IJMLNCE JOURNAL

International Journal of Machine Learning and Networked Collaborative Engineering

Journal Homepage: http://www.mlnce.net/home/index.html

DOI : https://doi.org/10.30991/IJMLNCE.2020v04i01.002

Revealing Brain Tumor Using Cross Validated NGBoost Classifier

[a]S. Dutta, [b]S. K. Bandyopadhyay[*],

[a]Faculty of Computer Science, The Bhawanipur Education Society College, Kolkata, India

[a]shawnidutta83@gmail.com

https://orcid.org/ 0000-0001-8557-0376

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

[b]1954samir@gmail.com

https://orcid.org/ 0000-0002-4868-3459

Abstract

Context: In the human body, one of the complicated and delicate anatomical structures is Brain. In most brain abnormalities, the brain tumor is most fatal, and it became carcinogenic in most of the cases. It is featured by abnormal and uncontrolled growth of brain cells. It can be benign or malignant.

Objective: It is challenging to detect a brain tumor, which is very crucial and challenging. Neurologist or Neurosurgeon needs to know the size and actual location of the tumor in the brain. It is required to know whether there is any swelling in the patient or compression of the brain. If it happens then, immediate attention is needed for the surgeon.

Method: Ensemble a strategy based Machine Learning (ML) algorithm that is exploited for revealing brain tumors. NGBoost algorithm is proposed to detect brain tumors of patients. Classifier is based on 5-fold stratified cross-validation method. It is compared with ensemble based other existing Classifiers.

Results: The proposed method outperforms baseline models with significantly improved efficiency. It is confirmed by the results obtained after the execution of the technique. The ranking is retrieved from the best classifier model based on interfering features. This 5-fold cross-validated NGBoost algorithm has reached an accuracy of 98.54%, and F1-score of 0.9917, the cohen-kappa score of 0.9302, and MSE of 0.0146.

Conclusion: An automated tool is approached in this paper for analyzing brain MRI image features, and the probability of brain tumor occurrences is detected. The predictive model has been designed in such a way it exhibits high accuracy and lowest error rate in terms of prediction results.

Keywords

Brain Tumor;

5-fold Cross-Validation;

NGBoost;

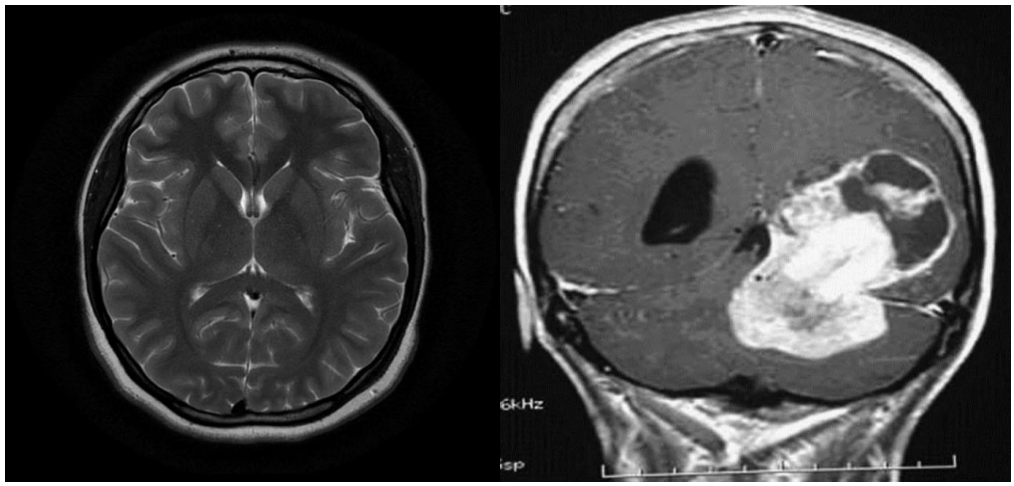
Ensemble Technique, Patient.

1. Introduction

In every sphere of human life ranging from communication, smart systems, and even medical diagnosis, the computer-aided technology plays a crucial role in the designing order for analysis of problems/new ideas in different areas. It is important today to analyze brain diseases from a dataset

[*]Corresponding author
SK.Bandyopadhyay
Email: 1954samir@gmail.com

containing features related to brain disease. The brain is one of the most complicated parts of the human body. For analysis of brain tumor, it is one of the difficult tasks for doctors to the diagnosis of it properly. The difficulty arises much more if cancer in the brain is carcinogenic, i.e., brain cancer. It is the actually uncontrolled growth of brain cells and creates a cranial cavity. The shape, size, position, and characteristics determine whether it is benign or malignant [1]. For better understanding, sample MRI images of brain tumor and without brain tumor is shown in figure.1.



