Predictive Modeling for Identifying the Human Brain Developmental Stages

Farhan Khan
Department of Electrical Engineering
Indian Institute of Technology
New Delhi, 110016
Email: Farhan.khan.0039@gmail.com

Tapan Gandhi
Department of Electrical Engineering
Indian Institute of Technology
New Delhi, 110016
Email: tgandhi@iitd.ac.in

Abstract - **The human brain undergoes various structural and functional changes as the age progress. These changes are however gradual and a substantial change is seen only after a certain period of years. This period of years of human life can be hypothesized as the developmental stage of the brain. The work described here proposes novel methodologies to find the developmental stages of the human brain. Functional magnetic resonance imaging(fMRI) techniques which have become widely popular in the research community are used in this study. This work uses resting state fMRI data of 1096 healthy subjects(age 9 to 83 years) to find the developmental stages of the brain. A classification model using SVM is also proposed which can classify the subject based on their fMRI data into the developmental stages. The proposed model is 90.4% accurate with specificity and sensitivity as 94.94% and 85.9%. This model can be used in cases of degenerative neurological disorders which lead to abnormal development of brain. For the subjects with a neurological disorder, their brain developmental stage can be obtained using the model and hence inferences can be drawn about the growth of the disorder.**

Keywords **: fMRI, BOLD, time series, developmental stage, functional connectivity, structural connectivity, K-means, SVMs, approximate entropy**

I. Introduction

The human brain is a complex organ the complete understanding of whose functioning is still unclear[1]. Many studies have been performed in the past to unravel the functioning behind brains working, however, nowadays neuroimaging techniques like fMRI[2] have become hugely popular in the research community to study the functioning of the human brain. It has been observed that the human brain undergoes various structural and functional changes as the age progress[3]. Different brain regions develop and the wiring between these regions strengthens. Change in functional connectivity in the human brain is, however, a gradual process with a substantial change is seen only after a certain period of years. During these years the brain's functional and structural connections remain almost constant. It can be hypothesized that brain develops with these period of years as the developmental stages. Identifying these developmental stages is of great importance in neuroscience as it can help in increasing our understanding of the human brain.

The work described here had two objectives. The first objective was to find the developmental stages of the human brain. The second objective was to develop model which can classify the subject into its developmental stage using his fMRI data. This model can also be used for predicting whether a subject has a neurological disorder or not.

II. Materials and Methods

A. Dataset Description

The data used in this project is the Resting state fMRI data which is obtained from Human connectome project (http://www.humanconnectomeproject.org/). It is the initiative taken by leading universities of the world to acquire fMRI data for research purpose. It consists of rs-fMRI data of 1096 "healthy" subjects with age ranging from 9 years to 83 years. The data for each subject was acquired on 3.0 Tesla Magnetom Tim Trio scanner. The resting state data was acquired for 500 seconds for each subject with the TR of 2.5 seconds

After preprocessing the data, BOLD activation time series[4] corresponding to each voxel region of the brain is obtained. 160 ROIs were then apriorly selected and Mean time series within the spherical range of 8 mm radius is calculated. Final data for each subject consist of 160 time series of length 190, a matrix of shape 190*160.

B. Statistical feature extraction

For the purpose of analysis, a new dataset is created from the fMRI data of subjects by extracting a feature from the BOLD activation time series of the ROIs. Approximate entropy(ApEn) is chosen as a feature to represent the time series. It is a "regularity statistic" that gives unpredictability of amplitudes in the time series[5].

978-1-5386-1887-5/17/$31.00

The presence of similar patterns of amplitudes makes a time series more predictable. ApEn gives the likelihood that similar amplitude values will not be followed by additional similar values. A time series with less amount of repetition has a smaller value of ApEn whereas a more complex one has higher ApEn value. The time series data used for the analysis has information in its amplitudes only thus ApEn can be used as a feature of the time series.

Mathematically it is given as,

$$ApEn = \ln\left[C_m/C_{m+1}\right] \tag{1}$$

where, *C* is the normalized number of repetitive subsequences in the time series and *m* is the length of subsequence.

