Cross-Validated AdaBoost Classifier used for Brain Tumor Detection

[1]Shawni Dutta and [2]Prof. Samir Kumar Bandyopadhyay

[1]Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.

[2] Academic Advisor, The Bhawanipur Education Society College, Kolkata, India

Abstract-

Brain Tumor is one of the severe diseases and occurrence of this disease threats human life. Detection of brain tumor in advance can secure patient's life from unwanted loss. Well-timed and swift disease detection and treatment strategy can lead to improved quality of life in these patients. This paper attempts to use Machine Learning based ensemble approaches for recognising patients with brain tumor. Ensemble technique based AdaBoost classifier and 10-fold stratified cross-validation method are assembled in single platform is proposed in this paper for prediction of brain tumor. This prediction is compared against three baseline classifiers such as Gradient Boost, Random Forest and Extra Trees classifier. Experimental result implies the superiority of this model with an accuracy of 98.97%, f1-score of 0.99, kappa statistics score of 0.95 and MSE of 0.0103.

Keywords: Brain Tumor, Machine Learning, Ensemble techniques, AdaBoost, Cross-Validation, Stratified technique.

Introduction-

Brain tumour, stroke, hemorrhage and multiple sclerosis (MS) disease always threat to human as the life threatening diseases in both male and female. The most common and widespread disease is Brain tumour amongst various brain diseases. It is necessary to detect early and accurate diagnosis of brain lesion for determining accurate treatment and prognosis. However, the diagnosis can only be performed by specialists in neuroradiology. It is a very challenging task. Various factors lead to abnormal brain lesion development include brain injuries, multiple sclerosis, hemorrhage, stroke, vascular disorders and brain tumours. Diagnosis for brain lesions depend on the type of lesion, the age and health condition of the patient, and how effective treatments are for the patient. For analysis brain tumour, specialists are required to examine and confirm of each medical report after proper investigations. Depending on the condition of the patient it is first necessary biopsy of the place and if exigency occurs surgery has to made to cure the disease.

The most common and aggressive primary brain tumour in adults is Glioblastoma multiforme (GBM). In severe condition of the patient it is required surgical resection followed by adjuvant radiotherapy with concurrent chemotherapy. In recent times the survival rate of patients with GBM has improved with advancements of treatment. Still the prognosis rate remains generally poor. The survival rates are in the range of 8%–12% [1].

 doi:10.20944/preprints202006.0351.v1

Using Machine Learning (ML) approaches, early prediction of any diseases can be performed accurately. Classification algorithm belongs to the category of ML approaches that maps input data into target class while learning and extracting features from training data. This paper aims to predict whether a patient can have tumors in brain or not. Patients' historical records are utilised for extracting hidden patterns for the purpose of recognising patients with abnormality in brain. An automated tool is recommended in this paper that captures interfering factors for occurrence of brain tumor disease and finally predicts brain tumor tendency of patients. This paper exemplifies the use of ensemble based [2] ML techniques for brain tumor detection. The reason of using ensemble based techniques is to produce improved accurate results over single classifier model. Ensemble techniques are known to be as meta-algorithms that assemble decisions from multiple base models into single predictive model [2]. Boosting is a technique that produces ensemble model which is implemented in this paper. This paper proposes AdaBoost algorithm [3] along with 10-fold stratified cross-validation methodology for the purpose of brain tumor prediction. For justifying the efficiency of the proposed classifier this paper also implements other ensemble techniques such as Gradient Boost [4], Random Forest [5] and Extra Trees classifier [6]. These classifiers create a baseline for comparing the performance of the proposed classifier. All these ensemble technique based classifier models are evaluated with some predefined metrics such as accuracy, Cohen-kappa score, f1-score and MSE. An efficient and accurate classifier model with minimised classification error is preferred in the domain of brain tumor detection.

Related Works-

In recent years, numerous researches have been carried out in order to detect tumor in brain MRI images. Applications of ML techniques, image enhancement techniques are employed by many researchers in the field of brain tumor detection.

Supervised ML algorithm, classification techniques such as naïve bayes, neural network, J48, Lazy-IBk implemented and applied on MRI image to detect whether it contains brain tumour or not [7]. For differentiating normal and abnormal brain tumors in the MRI images, techniques like segmentation, feature extraction and feature reduction mechanisms are applied. For segmentation, k-means clustering algorithm is applied. Using Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA), feature extraction and feature reduction techniques are implemented respectively. A classifier model Support vector machine (SVM) is applied that classifies the abnormal brain tumors into LGG and HGG [8].

