer and hemorrhage. Then, decision tree, random forest, and KNN algorithms were utilized for training and testing to compare accuracy for selecting the most effective model for clinical brain disease detection and prediction. The final experimental results demonstrated the accuracy of the ML models, providing valuable insights for future practical applications in identifying brain diseases using both CT and MRI scans.

2. Method

2.1 Dataset Preparation

In this paper, two different datasets of brain CT images and one MRI dataset are used, which are all from Kaggle [9-11]. Specifically, the first dataset identified as dataset-I about brain CT hemorrhage has been obtained in the platform [9]. It includes 512X512 pixel pictures of 27 directories normal and 18 directories hemorrhagic CT scans collected from the Near East Hospital, Cyprus. Another CT dataset includes 259 brain scans with the same pixel total with cancer, tumor, and aneurysm, each of which shows a detailed image of the brain of a patient captured through CT; in this study, it is referred to as dataset-II [10]. Finally, the MRI dataset classifies 3264 various pixels MRI images into four categories: meningioma tumor, glioma tumor, pituitary tumor and no tumor [11]. This data is referred to as dataset-III in this research.

In terms of data preprocessing, given the existence of duplicate features, no tumor images from dataset-III were merged with the normal images of dataset-I. Additionally, the tumor classification in dataset-II was omitted, as dataset-III provided a more detailed categorization of tumors into three specific subtypes. Subsequently, these three datasets were combined into one dataset with seven categories, such as aneurysm, cancer, glioma tumor etc., which was cleaned to avoid image corruption or duplication. After labeling process, all of images were converted to grayscale and resized to a uniform size of 128X128 pixels reducing complexity of data. Then, in order to focus on the relevant features and patterns noise, a reduction technique was applied to enhance image quality minimizing artifacts and random variations. The next step was normalizing the cleaner pixel values to [0, 1] and encoding labels into numerical format. Additionally, the dataset was split 80-20 for training and testing. Lastly, to expand the quantity and variety of the training dataset, data augmentation was employed by random transformations such as rotation, shearing, zooming and flipping. These operations also improved model generalization and Fig. 1 displays the performance of the original images, images after noise reduction, and after data augmentation.

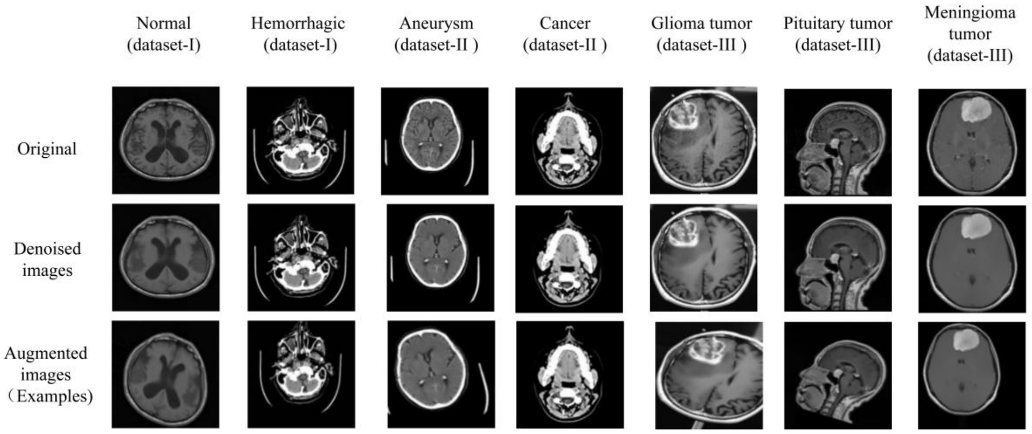

Fig. 1 Categories of brain scans from three datasets, including original images, images after noise reduction, and images after data augmentation (Photo/Picture credit: Original)

2.2 Machine Learning Models-based Prediction

This study used decision tree, random Forest and KNN to make categorical predictions of brain CT and MRI. Several key libraries and modules for data processing and machine learning tasks were used, such as sklearn, pandas, numpy, seaborn, PIL and cv2. At last, the performance of the models was evaluated with classification reports and confusion matrices, which facilitated comparison of the effectiveness.

