Sys Rev Pharm 2021; 12(5): 378-381
A multifaceted review journal in the field of pharmacy
E-ISSN 0976-2779 P-ISSN 0975-8453

Research Article

Detection of Exact Location of Brain Tumor from MRI Data Using Big Data Analytics

Lilly Sheeba, Anideepa Mitra, Saurav Chaudhuri*, Swarna Deep Sarkar
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

Article History: *Submitted: 23.04.2021* *Accepted: 07.05.2021* *Published: 14.05.2021*

ABSTRACT

MRI is the imaging technique most often used to detect brain tumor. A brain tumor is a knot, or mass, of abnormal cells in parts of the brain. Brain tumors can be either malignant or benign and can be located in the tissues of the brain. In this research study, a computerized approach has been presented where MRI grayscale images were assimilated for the detection of brain tumor. This study suggested a computerized approach that involves improvement at the elementary stage to reduce the gray-scale color variations. Filter operation was used to eliminate undesired noises as much as feasible to accommodate better segmentation. As this study test grayscale images therefore; threshold-based OTSU segmentation was used instead of color segmentation. Finally, specialists in the field of pathology provided feature intelligence that was used to recognize the zone of interests for brain tumor. This study pertained a novel architecture, named Xception, which permitted both elevated presentation, diminished expanse and estimated charge of deep neural networks employing depth wise separable convolution to establish high performance computer aided diagnosis system for brain tumor detection from MRI. Preparatory appraisal for the Xception model employing transfer learning exhibited exceptional performance with immense efficiency and prediction probability. Fascinatingly, prediction probabilities were distinct when various layers were reviewed.

Key words: Brain tumor, MRI, Big data, Prediction, Convolution, Segmentation

***Correspondence:**
Saurav Chaudhuri, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India, E-mail: saurav.chaudhuri11@gmail.com

INTRODUCTION

The brain is the most complex part of our body. Its functionality includes that of controlling our body functions, interpreting the environment around us-so that we can take the necessary actions, controlling our behavior and our responses and the seat of our intelligence. Given all of this, it is also the most mysterious part of our body, which makes it very difficult to diagnose any diseases that are related to it.

One such disease is the brain tumor. It is a collection of abnormal cells in any part of the brain, which can be cancerous or non-cancerous. Considering the location of the brain inside our skull, which is very rigid, a brain tumor can be hard to deal with or even locate (Makkie M, *et al.*, 2018).

The recent developments in the field of MRI and neuroimaging have been very helpful in dealing with these small masses of cells, though, it still has a very long way to go. With the advent of big data, it just might be possible. Big data, is used to analyze data sets that are too large to navigate through on a normal database. Generally, MRI scans and neuroimaging provides a lot of data and big data is very useful in this navigation of datasets. This, along with machine learning, where one can train the computer to deal with anything, can help in locating these small masses of cells, which are known as brain tumor.

The advantages behind using such a system includes model sequencing so that each sample can be assumed to be dependent on previous ones, convolutional layers can be employed to expand the effective pixel neighborhood, efficient handling of both linear and non-linear data, improving the performance in the target domain besides adaptively extracting specific features of interest (Siegal R, *et al.*, 2014).

MATERIAL and METHODS

Existing system

When the Human Brain Project and the Brain Initiative started out on its course to make neuroscience more approachable, it never considered big data as a possible mode of research and thus, very minor attempts were made in utilizing it. These projects, whose goal was to understand the complex structure of the brain along with its functionalities to make diagnoses in the fields easier, was generating large quantities of data. This vast amount of data put up a considerable challenge as it took a lot of time to decipher it. Though, Big Data was still an emerging name in the field of technology, it still wasn't developed enough that it could be adapted to this kind of large scale research. This paper discusses the on-going challenges in the field of neuroscience along the lines of big data. It also proposes a faster algorithm for processing the fMRI datasets that utilizes Spark and Hadoop, along with a data management system that can process such kind of vast data. It also proposes methods of advancements in executing a faster distributed dictionary learning.

New methods that can compile sophisticated analytical processes rapidly with more ease is seen in the existing system. Thus, allowing expert programmers to load and analyze data quickly and efficiently. In the already present system, we have seen that in the testing stage a portion of the datasets are compiled but for executing such large size of data that's not a convenient way. So, the process was not efficient (Jain S, 2013).

