Brain tumor detection through MRI image recognition

Georgios Angelos Mathioudakis
Computer Science Department - University of Crete, Voutes Campus
700 13 Heraklion Crete, Greece
med7p1120065@med.uoc.gr

Abstract

This research investigates the predictive ability of various machine learning algorithms on identifying brain tumors based on data extracted from brain MRIs. This research aims to demonstrate the optimal algorithm for brain tumor detection and present the most important features that could indicate the existence of this disease.

1 Introduction

Nowadays cancer is one of the diseases with highest mortality rate among the total population, despite the numerous efforts conducted through the years to treat it on an early stage. Especially brain tumor is one of the most lethal forms of cancer among people of all ages and characteristics. A brain tumor is a mass or growth of abnormal cells in your brain that can cause severe symptoms and very often death, unless it is diagnosed in sort time after the appearance.So Artificial Intelligence methods are being utilized to recognize patterns and hidden relationships among the image recognition data, that could lead to a primarily detection of such type of cancer and potentially to a successful treatment.

2 Research Objective

In order to encounter the disease a common practice is the usage of MRI, since it is considered the best practice to detect brain tumors by creating pictures of the soft tissue body parts. In this project the application of machine learning methods will be explored in terms of analyzing tabular data extracted from brain tumor images recognition to detect a brain tumor on an early stage. In addition, the significance of the independent variables will be discussed towards their contribution on the prediction. The dataset is composed by 3762 observation and 15 features which are mainly statistics regarding the brain MRI characteristics between the given images. The individual characteristics can be seen listed bellow:

1. Image : Image identifier
2. Class : Target variable of 1 and 0 of whether having a tumor or not
3. Mean : Gives the contribution of individual pixel intensity for the entire image
4. Variance : Used to find how each pixel varies from the neighbouring pixel
5. Standard Deviation : measures the deviation of measured Values or the data from its mean.
6. Skewness : measures of symmetry, or more precisely, the lack of symmetry.
7. Kurtosis : describes the peakedness of e.g. a frequency distribution
8. Entropy

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

9. Contrast : the difference in luminance or colour across the image
10. Energy : It's the rate of change in the color/brightness/magnitude of the pixels over local areas.
11. ASM (Angular second moment) : is a measure of textural Uniformity of an image
12. Homogeneity : homogeneity expresses how similar certain elements (pixels) of the image are.
13. Dissimilarity : is a numerical measure of how different two data objects are.
14. Correlation : Correlation is the process of moving a filter mask often referred to as kernel over the image and computing the sum of products at each location(CNN alike)
15. Coarseness : Describes the roughness/harshness of a texture

The analysis is a classification problem, in which the target variable will be predicted by using supervised machine learning algorithms. The process involves the model development, algorithms training on a train set and model implementation on unseen data to evaluate its performance and interpret its results. Three machine learning algorithms will be employed and evaluated in this project, while a tree-based model will be constructed with built in variable selection feature in order to distinguish which predictors are the most important. Random Forest is a supervised machine learning algorithm that can be used in either regression and classification and it is considered relatively powerful and yet not much computational cost required. It operates based on the ensemble approach which means combining multiple models to optimize its results. The algorithm generates several decision trees by using the bagging revamping techniques. Final output is the Majority Average of those classifiers decision trees.

Support Vector Machines (SVMs) is an advanced supervised algorithm that can be also employed for classification and regression problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. According to the SVM algorithm it finds the points closest to the line from both the classes. These points are called support vectors. The distance between the line and the support vectors is being computed. This distance is called the margin. The goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane. Thus, SVM makes a decision boundary in such a way that the separation between the two classes is as wide as possible.

K-Nearest Neighbors (K-NN) is a simple easy to interpret machine learning algorithm that is used for supervise and unsupervised problems. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (datapoint) to predict the class or continuous value for the new datapoint. It works based on three approaches to discriminate among the data.

1. Instance-based learning: Here we do not learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data.
2. Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance.
3. Non -Parametric: In KNN, there is no predefined form of the mapping function.

For this machine learning application, we will use Python programming language to develop the models and perform the predictions. It is considered one of the best tools in the area of data mining, since it offers great libraries and frameworks for AI and Machine Learning. Due to its simple syntax, the development of applications with Python is fast when compared to many programming languages. This experiment will be carried out using the Jupiter notebook provided by anaconda and libraries such as NumPy, scikit-learn and SVM will be used.