2.2.1 Decision tree

Decision Tree (DT) is widely used in ML for making decisions or predictions due to interpretability. It shows a tree structure in which internal nodes represent decisions based on specific features, branches indicate the results of those decisions, and leaf nodes show the final classification or prediction.

The main characteristic of DT is the recursive process of dividing a dataset determined by the values of input features, in order to create a tree structure that leads to the most accurate predictions or classifications. By evaluating different features and values, it aims to segment the data in a way that maximizes the homogeneity of the resulting subsets, ultimately facilitating precise and reasonable decision-making. Furthermore, the use of greedy algorithm in Decision Trees involves making locally optimal decisions at each node based on the best available split at that point, rather than evaluating the overall structure of the tree. However, this approach can lead to overfitting, where the model excels with training data but may underperform with new data. To address this, pruning techniques are employed to remove branches that contribute minimally to accuracy, simplifying the tree and enhancing robustness.

In this case, except for specifying the random state parameter to 42, the classifier uses these default settings for all other parameters. By setting a fixed random seed, the randomness involved, such as data shuffling and bootstrapping, always produces the same result, which makes it easier to debug and replicate the model.

2.2.2 Random forest

Random Forest (RF) is an ensemble learning approach by aggregating the outputs of multiple decision trees, which enhances prediction accuracy and control overfitting. This technique significantly improves robustness and generalization when compared to individual decision trees.

The fundamental principle of RF is to build a large number of decision trees, with each tree being trained on a different bootstrap sample of the data through the bagging method. Bagging forms several data subsets through random sampling with replacement and trains each tree on a different subset. Additionally, each tree is trained using a random subset of features to improve further variability. The final prediction is obtained by aggregating the outputs from all the trees, either through majority voting for classification or averaging for regression. This bagging technique is key to enhancing accuracy and stability because of reducing variance and increasing the diversity of the trees.

In this case, the Random Forest Classifier was configured with the following parameters: 100 trees were used to balance accuracy and computational expense. Trees were allowed to grow without a maximum depth, expanding until all leaves had fewer than the minimum samples needed to split. The minimum number of samples needed to split an internal node was set to 2, while the minimum required for a leaf node was set to 1. Lastly, class imbalances were handled by setting the class weight to ‘balanced’ and a random seed of 42 was used to ensure reproducibility.

2.2.3 K-nearest neighbors

The KNN algorithm functions by locating the ‘k’ closest data points to a target instance within a feature space, using a chosen distance metric, such as Euclidean distance for continuous features. For classification tasks, KNN allocates the target instance to the class that appears most frequently from its ‘k’ nearest neighbors. This plurality voting mechanism makes KNN particularly effective for classifying data points in well-separated datasets. Besides, since KNN is a lazy learner, there is no explicit training phase. The computational cost is shifted to the prediction phase, which involves calculating distances and finding nearest neighbors.

In this case, the primary hyperparameter for KNN classifier is ‘k’, which was optimized using grid search. The search range for ‘k’ was set from 1 to 20, and each value was evaluated using 5-fold cross-validation. Finally, the optimal value of ‘k’ was found to be 1, with a best cross-validation score.

3. Results and Discussion

3.1 The Performance of Different Models

Table 1. Classification reporting performance of different models in the detection of brain diseases by CT and MRI

Methods	Precision	Recall	F1-score	Accuracy
DT	0.91	0.90	0.91	0.90
RF (100 trees)	0.96	0.96	0.96	0.96
KNN (k=1)	0.96	0.96	0.96	0.96

By analyzing the classification report of the three models, Table 1 presents the precision, recall, f1-score and overall accuracy of each model in the detection of brain diseases by CT and MRI. Specifically, RF and KNN outperform DT significantly, as indicated by their higher accuracy, precision, recall, and f1-scores, all reaching 0.96. In contrast, DT exhibits an accuracy of 0.90, with slightly lower precision and recall (0.91 and 0.90, respectively). These differences suggest that DT, while able to capture broader trends, struggles to provide the same level of classification precision as RF and KNN, particularly when dealing with complex dataset. Additionally, Fig. 2, Fig. 3 and Fig. 4 summarize the performance of three models by comparing predicted labels with the actual labels respectively.