In the existing system the algorithm which was developed was known as HELPNI. The use of HELPNI was to store the data and it also compiled all the data (consisting of neuro images). HELPNI's another major task was to make sure the dataset (which consisted of multiple complex images) run in a controlled manner which can be controlled by the user easily (with the help of a simple User Interface). The next use of the said algorithm is to provide parallel and distributed accessibility to the developers, thus enabling them to enhance the model by introducing new methodologies using HELPNI. Thus, this algorithm was able to give information on neuro imaging using big data also giving the complexity of the system. It also facilitated feeding any form of data whether linear or nonlinear to the system. Moreover, we can say that due to this flexibility more methods can be added to the parallel system. The algorithm HELPNI consists of

five parts, which are data sharing tools, data storage, pipelining engine, user interface and data management tools. The system already existing is based on APACHE TOMCAT server 6.0, and the built is WAR process. This program uses basic JAVA commands such as GET, GIVE, DELETE and PUSH, which is used for managing the data. For installation and updating MAVEN is used. For generating reports, it uses turbine technique. The algorithm uses pipelining which is done by XML schema which is used to receive the attributes and resources with the help of a JAVA parser. Thus, the end result is achieved which includes a workflow of numerous applications and procedures.

Drawbacks

- Still missing-features problem exists
- Input is vastly different from the mean and therefore potentially erroneous
- Observations with large number of incomplete records
- Computational analysis impractical
- Do not generalize well across problem domains

Proposed system

Even with the new emerging technologies and the ground breaking research that is taking place in the field of medicine, it is still very difficult to diagnose a patient without any human interference. This is very evident in the field of neuroscience, especially when it comes to detection of the brain tumor. Though, advancing technology is still used to make the process of detection easier. One such technique is the process of Segmentation. This process is used to glean out the distinct features of the MRI brain images, so that they can be looked into for the purpose of interpreting it. The process of segmentation is used to identify and detect any abnormality that is present in the brain (Nandpuru HB, *et al.*, 2014).

An important step in the process of segmentation is identifying the image threshold, which is then used to find the histogram of the image. If the threshold has a symmetric distribution, then the Gaussian distribution is used to calculate the histogram, else if the threshold is non-symmetric, then, a Gamma distribution is used to find the histogram. The MRI brain images get approved by the proposed method.

The aim is to make the process of feature selection less cumbersome, such that it occupies less memory space and takes up less time to compute. Feature selection is used to extract the best features which has the most chances of yielding the results we are searching for. This is then used to calculate the variance and the feature with highest variance is checked out first *(Figures 1 and 2)*.

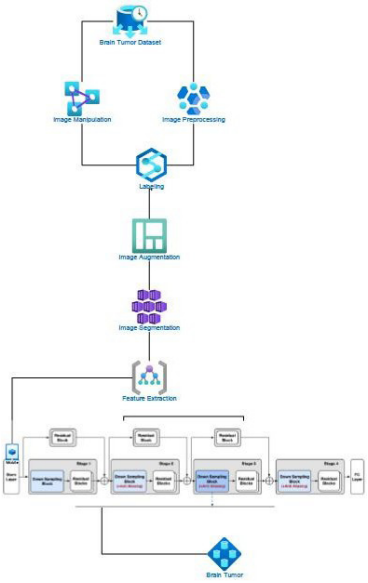

Figure 1: Architectural diagram

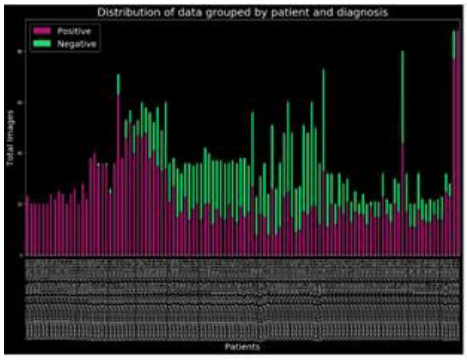

Figure 2: Type of diagnosis

RESULTS and DISCUSSION

Data import and processing

This process is used to work on images, here both the output and input are of the lowest level with respect to strength of the images. The basic use of this process is to improve the data which we receive from the images as the data includes a lot of unwanted features in the images which are not required (Guo L, *et al.*, 2011).

Color images are converted to grayscale to reduce complexity in computation. It is found that in certain problems, the main use of it is to remove unwanted information, thus reducing complexity *(Figure 3)*.

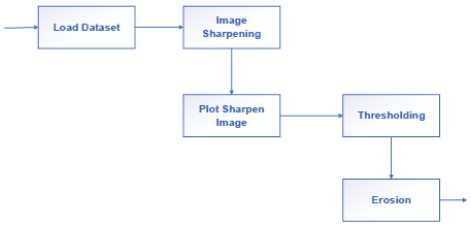

Figure 3: Data processing diagram

The images are converted to monochrome or greyscale as in this model color is not necessary preventing from unwanted data. In the research it has been found that non- greyscale images can cause a lot of unnecessary information which results in increased complexity, thus occupying a lot of space.