(a) (b)

Figure 1. Brain MRI image (a) without tumor (b) with tumor

Using Machine Learning (ML) approaches, early prediction of any diseases can be performed accurately. This paper aims to predict whether a patient can have cysts in the brain or not. Artificial Intelligence (A.I.) is a superset of ML. A.I. is based on a learning system like a human being that enables automatic learning of operations without being explicitly programmed. The facility for self-learning of order by gathering knowledge from experience is the aim of ML. This paper exemplifies the use of ensemble-based ML techniques [2] to detect brain tumors. The brain tumor predictive model should be designed to achieve maximized accuracy with lower prediction error. Thus the ensemble-based predictive model is approached in this context. The ensemble-based techniques produce improved, accurate results over a single classifier model. Ensemble techniques are known to be meta-algorithms that assemble decisions from multiple base models into a single predictive model [2]. Boosting is a technique that produces an ensemble model that is implemented in this paper. A classifier is NGBoost, followed by 5-fold stratified cross-validation for the detection of brain abnormalities [3]. The comparison of the proposed method is made against other ensemble algorithms such as AdaBoost [4], Gradient Boost [5], Random Forest [6], and Extra Trees [7] classification techniques. All the classifier models are implemented by adjusting their parameters for achieving maximized efficiency. However, the classifier models are applied to the transformed dataset after applying necessary data cleaning operations.

2. Literature Survey

Computer-aided detection of Brain tumors, stroke lesions, hemorrhage lesions, and multiple sclerosis lesions are difficult for diagnosis. Computer-aided techniques can only help doctors to show the segmentation of abnormal tissues in brain MRI [8]. Usually, brain injuries, due to some reasons, distort typical tissue structures. Actually, in healthy tissues, the Intensity distribution is complicated, and also some overlaps occur between different types of tissues. Presently e-health care system based on information technology can make excellent support for early detection of health abnormalities of patients. The brain tumor is indeed a life-threatening disease for humans due to abnormal growth of tissues in the brain. Still, it is not known the exact reason for the occurrence of such abnormalities in the brain. It hampers proper brain function, so appropriate treatment is necessary at an early stage; otherwise, it is difficult to save the life of humans. A brain tumor can be as benign and malignant. A malignant tumor is fast developing and harmful than mild since it grows slowly and less harmful [9-12].

Early-stage cancer detection may not always be feasible, and early detection can prevent death. The tumor could be benign, pre-carcinoma, or malignant. If a patient is detected with a benign tumor, then it can be surgically removed since it does not spread to other organs and tissues [13]. The same is not valid for malign. Brain tumors are classified as gliomas, meningiomas, and pituitary tumors. Nerve cells and blood vessels do not create Gliomas, but brain tissues form it. These are generally seen in patients after tumor detection. If the tumor covers the brain and surrounding areas of the central nervous system, then the tumor is known as Meningiomas and, this type of tumor slow-going tumor[14]. Lumps created inside the skull are diagnosed as pituitary tumors. Gliomas are most commonly malignant, but meningiomas are typically benign. Some tumors with benign such as pituitary tumors can cause damage to other parts of a human. Magnetic resonance imaging (MRI) is the only basis of diagnosing a patient by doctor since the image shows the true picture of the tumor. It is difficult to decide whether the surgery of the tumor has to be made or not by the computer-aided technique. The entire decision depends on the radiologist [15]. Usually, the biopsy is done if there is any abnormality found in the MRI. The biopsy of the patient determines whether the tissue is benign or malignant. It indicates the future actions to be made for the patient.

In medical fields, A.I. and ML play an important supporting tool for the initial diagnosis of disease. ML can indeed be used to detect brain tumor subject to the final decision will be taken by radiologists. Methods based on ML using image databases have been designed and found in many research papers [10, 16-18]. Researchers carried out for classification of tumors on small databases [19][20][21]. Sixty-six images using Deep Neural Network (DNN) are applied to classify patients with or without tumor. The accuracy is 96.97% [22]. Other algorithms based on neural networks have been made on various medical databases [23].

Some researchers detected the tumor types based on the augmented tumor region. The accuracy of 91.28% is obtained using these methods [24]. It has also used Convolutional Neural Networks (CNN) to extract and classify features for the detection of the zone of the tumor. In CNN, pre-processing and feature extraction is not required since it is mostly found from the testing process [24].