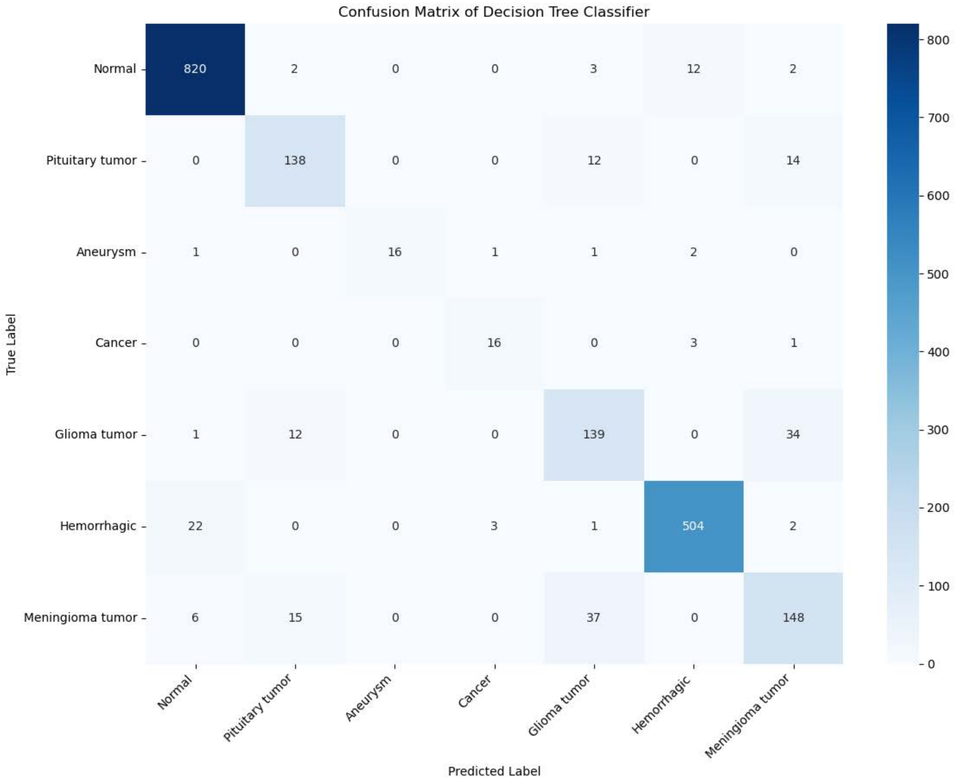

Fig. 2 Confusion Matrix of DT on identification of brain diseases by CT and MRI (Photo/Picture credit: Original)