C. K-means clustering

K-means is one of the unsupervised machine learning algorithms which is used for clustering problems. The algorithm aims at classifying the sample points into different clusters decided priorly. K-means uses Expectation Maximization(EM) algorithm to determine cluster centers and to assign cluster labels to each sample.

The k-means algorithm[6] is described in following steps:

1) Initialize cluster centers $\mu_1, \mu_2, \mu_3, ...\mu_k \in \mathbb{R}^n$ randomly.
2) Repeat till convergence:

   For each *i*,
   $$c_i = \underset{j}{agrmin}||x_i - \mu_j||^2 \tag{2}$$
   For each *j*,
   $$\mu_j = \frac{\sum_{i=1}^{m}(c_i = j)x_i}{\sum_{i=1}^{m}(c_i = j)} \tag{3}$$

In above algorithm, *k* is the number of clusters decided a priorly and μ_j represents cluster centers which are decided by choosing *k* samples points randomly. At each iteration of step 2, the cluster centers are updated and new clusters are assigned to the sample points. The algorithm converges when no change in cluster center is seen.

D. Support Vector Machines(SVMs)

Support Vector Machines is a set of supervised machine learning algorithms which can be used for both classification and regression problems. SVMs performs classification by creating hyperplane in a multidimensional space such that the hyperplane separates the samples belonging to different classes. SVMs also perform well in the settings where the data is not linearly separable with the introduction of kernel functions. These functions transform the data into higher dimensional space in which the data is linearly separable.

The model for SVM[7] is given as,

$$y(w, x) = sign(w^T x + w_0) \tag{4}$$

where, $w^T x + w_0$ is the equation of the hyperplane defined by the weight vector $w \in \mathbb{R}^d$. The weight vector *w* is obtained by minimizing the cost function given as,

$$\frac{1}{2}w^T w + c\sum_{i=1}^{N}\xi_i \tag{5}$$

subject to the constraints:

$$y_i(w^T\phi(x_i) + w_0) \geq 1 - \xi_i, i = 1, 2....N \tag{6}$$

$$\xi_i \geq 0, i = 1, 2....N \tag{7}$$

The above equations are solved using Lagrange and SMO algorithm to obtain w vector.

E. Performance measures

Consider the confusion matrix of any classification model

Table I
CONFUSION MATRIX

TP	FN
FP	TN

The targets predicted by the classification model can be divided into four different categories-

1) **True Positive(TP)** : The number of samples which belong to the class and are correctly classified as such.
2) **False positive(FP)** : The number of samples which do not belong to the class and are wrongly classified as such.
3) **True Negative(TN)** : The nuber of samples which do not belong to the class and are correctly classified as such.
4) **False Negative(FN)** : The number of samples which belong to the class and are wrongly classified.

Following performance measures were calculated for the model-

- **Accuracy** : It is the measure of number of samples correctly classified by the model. It is given by,
  $$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$
- **Sensitivity** : It is the measure of models capability to avoid the false negatives. Its higher value indicates lesser false negative rate of the model. It is given by,
  $$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$
- **Specificity** : It is the measure of models capability to avoid the false Positives. Its higher value indicates lesser false positive rate of the model. It is given by,
  $$Specificity = \frac{TN}{TN + FP} \tag{10}$$

III. Experiments

A. Age bins analysis

A novel methodology is adopted to obtain the period of years(age bins) or developmental stages over which the brain functionality remains almost similar. This section describes a framework which uses K-means clustering for obtaining the age bins from the fMRI data of the subjects.