In convolutional neural network, the main task is to reframe all the images into a single dimension which is standard for all images present in the data. Thus all the data which is given as input has to processed and scaled before the algorithm studies the data *(Figure 4)*.

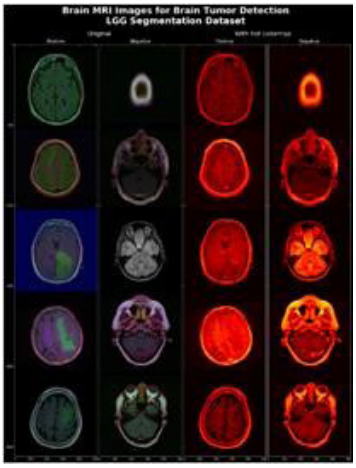

Figure 4: Brain segmentation

Data augmentation

To effectively implement the methods, a powerful image classifier is built, with the help of very few training examples, that is, a few hundred or thousand pictures taken from each class that needs to be recognized. In order to make the most out of the examples, it needs to be augmented via a number of random transformations, so that the same image is not checked twice. This prevents overfitting and makes generalization easier (Kumari R, 2013).

The tool used for image classification is a convent, which trains the data as an initial baseline. Given that there are only a few examples to be considered, the main concern here should be that of overfitting. A situation where the model reads very few samples and the resultant pattern cannot categorize the new data, is categorized as overfitting. This leads the model to start using irrelevant features to make predictions. An example of this would be, that if one was presented with three pictures of cats and three pictures of dogs and their basis of generalization is that, if the animal has whiskers, then it is a cat as opposed to a dog, then, they would categorize a dog as a cat. This leads to wrong classification and ultimately, to produce wrong data *(Figure 5)*.

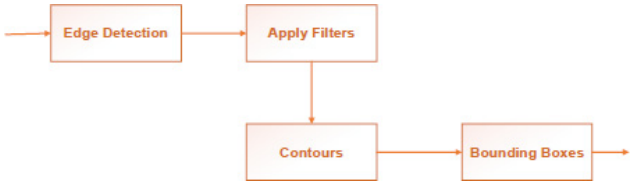

Figure 5: Data augmentation diagram

Data augmentation is used to deal with overfitting, but it loses its credibility, if the augmented samples are highly correlated. The main focus to counter overfitting should be on the entropic capacity of the model, i.e., the amount of information the model is allowed to store. A model with a lot of information has the potential to be more accurate, but it also faces the risk of storing irrelevant features. On the other hand, a model which stores less features, will have focus on the most significant details present in the data, and thus, is more likely to generalize better.

The entropic capacity has to be modulated for better results. The main procedure to modulate it, is to do the selection based on the number of parameters present in the model, i.e., the number of layers present and the size of each layer. Another way to modulate is by utilizing weight regulization, which is implemented by forcing the model weights to take up smaller values. The weight regulization can be seen in the L1 or L2 regulization (Gordillo N, *et al.*, 2013).

Model building

The cov1 layer is of a dimension of 224 x 224 RGB image. The image of the fixed size is then sent within a convolutional layer, which has tiny receptive filters of size 3 X 3, the alternate has a 1 X 1 dimension, also known has linear transformation of input channel. The next step is non-linear transformation. The stride has a fixed size of 1 pixel. The layer also consists of a padding of 1 pixel. The contents of spatial pooling are 5 pooling layer and then conv. Layer, the size of the window is 2×2.

After this, a Model Checkpoint is used to monitor a specific parameter of the model to try and save it. In the training example, the validation accuracy was monitored by passing the val_acc to the Model Checkpoint. If the validation accuracy in the current epoch is found to be greater than that of the last epoch, then the model gets saved to the disk *(Figure 6)*.

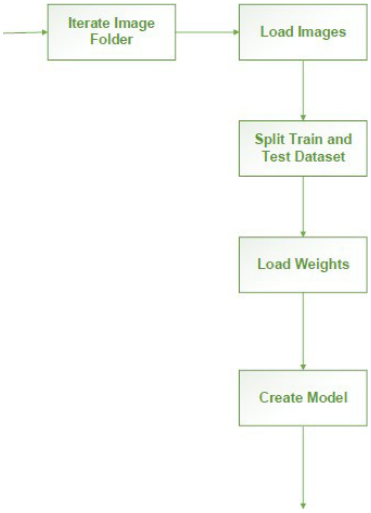

Figure 6: Model building diagram

Model performance

While training the classification predictive model, it is very important to assess the quality of the model. There are many different ways of evaluating the performance, of which most data scientists use Python package called Scikit-learn for predictive modelling. It contains many built-in functions that can be used to analyze the performance of the models (Demirhan A, *et al.*, 2014).

