Machine Learning – Final Project

Brain Tumor Classification

Noa Amichai 206996381
Avi Ostroff 327341590

GitHub Repository: Brain Tumor Classification GitHub Repository

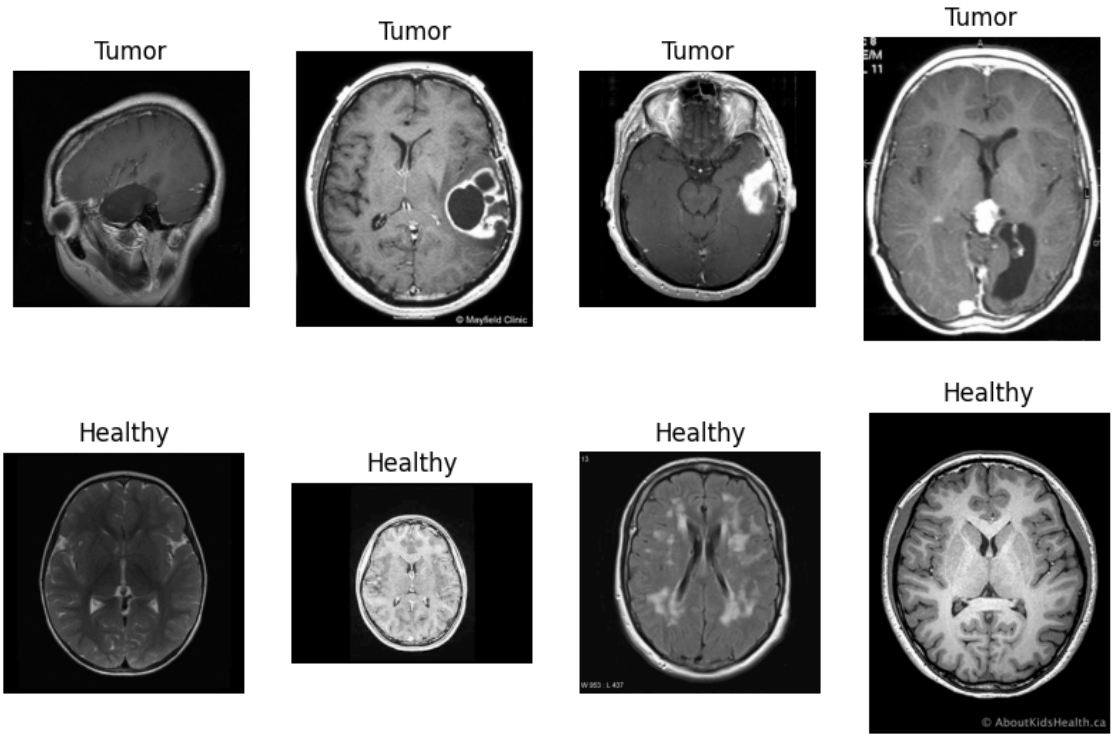
---

Dataset

Our chosen dataset "Brain Tumor" consists of MRI images divided into two categories: "tumor" and "healthy." The images are sourced from the 'Brain Tumor Data Set/Brain Tumor' and 'Brain Tumor Data Set/Healthy' directories.

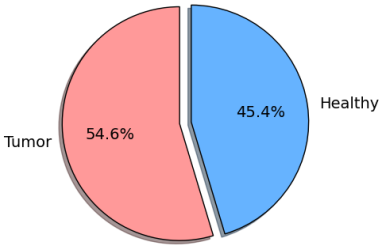
- **Tumor Images**: Images depicting various types of brain tumors.
- **Healthy Images**: Images depicting healthy brains without tumors.

Sample images from the dataset:



Data Distribution

The data is equally distributed between the "tumor" and "healthy" categories, ensuring a balanced dataset for training and evaluation.