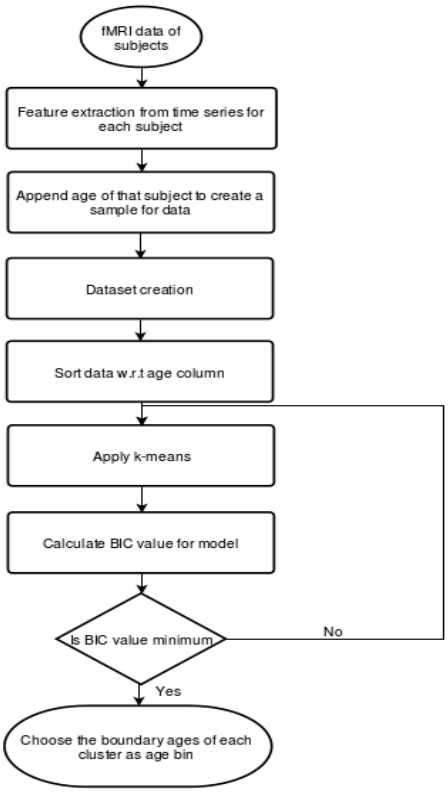

Figure 1. Framework to obtain age bins

Above figure shows the framework adopted to find the age bins. Firstly a new dataset is created using fMRI data of subjects by calculating Approximate entropy(ApEn) from each BOLD activation time series. A new entry named ”Age”, which has the age of the subjects, is also included in the dataset.

The samples of the data are first sorted according to the age column and k-means clustering with BIC as a measure for model selection is performed. BIC is based on calculating the likelihood function and it also includes penalty term for the number of parameters. The model with the lowest BIC value is chosen as the best model. The number of clusters while performing k-means clustering are increased in each iteration till the BIC score is minimum. Model with minimum BIC score is selected and ages of the subjects which are lying at the boundary in each cluster are taken as age regions or age bins.

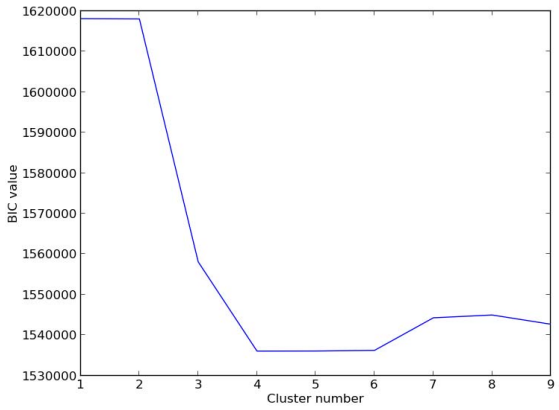

Figure 2. BIC curve for model selection

The above BIC curve shows that the best model is the one which has 4 clusters.

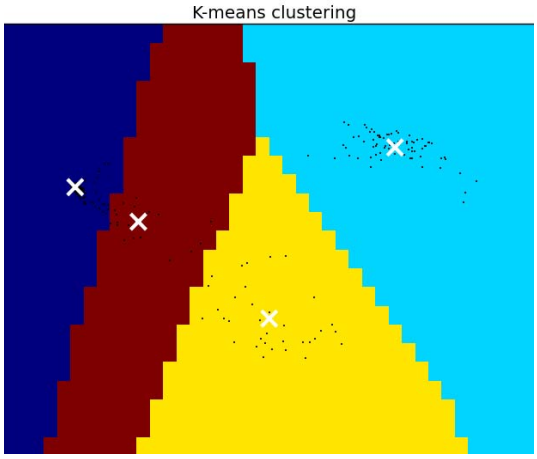

Figure 3. Clustering of dataset

K-means clustering is performed using 4 as cluster size and different age bins are obtained using the framework described earlier in this section. Above figure shows clustering of dataset with 4 clusters. The obtained results are - **9-18, 19-32, 33-51, 52-83** years.

B. Prediction of brain development

Degenerative neurological disorders lead to abnormal development of human brain. This fact can be used to create a model which can predict the brain development of a subject and hence inferences can be developed to predict the growth of disorders. This section describes a framework which uses SVM as a classification algorithm to obtain the developmental stage of brain for a subject.