A combinational algorithm of FCM clustering and SVM classifier is proposed in [8] for classification of the tumors in combination with BCFCM for bias field correction and HAAR wavelet transform for feature extraction. The proposed method achieves promising accuracy of 98.2% [9]. Using image processing and feature extraction techniques are employed in [10] for brain tumour classification and segmentation. Image processing techniques are applied for noise elimination and image enhancement process. Textual

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2020 doi:10.20944/preprints202006.0351.v1

features are extracted and reduced using PCA methodology. Experimental results presented Dice score of 0.95 as prediction results [9]. Some researches such as [11-12] demonstrated the use of unsupervised algorithm such as clustering. K-Means clustering algorithm is used by [11] whereas [12] utilised fuzzy clustering algorithm for segmentation purpose.

Deep Learning (DL) techniques such as Convolutional Neural Network (CNN) is utilized in [13] with a kernel size of 3x3, automatic classification and segmentation of brain tumor is detected from MRI images. Another study [14] used CNN for automatic brain tumor detection with successful classification rates of 97.5%.

Dataset Used-

This paper collects *Brain Tumor Dataset* available at kaggle [15]. This dataset consists of 1644 number of patient's records and each record is formulated as collection of 18 attributes. The dataset includes five first-order feature and eight texture feature and four quality assessment parameters with the target level. The first-order feature set contains attributes such as Mean, Variance, Standard Deviation, Skewness, and Kurtosis. Contrast, Energy, ASM (Angular second moment), Entropy, Homogeneity, Dissimilarity, Correlation, Coarseness are the attributes included in second-order texture feature set. There are four quality assessment parameters such as PSNR (Peak signal-to-noise ratio), SSIM (Structured Similarity Index), MSE (Mean Square Error), and DC (Dice Coefficient). All these features are extracted from MRI images. Infinite values and Not a Number (NaN) values are present in this dataset. Presence of these values will change the prediction efficiency. However, the presence of missing values can be ignored or deleted when the number of missing values is less in percentage. In some cases, it is required to consider unknown or missing values present in the dataset since these may contribute to the disease. In our implementation, missing values are handled by replacing zeroes. Table 2 summarises the occurrence of missing values in the dataset.

Attribute Name	Number of missing values
Skewness	369
Kurtosis	369
PSNR	98
SSIM	369
DC	98

Table 2: Presence of missing values in the dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2020 doi:10.20944/preprints202006.0351.v1

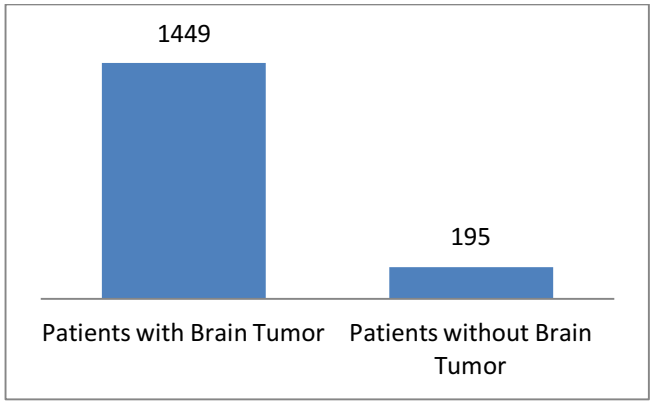


Fig1: Distribution of Target variable in the dataset

Proposed Methodology-

This paper attempts to use classifier model based on ensemble techniques [1]. The target of any classifier is to associate input variables into target variables considering the training dataset. The proposed classifier employs boosting techniques which is a type of ensemble approaches in order to identify whether a patient has brain tumor or not. Basically, binary classification problem is addressed in this paper. Boosting techniques [2] are capable of obtaining highly efficient prediction results by combining weak and inaccurate learners. AdaBoost is known to be the first boosting technique proposed by Freund and Schapire [2]. This algorithm also belongs to the category of interpolating classifiers which defines algorithmic property of fitting the training data completely without error. This algorithm also exhibits the property of self-averaging property. Considering these two properties will assist in obtaining low generalization error. This classifier is also known as a meta-estimator that proceeds by fitting a classifier on the original dataset and additional copies of the classifiers are fitted after re-weighting the incorrectly classified instances in such a manner that the classifier is capable in handling more difficult cases [2].