Research Questions

The primary questions addressed in this project are:

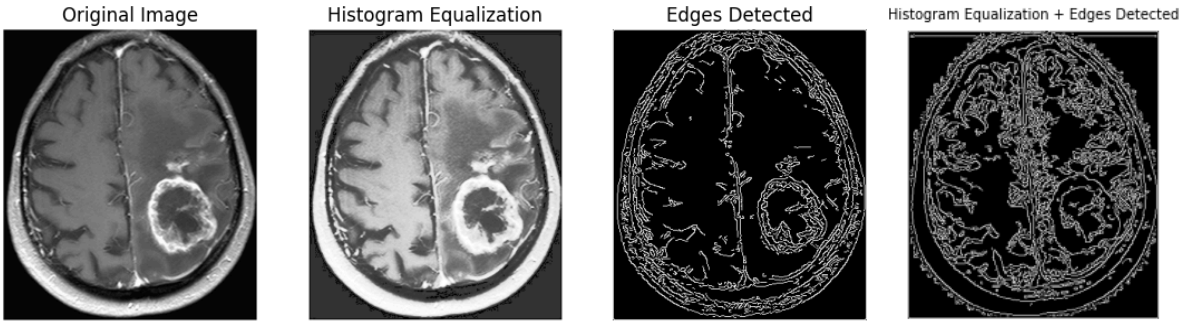
1. Can we accurately classify brain tumors using image preprocessing and machine learning techniques?
2. Which preprocessing and feature extraction techniques yield the best classification performance?
3. What is the best way (for our data) to convert an image to a vector?
4. Will the color palette (RGB or grayscale) significantly change the accuracy?

Preprocessing

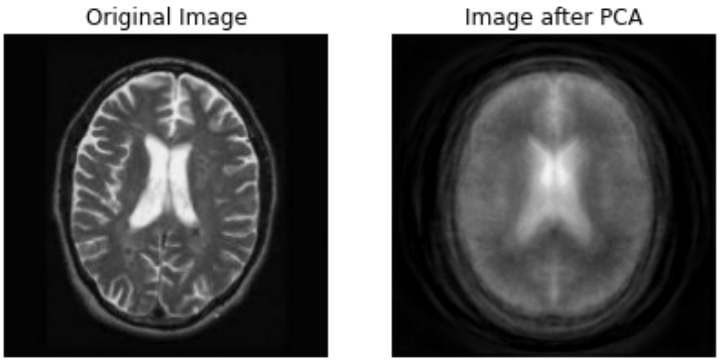
We experimented with several preprocessing techniques and evaluated the model's performance with each technique.

1. **Histogram Equalization**: Enhancing image contrast.
2. **Canny Edge Detection**: Detecting edges to highlight significant features.
3. **Combining Histogram Equalization followed by Canny Edge Detection**

Sample preprocessing results:



4. **Principal Component Analysis (PCA)**: Reducing feature dimensionality.



5. **Feature Extraction Using Pre-trained VGG16 Model**
6. **Feature Extraction Using Pre-trained ResNet50 Model**

Baseline Model

Initially, we trained a Dummy Classifier to establish a baseline accuracy. The Dummy Classifier predicted the most frequent class in the training set. Our subsequent models should outperform this baseline.

The accuracy of the dummy model is 0.546.

Model Training

We trained several machine learning models using the preprocessed images and extracted features:

1. K-Nearest Neighbors (KNN)
2. Logistic Regression
3. Support Vector Machine (SVM)
4. Random Forest
5. XGBoost
6. Convolutional Neural Network (CNN)

Thought Process During Project Development

Our approach to the project began with a straightforward and familiar method: minimal preprocessing of the images before feeding them into the model. This simple method (using both RGB and grayscale images) yielded decent results, with an accuracy of at least 91% on the validation set. Seeking higher accuracy, we decided to experiment with PCA. However, adding PCA without additional processing did not significantly improve the results.

The next step was to use Canny edge detection, thinking that identifying the edges could help in better recognizing the tumors and thus improving our accuracy. The SVM model showed improvement, achieving an accuracy of 96%. On the other hand, the KNN model's performance dropped significantly, achieving only 60% accuracy, which is barely better than our dummy model. We hypothesize that the loss of substantial information due to edge detection made it challenging for KNN to classify the images accurately.

Subsequently, we tried histogram equalization, which showed a slight improvement. By combining histogram equalization with PCA for the XGBoost model, we achieved further accuracy gains.