Considering an actual label and a predicted label, the confusion_matrix divides the samples into 4 different buckets:

- For True positive bucket, both the actual label and the predicted label is found to be 1.
- For False positive bucket, the actual label is found to be 0 and the predicted label is found to be 1.
- For False negative bucket, the actual label is found to be 1 and the predicted label is found to be 0.
- For True negative bucket, both the actual label and the predicted label is found to be 0 *(Figures 7-8)*.

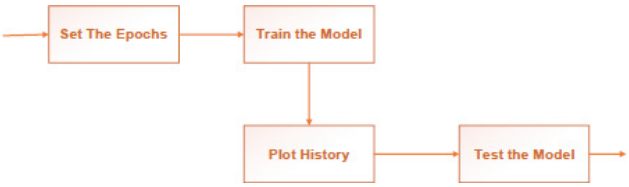

Figure 7: Model performance diagram

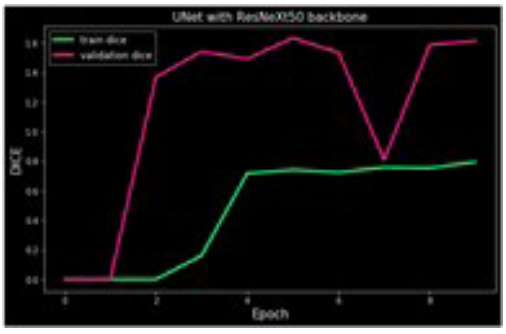

Figure 8: Feeding data and training it

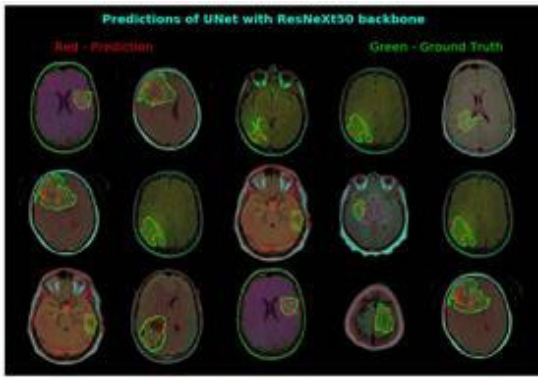

Figure 9: Final prediction of location of brain tumor

CONCLUSION

This paper proposes a faster method for the organization and identification of the brain tumor that might be present in the MRI images. It proposes Faster R-neural network for this. This is better at classifying data and obtaining the required information than the previous R-neural network. This paper also proposes a better method for image thresholding. Instead of using the Gamma distribution that is already being used in other major research, this paper uses the Gamma distribution to classify the image threshold and find the variance of it. This increases the probability in detecting a potential brain tumor.

The final output shows the predicted tumor locations that has been provided by the algorithm. The lines marked in red are the locations that the algorithm predicted to be a potential tumor whereas, the lines marked in green are the actual locations of the tumor. The output provided goes on to show that the algorithm is good at predicting the locations and with a few more rectifications, it would point out the tumors with more ease.

ACKNOWLEDGMENTS

Doing this project would not have been accomplished if we didn't have the support of Dr. Lilly Sheeba who has guided us in every part of project and make it a success. While working on this project we gained a lot of knowledge about the rapid growing fields of Big Data and Machine Learning and its contributions towards the neuroscience and neuroimaging fields along with its utilization in medical field and Brain diagnosis projects.

REFERENCES

1. Makkie M, Li X, Quinn S, Lin B, Ye J, Mon G, *et al.* A distributed computing platform for fMRI big data analytics. IEEE transactions on big data. 2018; 5(2): 109-119.
2. Siegal R, Miller KD, Jemal A. Cancer statistics, 2012. Ca Cancer J Clin. 2014; 64(1): 9-29.
3. Jain S. Brain cancer classification using GLCM based feature extraction in artificial neural network. International Journal of Computer Science and Engineering Technology. 2013; 4(7): 966-970.
4. Nandpuru HB, Salankar SS, Bora VR. MRI brain cancer classification using support vector machine. 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science. 2014; 1-6.
5. Guo L, Zhao L, Wu Y, Li Y, Xu G, Yan Q. Tumor detection in MR images using one-class immune feature weighted SVMs. IEEE Trans Magn. 2011; 47(10): 3849-3852.
6. Kumari R. SVM classification an approach on detecting abnormality in brain MRI images. Int J Eng Res Appl. 2013; 3(4): 1686-1690.
7. Gordillo N, Montseny E, Sobrevilla P. State of the art survey on MRI brain tumor segmentation. Magnetic resonance imaging. 2013; 31(8): 1426-1438.
8. Demirhan A, Törü M, Güler İ. Segmentation of tumor and edema along with healthy tissues of brain using wavelets and neural networks. IEEE J Biomed Health Inform. 2014; 19(4): 1451-1458.