This AdaBoost classifier is implemented using 500 base estimators. The base estimator considered in this case is Decision Tree (DT) classifier. Basically, the target of this implementation is to enhance the efficiency of DT classifier. The learning rate of this classifier is set to 1.0. Using discrete boosting algorithm known as SAMME algorithm, this model is constructed. The description of the implementation is summarised in Table 1.

Parameters Used	Values
Base Estimator Used	Decision Tree (DT) Classifier
Number of Base estimators	500
Learning Rate	1.0
Algorithm Used	SAMME

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2020 doi:10.20944/preprints202006.0351.v1

Random State	1

Table 1: Implementation details of AdaBoost Classifier.

Once the model is implemented, it is followed by 10-fold cross-validation method [14] in order to estimate the skill of the model. It is a resampling methodology where the dataset is partitioned into 10 groups and in each iteration one group is considered as the test data and the remaining nine folds are considered as training data. The above mentioned model is fitted into the training dataset and it is evaluated against the test dataset. Later evaluation scores for each of these iterations are accumulated and mean score is calculated. The implementation of cross-validation ensures stratified mechanism which enforces that the distributions of all folds are necessarily similar to proportion of all labels in the original data [14].

Baseline Classifier-

This section describes ensemble technique based classifiers such as Gradient Boosting classifier, Random Trees, and Extra Trees classifier. The proposed classifier is justified against these mentioned classifiers. Hence, they are providing a baseline platform for comparing the prediction performance of proposed classifier.

Random forest (RF) [4] exemplifies the concept of ensemble learning approach and applies regression technique for classification based problems. This classifier is a combination several tree-like classifiers which are applied on various sub-samples of the dataset and each tree cast its vote to the most appropriate class for the input.

Extra Trees Classifier [5] belongs to the category of ensemble learning technique. It aggregates the outcomes of various de-correlated decision trees collected in a "forest" and delivers output as classification result. The Extra-Trees algorithm creates an ensemble of un-pruned decision or regression trees. It has two main differences with other tree-based ensemble methods. The splitting of nodes is done by randomly choosing cut-points. Then uses the whole learning sample (rather than a bootstrap replica) to grow the trees.

Gradient Boost (GB) algorithms [3] another boosting algorithm which are suitable in fitting new models to obtain maximised efficiency while estimating response variable. The objective of this algorithm is to construct new base learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. This algorithm is highly customizable to any domain which provides freedom in model designing. One of the important issues of this algorithm is identifying and incorporating loss function to this algorithm which is subject to change as a matter of trial and error [3].

Implementation of baseline classifiers-

The baseline classifiers are applied on the dataset after partitioning it into training and testing dataset with a ratio of 7:3. The training set is fitted to these classifier models and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2020 doi:10.20944/preprints202006.0351.v1

later prediction is retrieved for the testing dataset. The GB classifier is implemented with 500 base estimators, learning rate of 1.0. The RF classifier is also implemented with 500 base estimators whereas; Extra Trees classifier is designed with 500 numbers of trees in the forest. These designed ensemble models with necessary tuning will assist in attaining best results.

Performance Evaluating Metrics-

While evaluating performance skill of a model, it is necessary to employ some metrics to justify the prediction results. The purpose of metrics is to pick up top models based on their performances. The abovementioned classifier models are compared with respect to the following performance evaluation metrics. Use of the following metrics will assist in attaining best problem solving approach.

1. Accuracy [16] is a metric that detects the ratio of true predictions over the total number of instances considered. However, the accuracy may not be enough metric for evaluating model's performance since it does not consider wrong predicted cases. Hence, for addressing the above specified problem, precision and recall is necessary to calculate.
2. Precision [17] identifies the ratio of correct positive results over the number of positive results predicted by the classifier. Recall denotes the number of correct positive results divided by the number of all relevant samples. F1-Score or F-measure [17] is a parameter that is concerned for both recall and precision and it is calculated as the harmonic mean of precision and recall. The best value of F1-score, precision, and recall is known to be 1.
3. Mean Squared Error (MSE) [17] is another evaluating metric that measures absolute differences between the prediction and actual observation of the test samples. MSE produces non-negative floating point value and a value close to 0.0 turns out to be the best one.
4. Cohen-Kappa Score [18] is also taken into consideration as an evaluating metric in this paper. This metric is a statistical measure that finds out inter-rate agreement for qualitative items for classification problem. The kappa statistic outputs value in the range of -1 to +1 and +1 indicates the maximum chance of agreement.