We also experimented with combining histogram equalization and Canny edge detection, expecting that the combination might help in better distinguishing the tumor regions. However, this combination did not improve the model's performance. We believe this may be due to the fact that combining these two techniques could remove important features from the images that are necessary for the model to make accurate predictions and may also introduce irrelevant noise.

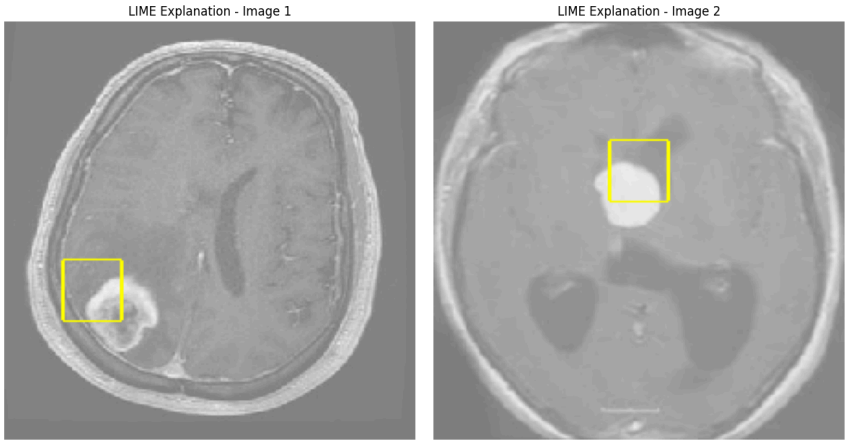
Our final step involved feature extraction using two pre-trained models, ResNet50 and VGG16. By extracting significant features from the images, we were able to substantially improve our model's accuracy, ultimately achieving an accuracy of 98.4% on our test set.

LIME Explanations

We used LIME on our CNN to:

1. Validate the regions of the MRI images that the model considered important for its predictions.
2. Ensure that the model was not making decisions based on irrelevant features or noise.

By using LIME, we could visualize which parts of the brain MRI were most influential in classifying an image as having a tumor or not.



Model Performance Summary

Image Processing Technique	Model	Validation Accuracy	Test Accuracy
RGB + Flatten	KNN	0.942	-
RGB + Flatten	Random Forest	0.963	-
RGB + Flatten	SVM	0.917	-
RGB + Flatten + PCA	Random Forest	0.962	-
Gray Scale + Flatten	Random Forest	0.947	-
Gray Scale + Flatten	XGBoost	0.957	-
Gray Scale + Flatten	CNN	0.953	0.960
Canny Edge Detection	KNN	0.600	-
Canny Edge Detection	SVM	0.960	-
Histogram Equalization	Random Forest	0.963	-
Histogram Equalization	XGBoost	0.966	-
Histogram Equalization + PCA	XGBoost	0.972	-
Histogram Equalization + Canny Edge Detection	SVM	0.953	-
Histogram Equalization + Canny Edge Detection	Random Forest	0.903	-
Histogram Equalization + Canny Edge Detection	XGBoost	0.932	-
VGG16 for Feature Extraction	Random Forest	0.972	0.967
VGG16 for Feature Extraction	XGBoost	0.989	0.971
VGG16 for Feature Extraction	**SVM**	**0.989**	**0.984**
VGG16 for Feature Extraction	KNN	0.972	0.965
VGG16 for Feature Extraction	Logistic Regression	0.989	0.979
ResNet50 for Feature Extraction	SVM	0.990	0.982
ResNet50 for Feature Extraction	KNN	0.989	0.978
ResNet50 for Feature Extraction	Logistic Regression	0.989	0.979

Results

The performance of different models was evaluated based on accuracy score. The best-performing model was the SVM using VGG16 for feature extraction, achieving an accuracy of 98.4%.

Conclusion

The Brain Tumor Classification project demonstrated that advanced image preprocessing and feature extraction techniques, coupled with robust machine learning models, can lead to accurate classification of brain tumors. This project highlights the potential of machine learning in medical diagnostics and provides a solid foundation for further research and improvement in this critical area.