The objective of this project is to detect brain tumors using numerical data extracted from brain MRI images using machine learning algorithms and various statistical modeling techniques. Due to the nature of the problem, data is very limited and challenging to collect. In this project we will compare the predictive power of the proposed algorithms and utilize the feature selection module of the random forest algorithm to establish which variable are the most important for the prediction results.

3 Experiment

3.1 Research questions

1. The aim of this project is to investigate which algorithm performs better on discriminating cancer on data observations that collected from brain image recognition (SVM, Random Forest, K-NN).
2. Also, we will estimate which one of the 15 statistical variables are the most significant predictors regarding the brain tumor detection.

3.2 Process Description

In general the data was extracted by using image recognition on brain MRIs so the dataset does not suffer from missing values or requires further cleaning, since the variables are statistical results of this procedure. The first phase of our analysis was an exploratory data analysis to understand the data and what methods to use in order to produce an accurate prediction.

Bellow can bee seen the most significant findings regarding the data distributions.

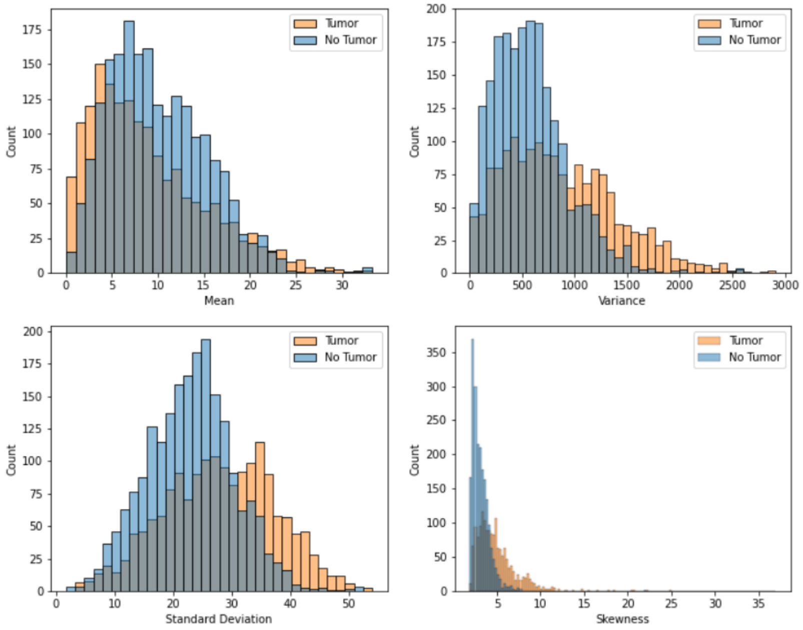

Figure 1: Density distributions for Mean,Variance, Skewness and Standard Deviation variables

As can be seen from comparison between observations with the indication of having a tumor and those without, have relatively normally distributed standard deviation while, the mean and the standard deviation are slightly skewed. On the other hand, the skewness column seems to be quite skewed and containing a significant number of potential outliers. Furthermore, it is clear that values in regard to the No tumor observations tend to have greater variance.

Similar distribution can be identified regarding the ASM and the Energy variables in which observations with tumor identifier seem to highly skewed and probably there are outliers as well. While the

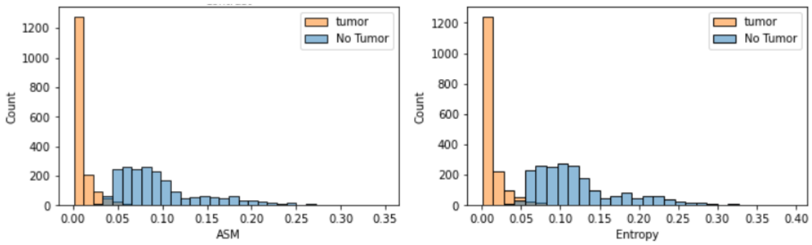

Figure 2: Density distributions for Mean,Variance, Skewness and Standard Deviation variables

no tumor observations are skewed marginally while potential outliers could be identified from this initial analysis. As part of the data preprocessing normalization was being applied on the numerical values of the dataframe since scaling is required in order for the model. Scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower. This is because the model investigates the distance among the features and not the absolute values. In general normalization means transforming the data, namely converting the numerical data such as the distribution centered around 0, with a standard deviation of 1.Furthermore, the target variable a fairly divided to about 55% no tumor to 45% tumor indication which is a beneficial for our objective.

Next feature selection with mutual information was applied on the numerical columns. Feature selection is the process of identifying and selecting a subset of input features that are most relevant to the target variable. After acquiring the optimal features to be used in the model we filter the dataset to remove the not important ones since they could cause confusion in the models.