Experimental Results-

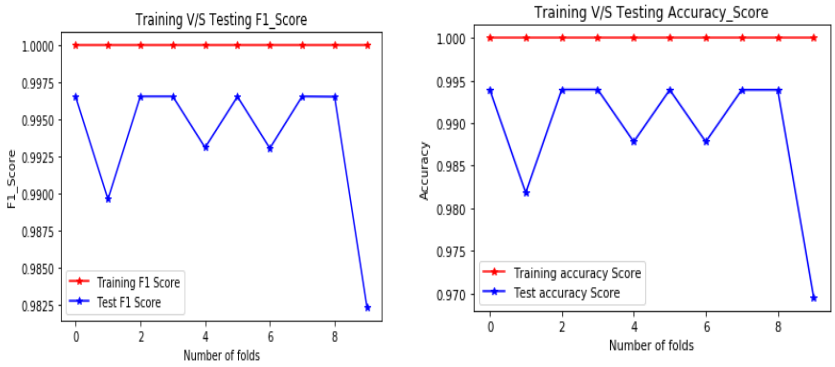
During training of the proposed cross-validated AdaBoost algorithm, training and testing score with respect to accuracy, f1-score, kappa-score and MSE is calculated for each fold. The scores obtained for this model during each fold is depicted in Fig2. In fig 2 (a), (b) and (d) higher values of training scores and lower values of testing scores are observed. Again in fig 2 (c), testing MSE values are greater than training dataset MSE values. The scores shown in Fig 2 for training and testing scores clearly indicate that the proposed model prevents itself from over-fitting. The obtained testing scores are collected for each fold and their

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2020 doi:10.20944/preprints202006.0351.v1

mean is calculated as the final testing score. These scores are shown in Table 1. A comparative study among all specified baseline classifiers is drawn with respect to specified metrics. As shown in Table 2, the proposed model achieves better result over other baseline classifiers such as Random Forest, Extra Trees, and Gradient Boost Classifier. In terms of all specified metrics, stratified cross-validated AdaBoost classifier provides the best predictive result. An accuracy of 98.97%, f1-score of 0.99, kappa statistics score of 0.95 and MSE of 0.0103 is indicated by the proposed model.

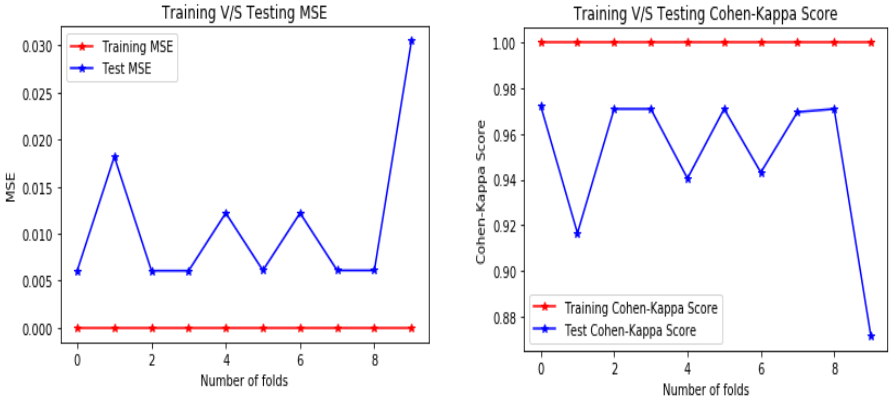
Performance Measure Metrics	Accuracy	F1-Score	Cohen-Kappa Score	MSE
Proposed Classifier Model				
Cross-Validated AdaBoost Classifier	98.97%	0.99	0.95	0.0103
Baseline Classifier Model				
Random Forest Classifier	98.18%	0.98	0.93	0.02
Extra Trees Classifier	94.33%	0.94	0.74	0.06
Gradient Boosting	90.69%	0.91	0.54	0.09

Table 3: Performance Summary of Ensemble Techniques



(a) (b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2020 doi:10.20944/preprints202006.0351.v1



(c) (d)