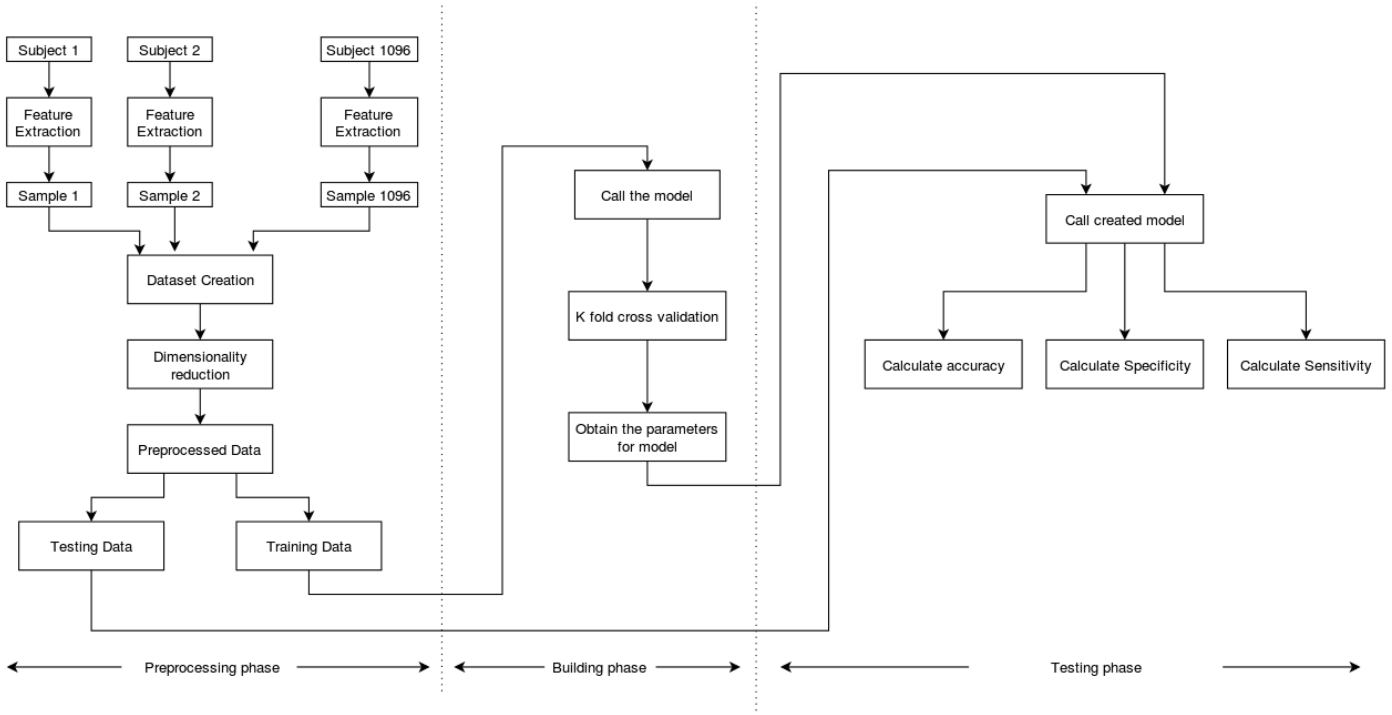

Figure 4. Model for Brain development prediction

Above shown figure describes the model used for predicting the brain development of the subject. After age bins analysis, 4 different age bins were obtained. Thus it can be hypothesized that human brain develops in 4 different stages of life. This model predicts the developmental stage of a subject using its fMRI data. It has three different phases Preprocessing phase, Building phase and Testing phase.

During Preprocessing phase, data is made ready to be used for analysis. From the fMRI data of subjects firstly a new dataset is created after feature extraction. This new dataset is then reduced in dimensions using Principal Component Analysis(PCA) by considering those dimensions which contain 90% of total variance in the data. After PCA the data is then split into two parts, Training data and Testing data.

During Building phase, the actual model is created. The model is firstly called on the training data and is then cross validated for obtaining the parameters for the model. Here in this work Support vector Machines(SVMs) is used as the algorithm for building the model. Cross validation is performed using k-fold cross validation method for setting the parameters for the model. Cross validation accuracy is taken as a measure for choosing the best model.

