**DATA PROCESSING AND PRELIMINARY LEARNING ALGORITHM**

**Data Preprocessing**

For the development of the project the initial dataset of Brain Tumor images was collected from The Cancer Imaging Archive (TCIA). These images were downloaded in Digital Imaging and Communications in Medicine DICOM format. DICOM images are the images produced by a MRI machine and is a 3-Dimensional portrayal of the body part. The DICOM image is then converted into JPEG image format using an online convertor which gives out the frontal view of the brain.

Preprocessing the DICOM images to extract the tumors or in turn their features adds to the computation of the system and it takes way longer time to process and generate results. It does though at times provide a better result. Hence in this case the program is chosen to work with JPEG images generated from the DICOM images.

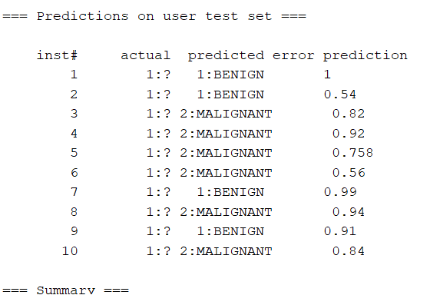
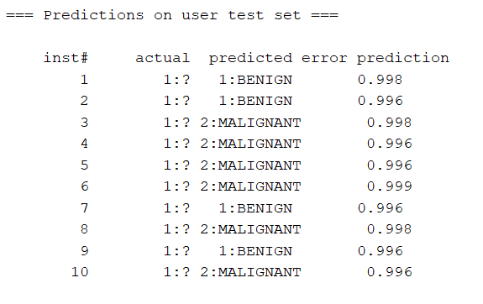
Any image before it can be processed to extract the regions of interests need to be preprocesses so as to make it suitable for usage for the computer. A series of preprocessing steps are applied to the MRI generated image like conversion to GreyScale, Histogram Equalization, Thresholding and the Contrast Enhancement using Mathematical Morphology. The Histogram Equalization uses the plot of greys in the image and spreading out most frequent intensity values on the graph hence effectively stretching the intensity range of the graph. Further the Contrast is corrected using a convolving function. Once enhanced the tumor image is resected out using the most recent and highly accurate Edge Detection technique called the Sub-Pixel detection method that draws a finer edge over the edges detected by other Edge detectors such as Canny or Devernay algorithm. Further the dissected edges are treated with the Gaussian Filter in order to extract the Tumor from the image. This Tumor image is then used to calculate a total of 13 mathematical features corresponding to an image. These 13 features act as attributes for further classification on the Brain Tumors into Malignant and Benign. These attributes include the Contrast(the measure of luminance and color that makes it possible to distinguish objects), Correlation(the computation of  sum of products at each location with respect to a filter being displaced over the image), Energy(captures the desired solution and performs descent function over the gradient in order to compute its lowest value), Homogeneity(the measure related to the changes in the intensity of a region in an image), Mean(the average pixel value), Standard Deviation(the measure of the deviation of the pixel values), Entropy(The measure of the intensity levels of corresponding states that individual pixels can adapt), Root Mean Square(The measure of the differences in the levels of the column vectors), Variance(A variance image is an image of the the squares of Standard Deviation), Smoothness(The information of pixel pattern generated after the removal of outliers within the pixel data), Kurtosis(The measure interpreted in combination of the resolution and noise in the pixel set of the data), Skewness(the measure of feature of image that determines it’s darkness, glossiness, matte or the lightness) and Inverse Difference Moment or IDM (the measure of local homogeneity of the image.

These values for all the images becomes the instances of ARFF file which will further be run on the classifiers to build a Machine Learning model to predict the class of the tumor from Malignant or Benign.

**Core Algorithm**

The ARFF file created was run on WEKA to get classifications for the test set of data using the Random Forest and the k-Nearest Neighbors algorithm. A separate test data set was created from the dataset containing 10 instances which were then added in WEKA as supplied test set after running the train data on cross validation for 10 folds. The screenshots for the predictions for both the algorithms run are as shown below.

According to the initial analysis on the dataset the k-Nearest Neighbors Algorithm gave a pretty good accuracy on the supplied dataset. A further analysis of the k-Nearest Neighbors with the Random Forest Algorithm was down to find out the Summary Statistics for the 2 algorithms for a deeper comparison of the algorithms.

Screenshot for Random Forest Screenshot for k-Nearest Neighbors

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F-measure** | **ROC** |  |
| IBK | 75.96% | 0.76 | 0.76 | 0.76 | 0.752 | Cross-Validation |
| IBK | 72.15% | 0.722 | 0.722 | 0.722 | 0.711 | Percentage Split |
| IBK | 99.14% | 0.992 | 0.991 | 0.991 | 1 | Use Training Set |
| Random Forest | 87.1245 % | 0.88 | 0.848 | 0.864 | 0.938 | Cross-Validation |
| Random Forest | 86.0759 % | 0.794 | 0.871 | 0.831 | 0.941 | Percentage Split |
| Random Forest | 99.1416 % | 1 | 0.982 | 0.991 | 1 | Use Training Set |

The above analysis proves that Random Forest Algorithm provides a better accuracy of the two algorithms for the dataset. A further analysis will be done with other algorithms to determine the best algorithm for classification of Brain Tumors into Malignant or Benign using the acquired dataset.

The features extracted for the project were done used by an Image Analysis project developed by one of the team mates Ayush Arora.