3. Proposed System and Methodology

This paper approaches an automated tool for recognizing brain tumors in patients. A classifier model is used for this purpose that can associate the input variable into the output class while learning from training data. The learning procedure gained by the classifier is evaluated in terms of acquiring prediction. Input to the proposed classifier model is a new set of features for the prediction of disease.

The improved version of the Gradient Boost algorithm is Naturals Gradient Boost (NGBoost) [3]. It is a probabilistic prediction scheme, and it trains the base learner to output for each training example, such as probability distribution, for minimizing the excellent score. In Natural gradient (N.G.) descent, an optimization algorithm, the base learner is a collection of weak learners using the boosting approach. The NGBoost algorithm combines a multi-parameter boosting algorithm with the natural gradient to estimate the variation of parameters of the presumed outcome distribution with the observed features.

The proposed model has training dynamics, which is a multi-parameter NGBoost approach. The advantages of the classifier are flexible, scalable, and easy-to-use. It handles classification, regression, survival problems, etc. using the same software package and interface. It thus trivially extends to a variety of use cases, such as negative binomial boosting (for counts), Gamma, or Weibull boosting (for survival prediction, with or without right-censored data), etc. It scales to large numbers of features or observations with the same favorable complexity of traditional boosting algorithms [3]. The NGBoost classifier is implemented using a particular distribution, which is a discrete probability distribution. The k-dimensional categorical distribution applies to the k-way distribution event. It provides the probabilities of potential outcomes of a single withdraw rather than multiple drawings.

Decision Tree Regressor of maximum depth of 3 is the basic building block of the classifier. It learns with 500 base estimators with a learning rate of 0.01. The description of the implementation of the NGBoost classifier is shown in table 1.

Table 1: Description of NGBoost Classifier

Probability Distribution	Categorical Distribution
Base estimator	Decision Tree of max depth of 3

Number of Base estimator	500
Learning rate	0.01
Criterion	Friedman MSE

5-fold stratified cross-validation helps to extend the methodology. It is a resampling methodology that is used to estimate the skill of a model. In this method, the dataset is shuffled randomly, and then it is partitioned into k groups. In this case, we consider the value of k as 5. One partition is regarded as a test or holds outset, and rest partitions of the datasets are considered as a training dataset. It is done for each group. It is then fitted to the training dataset, and evaluation is processed on the testing dataset. Ensuring accuracy evaluation scores and mean score is calculated for the final result of each of these folds. It provides a stratified mechanism with proper cross-validation [25]. The exact workflow of the implemented method is given in Figure 2.

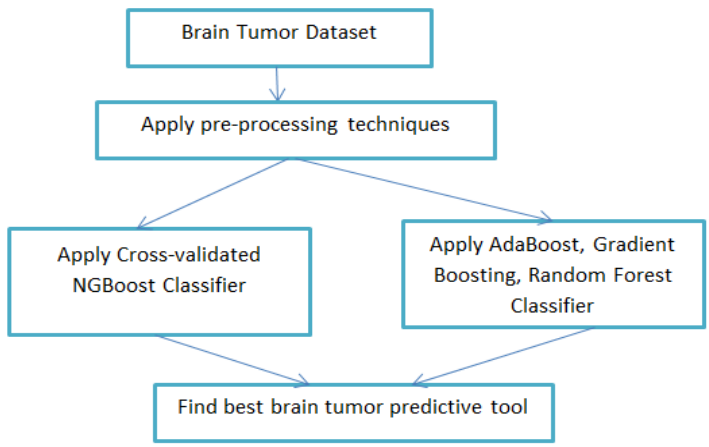


Figure2: System Workflow

3.1 Baseline Machine learning Classifiers

Ensemble technique based other classifiers are compared with the proposed. It justifies the performance of the method. It is used for the prediction of brain tumor of the patient.

AdaBoost is considered to be the first boosting technique proposed by Freund and Schapire [4]. This algorithm also belongs to the category of interpolating classifiers, which defines an algorithmic property of fitting the training data entirely without error. For providing a classifier on the original dataset, a meta-estimator is used first. Next, replicas of the classifiers are fitted for re-weighting the incorrectly classified instances for handling more complicated cases [4]. The presented classifier is also known as meta-estimator.

Another classifier is suitable for fitting new models for maximized efficiency by estimating the response variable. It is a Gradient Boost (G.B.) algorithms [5], used to construct new base learners to make maximally correlated with the negative gradient for the loss function. It provides freedom in model designing. Based on trial and error, this method finds a loss function [5].