During Testing phase, the model is tested on the unseen testing data. The goodness of the model is decided through performance measures such as accuracy, specificity and sensitivity

SVMs are known for their versatility because they can be used in different settings. The dataset used here was skewed with different classes having a different number of subjects. It has 4 classes each having subjects from the age bins obtained in earlier study. Class1 - 216, Class2 - 691, Class3 - 113, Class4 - 76. SVM support classification of skewed dataset by minimizing the cost of misclassification and assigning different weights to classes.

Cross validation with SVM as a model along with misclassification cost minimization is performed for model selection by 10-fold cross validation method. Obtained cross validation accuracy is 92%. Model is then tested on testing data and the obtained ROC curve and confusion matrix are shown below

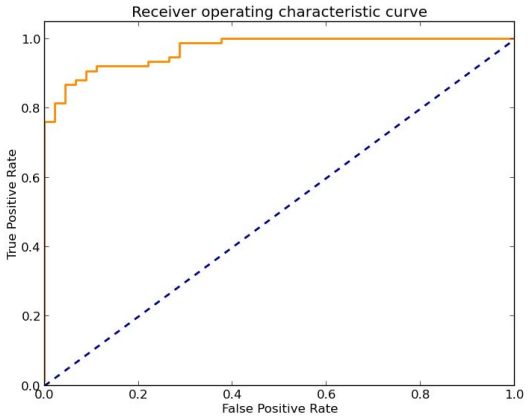

Figure 5. ROC curve for SVM

Table II
CONFUSION MATRIX FOR SVM

65	8	4	0
5	244	2	0
1	2	36	1
4	8	3	13

From the above confusion matrix the obtained accuracy, specificity and sensitivity are **90.4%**, **94.94%** and **85.9%** respectively.

IV. DISCUSSION AND CONCLUSIONS

The human brain is a complex organ the complete understanding of whose functioning is still unclear. It has been observed that with age the human brain undergoes various structural and functional changes. However, there occurs a certain period of years in which the brain's structural and functional connection remain almost constant. It can be hypothesized that the human brain develops in stages. This period of years in which brain's structural and functional connections remain almost constant are termed as developmental stage of brain. This work proposes a novel methodology for obtaining these age bins or developmental stages. It was found that human brain develops in 4 different stages with period of years as - **9-18, 19-32, 33-51, 52-83** years.

In this work a model is also proposed which can predict the brain developmental stage of a subject based on its fMRI data. This study is useful for predicting the growth of degenerative neurological disorder in a subject. These type of disorders can lead to abnormal development of the brain. By predicting the brain developmental stage of a subject some inferences can be developed about the growth of the disorder. The proposed model uses SVM as classification algorithm and gives **90.4%** accuracy, **85.9%** sensitivity and **94.94%** specificity.

REFERENCES

[1] Van Den Heuvel, Martijn P., and Hilleke E. Hulshoff Pol. "Exploring the brain network: a review on resting-state fMRI functional connectivity." European neu- ropsychopharmacology 20.8 (2010): 519-534.

[2] Lindquist, Martin A. "The statistical analysis of fMRI data." Statistical Science (2008): 439-464.

[3] Ferreira, Luiz Kobuti, and Geraldo F. Busatto. "Resting-state functional connec- tivity in normal brain aging." Neuroscience Biobehavioral Reviews 37.3 (2013): 384-400.

[4] Di, Xin, et al. "Calibrating BOLD fMRI activations with neurovascular and anatomical constraints." Cerebral Cortex 23.2 (2013): 255-263.

[5] Pincus, Steven M. "Approximate entropy as a measure of system complexity." Proceedings of the National Academy of Sciences 88.6 (1991): 2297-2301.

[6] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[7] Scholkopf, Bernhard, and Alexander J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

[8] Dosenbach, Nico UF, et al. "Prediction of individual brain maturity using fMRI." Science 329.5997 (2010): 1358-1361.

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[10] Hedden, Trey, and John DE Gabrieli. "Insights into the ageing mind: a view from cognitive neuroscience." Nature reviews neuroscience 5.2 (2004): 87-96.

[11] Biswal, Bharat B., et al. "Toward discovery science of human brain function." Proceedings of the National Academy of Sciences 107.10 (2010): 4734-4739.