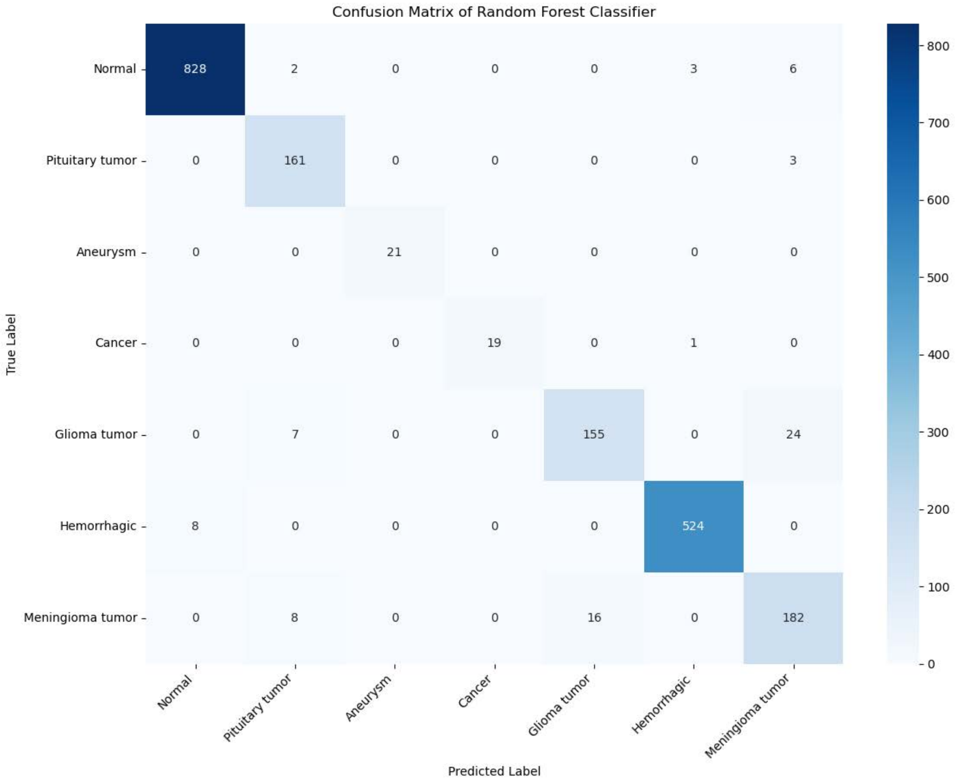

Fig. 3 Confusion Matrix of RF on identification of brain diseases by CT and MRI (Photo/ Picture credit: Original)

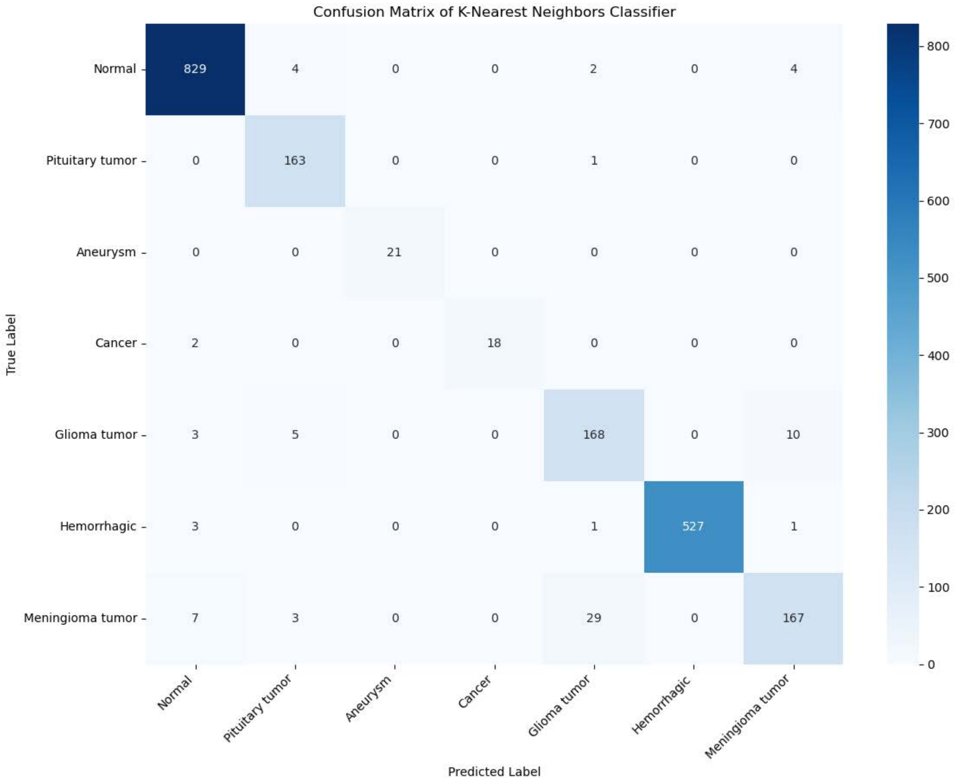

Fig. 4 Confusion Matrix of KNN on identification of brain diseases by CT and MRI (Photo/ Picture credit: Original)

From the confusion matrices in the figures, which provides a detailed view of how well each model performs in classifying brain diseases using CT and MRI scans. The DT model shows notable misclassifications, particularly between meningioma tumor and glioma tumor. A significant number of meningioma tumor instances are misclassified as glioma tumor, indicating that DT may have difficulty distinguishing between these categories. This suggests that decision boundaries of DT are less precise, leading to errors when features overlap or are closely related between classes. On the other hand, RF demonstrates a more robust classification across all categories, with fewer errors, including the previously problematic meningioma tumor and glioma tumor. This highlights the benefit of the ensemble approach employed by RF, which aggregates the predictions of multiple decision trees to enhance its generalization ability and minimize overfitting, thus performing better in distinguishing between complex, overlapping categories. Lastly, while achieving results nearly identical to RF in terms of overall metrics, KNN also faces some issues with misclassification, particularly in the meningioma tumor and glioma tumor as well. This may be due to k=1, where the model relies heavily on the nearest neighbor, which is at risk of being affected by local noise or slight variations.