Fig2. Performance of Proposed Classifier during each fold

Conclusions-

Detection of presence of abnormalities in brain are very important for proper diagnosis of diseases. The proposed system can be further enhanced to classify the types of abnormalities and other types of tumours with few modifications. The aim of this paper is to detect feasibility of applying machine learning techniques for recognizing brain tumors of patients. In this paper, stratified cross-validated AdaBoost classifier is introduced in order to detect brain tumor of patients. A list of interfering factors is fed into the classifier model for examining patients with brain tumor. Applying pre-processing techniques to the brain tumor dataset and fine-tuning the hyper-parameters of the proposed model, maximized performance is attained in terms of classification. High accuracy and low classification error rate in classifier model's performance is favored in the field of brain tumor detection.

References-

[1] Ghosh M, Shubham S, Mandal K, Trivedi V, Chauhan R, Naseera S. "Survival and prognostic factors for glioblastoma multiforme: Retrospective single-institutional study". Indian J Cancer [serial online] 2017 [cited 2020 Jun 26]; 54: 362-7. Available from: http://www.indianjcancer.com/text.asp?2017/54/1/362/219544

[2] R. Maclin, "Popular Ensemble Methods : An Empirical Study Popular Ensemble Methods : An Empirical Study," vol. 11, no. July, pp. 169–198, 2016, doi: 10.1613/jair.614.

[3] R. E. Schapire, "Explaining adaboost," *Empir. Inference Festschrift Honor Vladimir N. Vapnik*, pp. 37–52, 2013, doi: 10.1007/978-3-642-41136-6_5.

[4] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[5] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[6] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[7] M. Al-Ayyoub, G. Husari, O. Darwish, and A. Alabed-Alaziz, “Machine learning approach for brain tumor detection,” *ACM Int. Conf. Proceeding Ser.*, no. April, 2012, doi: 10.1145/2222444.2222467.

[8] Han’guk Chŏngbo Kwahakhoe, IEEE Computer Society, and Institute of Electrical and Electronics Engineers, “The 32nd International Conference on Information Networking (ICOIN 2018) : January 10 (Wed.)-12 (Fri.), 2018, Holiday Inn Chiang Mai, Chiang Mai, Thailand,” pp. 473–475, 2018.

[9] A. Batra and G. Kaushik, “SECTUBIM : Automatic Segmentation And Classification of Tumeric Brain MRI Images using FHS (FCM , HWT and SVM),” vol. 7, no. 6, pp. 13190–13194, 2017.

[10] K. Abbas, P. W. Khan, K. T. Ahmed, and W. C. Song, “Automatic Brain Tumor Detection in Medical Imaging using Machine Learning,” *ICTC 2019 - 10th Int. Conf. ICT Converg. ICT Converg. Lead. Auton. Futur.*, pp. 531–536, 2019, doi: 10.1109/ICTC46691.2019.8939748.

[11] J. Batista and R. Kitney, “Extraction of tumours from MR images of the brain by texture and clustering,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 974, pp. 235–240, 1995, doi: 10.1007/3-540-60298-4_264.

[12] N. Behzadfar and H. Soltanian-Zadeh, *Automatic segmentation of brain tumors in magnetic resonance images*. 2012.

[13] G. Hemanth, M. Janardhan, and L. Sujihelen, “Design and implementing brain tumor detection using machine learning approach,” *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, vol. 2019-April, no. Icoei, pp. 1289–1294, 2019.

[14] J. Seetha and S. S. Raja, “Brain tumor classification using Convolutional Neural Networks,” *Biomed. Pharmacol. J.*, vol. 11, no. 3, pp. 1457–1461, 2018, doi: 10.13005/bpj/1511.

[15] R. H. Kirschen, E. A. O’Higgins, and R. T. Lee, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *Am. J. Orthod. Dentofac. Orthop.*, vol. 118, no. 4, pp. 456–461, 2000, doi: 10.1067/mod.2000.109032.

[16] Jakesh Bohaju, “Brain Tumor.” Kaggle, doi: 10.34740/KAGGLE/DSV/955413.

[17] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, “Assessing the accuracy of prediction algorithms for classification: An overview,” *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000, doi: 10.1093/bioinformatics/16.5.412.

[17] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification

 doi:10.20944/preprints202006.0351.v1

Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[18] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," *2010 IEEE World Congr. Comput. Intell. WCCI 2010*, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447.