We utilize stratified cross validation along with hyperparameters tuning in order to estimate the best parameters for the proposed algorithms (Random Forest, SVM and K-NN). Cross validation is an advanced method to ensure that the model is not overfitting on the training data. For the model tuning we choose an exhausted Grid searching method which is the process of scanning the data to configure optimal parameters for a given model.

Then, we split the data to test and train sets for the model implementation. We are using the optimal parameters extracted in previous phase and model evaluation based on classification metrics according to the confusion matrix ratios. As mentioned above the target variable is divided almost equally to the two classes, so as there is no class imbalance problem and the accuracy ratio can be an efficient measure of model evaluation.

4 Results-Discussion

Our findings reveal that Random Forest appears to be the most accurate algorithm since it produces an accuracy score of 0.99, followed by the K-NN with a ratio of 0.86, see Table 1.

Table 1: Performance Metrics results

Model	Accuracy
Random Forest	0.99
SVM	0.57
K-NN	0.86

Those results seem appropriate based on the existing literature as the Random forest tends to outperform other data mining algorithms as far as medium to small size datasets are concerned. In addition we have to mention the that the SVM and K-NN models are by far more computationally expensive that the Random Forest, as a result with conclude that based on the given historical data Random Forest is superior than its counterparts both in predictive power and efficiency.

Moreover we investigated the predictors significance according to the variables selection feature of the Random forest algorithm.

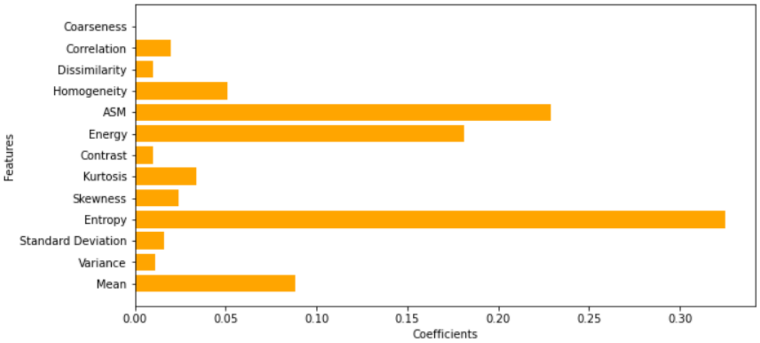

Figure 3: Features importance

Based on Figure 3 it is quite clear that Entropy, ASM and Energy appear to be the most significant features in regard to the brain tumor detection.In addition the Mean seems to be an a strong predictor as well while not the aforementioned ones. However, some of the features with low significance could be omitted in order to make the models less complex and possibly provide a more accurate prediction.The variables with the least importance seem to be Coarseness and Contrast.

Nevertheless, there are some limitations in our project due to the limited data since their collection occurs based on real life patients information and this makes the collection process quite challenging.

References

[1] Keerthana, A., Kavin Kumar, B., Akshaya, K. and Kamalraj, S., 2021. Brain Tumour Detection Using Machine Learning Algorithm. Journal of Physics: Conference Series, 1937(1), p.012008.

[2] Medium. 2022. Machine Learning Basics with the K-Nearest Neighbors Algorithm. [online] Available at: <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761> [Accessed 7 February 2022].

[3] Kaggle.com. 2022. Brain MRI Images for Brain Tumor Detection. [online] Available at:<https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection/activity> [Accessed 7 February 2022].

[4] Analytics Vidhya. 2022. Random Forest | Introduction to Random Forest Algorithm. [online] Available at:<https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/> [Accessed 7 February 2022].

[5] Kumar, T., Rashmi, K., Ramadoss, S., Sandhya, L. and Sangeetha, T., 2017. Brain tumor detection using SVM classifier. 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS),.

[6] Angel Viji, K. and Hevin Rajesh, D., 2020. An Efficient Technique to Segment the Tumor and Abnormality Detection in the Brain MRI Images Using KNN Classifier. Materials Today: Proceedings, 24, pp.1944-1954.

[7] International Journal of Science and Research (IJSR), 2016. Detection of Brain Tumor Using K-Means Clustering. 5(6), pp.420-423.

[8] Płoński, P., 2022. Random Forest Feature Importance Computed in 3 Ways with Python. [online] MLJAR. Available at: <https://mljar.com/blog/feature-importance-in-random-

forest/: :text=Random%20Forest%20Built%2Din%20Feature%20Importancetext=It%20is%20a%20set%20of,sets%20with%20s
[Accessed 7 February 2022].