The differences in performance among these models can be explained by inherent characteristics. The misclassification of meningioma tumor into glioma tumor, and vice versa, suggests that the proximity of these samples in the feature space could be affecting ability of models accurately distinguish between them. Superior performance of RF stems from ensemble nature, where multiple decision trees work together, which reduces overfitting and improves generalization, especially in complex, noisy datasets. Additionally, KNN relies heavily on local data structure, which makes it effective when class distributions are distinct and more vulnerable to errors if classes overlap, or data variations are subtle. However, the Decision Tree is more prone to overfitting and struggles with precise boundary distinctions. It leads to more misclassifications in overlapping classes. Thus, robustness in handling diverse patterns and minimizing errors makes it the most reliable model in this study.

In summary, for the detection of brain diseases using CT and MRI, RF provides the most accurate and reliable classification due to high accuracy and minimal confusion across classes. KNN performs similarly well, but its reliance on local data structures causes some limitations in specific cases. Though it is simple to interpret and fast to train, DT shows weaknesses in handling more complex class distinctions. To Improve model performance and exploring ensemble methods that could combine the strengths of these algorithms for even better accuracy and generalization.

4. Conclusion

This paper compares the performance of three machine learning models including DT, RF, and KNN in classifying brain diseases from CT and MRI scans, which aims to assist in clinical diagnosis. After merging and cleaning the datasets, the models were applied, and their performance was evaluated using classification reports and confusion matrices. RF demonstrated the best overall performance, followed by KNN, both outperforming DT. However, the study has some limitations. Firstly, the datasets may not fully capture the diversity of brain diseases across different populations, limiting the generalizability of the findings. Secondly, model hyperparameters, such as the number of trees in RF and the k value in KNN, were might not fully optimized, which possibly affects the results. Future work could include additional data sources, such as clinical records or genetic information, to enhance predictive accuracy. Additionally, advanced models like CNNs could be explored to improve classification performance further. These improvements may lead to more effective tools for diagnosing brain diseases in clinical practice.

References

[1] Brain Diseases: Definition & Types. Cleveland Clinic. 2022 Apr 26. Available from: https://my.clevelandclinic.org/health/diseases/22934-brain-diseases.

[2] Brain research. European Commission. Available from: https://research-and-innovation.ec.europa.eu/research-area/health/brain-research_en.

[3] WHO. Over 1 in 3 people affected by neurological conditions, the leading cause of illness and disability worldwide. 2024 Mar 14. Available from: https://www.who.int/news/item/14-03-2024-over-1-in-3-people-affected-by-neurological-conditions--the-leading-cause-of-illness-and-disability-worldwide.

[4] Haq EU, Huang J, Kang L, et al. Image-based state-of-the-art techniques for the identification and classification of brain diseases: a review. Med Biol Eng Comput. 2020;58(11):2603-2620.

[5] Dixit S, Kala R. Early detection of heart diseases using a low-cost compact ECG sensor. Multimed Tools Appl. 2021;80(21-23):32615-32637.

[6] Ahammed M, Mamun Md A, Uddin MS. A machine learning approach for skin disease detection and classification using image segmentation. Healthcare Analytics. 2022;2:100122.

[7] Woźniak M, Siłka J, Wieczorek M. Deep neural network correlation learning mechanism for CT brain tumor detection. Neural Comput Appl. 2021.

[8] Rahman T, Islam MS. MRI brain tumor detection and classification using parallel deep convolutional neural networks. Measurement: Sensors. 2023;100694.
[9] Abdulkader Helwan. Brain CT Hemorrhage Dataset. Kaggle.com. 2018. Available from: https://www.kaggle.com/datasets/abdulkader90/brain-ct-hemorrhage-dataset/data.
[10] Data T. Computed Tomography (CT) of the Brain. Kaggle.com. 2023. Available from: https://www.kaggle.com/datasets/trainingdatapro/computed-tomography-ct-of-the-brain/data.
[11] Brain Tumor Classification (MRI). Kaggle.com. Available from: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.