The concept of the ensemble learning approach in Random forest (R.F.) [6] is exemplified to use the regression model and is a combination of several tree-like classifiers. This approach makes each tree cast in such a way that it votes to the most appropriate class for the input [6].

Some researchers use other trees classifier as the primary function for classification [7]. This classifier is used to form an ensemble regression tree. The top-down procedure is the basis of this classifier [7].

3.2 Implementation of Baseline models

Dataset fed into these baseline classifiers is initially partitioned into two classes, namely training and testing dataset with a ratio of 7:3. For extracting and learning of hidden patterns, the first set is the input to the classifier model. The learning evaluated testing dataset and prediction result is retrieved. The G.B.

classifier is implemented with 500 base estimators, a learning rate of 0.9. The R.F. classifier is also implemented with 500 base estimators, whereas; Extra Trees classifier is designed with 100 numbers of trees in the forest. These designed ensemble models with necessary tuning will assist in attaining the best results.

3.3 Dataset and Pre-processing

This paper collects Brain Tumor Dataset available at Kaggle [26]. 1644 number of patients records are in the dataset, and each has 18 attributes. Brain tumor image data used in this article were obtained from the MICCAI 2013 Challenge on Multimodal Brain Tumor Segmentation. The database consists of anonymous images retrieved from the Cancer Imaging Archive. From these images, features are extracted and summarized into one CSV file. This file is available at Kaggle [26]. The dataset includes five first-order features and eight texture features and four quality assessment parameters with the target level. These features are discriminating properties by which are analyzed by any ML algorithms for identifying brain tumors.

The presence of Infinite values and Not a Number (NaN) in the data set will not always impact the efficiency of the classification of the tumor. The missing values in the proposed method are handled by replacing them with the mean values of missing values. The exact numbers of missing values are shown in Table 2.

Table 2: Missing Values in Dataset

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

The different feature sets used here are: Mean, Variance, Standard Deviation, Skewness, and Kurtosis. Contrast, Energy, ASM (Angular second moment), Entropy, Homogeneity, Dissimilarity, Correlation, and Coarseness, Quality assessment parameters are PSNR (Peak signal-to-noise ratio), SSIM (Structured Similarity Index), MSE (Mean Square Error), and D.C. (Dice Coefficient). Figure 3 shows the distribution of patients with tumors or without tumors.

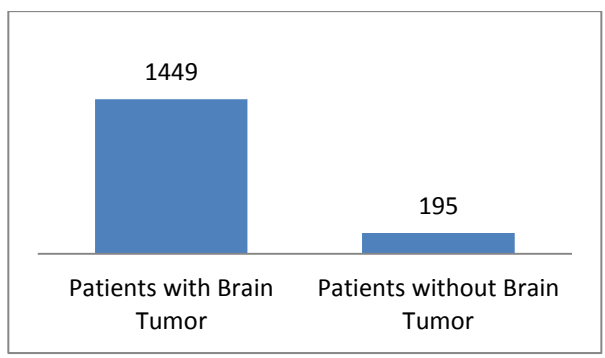


Figure 3. Distribution of brain tumor patients in the dataset

4. Experimental Results

The performance measures based on conventional parameters are calculated for the proposed method [27-28]. Mean Squared Error (MSE) [27] is also calculated. During the training of the proposed cross-validated NGBoost algorithm, the training and testing process is evaluated against all the parameters described in the previous section. In the proposed method, 5-fold cross-validation is used. The accumulated

scores for each fold for testing and training are shown in figure 4. Accuracy, f1-score, Kappa score, and MSE for each fold are shown in Figure 4 (a), (b), (c), and (d), respectively.

In figure4 (a), (b) and (d) higher values of training scores and lower values of testing scores are observed. Again in figure 4 (c), testing MSE values are more significant than training dataset MSE values. The scores are shown in Figure 4 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 mean is calculated as the final testing score. The model described in the paper shows better results than all other classifiers. The results of all performance matrixes are shown in table 3. Features obtained for predictive analysis are shown in figure 5. The proposed model, along with baseline classifier models, are implemented in Python with scikit-learn package [29].

Table 3: Performance Summary of Ensemble Techniques

Performance Evaluating Metrics	Accuracy	F1-Score	Cohen-Kappa Score	MSE
Proposed Model				
Cross-Validated NGBoost Classifier	98.54%	0.9917	0.9302	0.0146
Baseline Model				
Gradient Boost	97.37%	0.97	0.89	0.03
AdaBoost	98.18%	0.982	0.92	0.02
Random Forest	97.98%	0.98	0.92	0.02
Extra Trees	94.13%	0.94	0.72	0.06

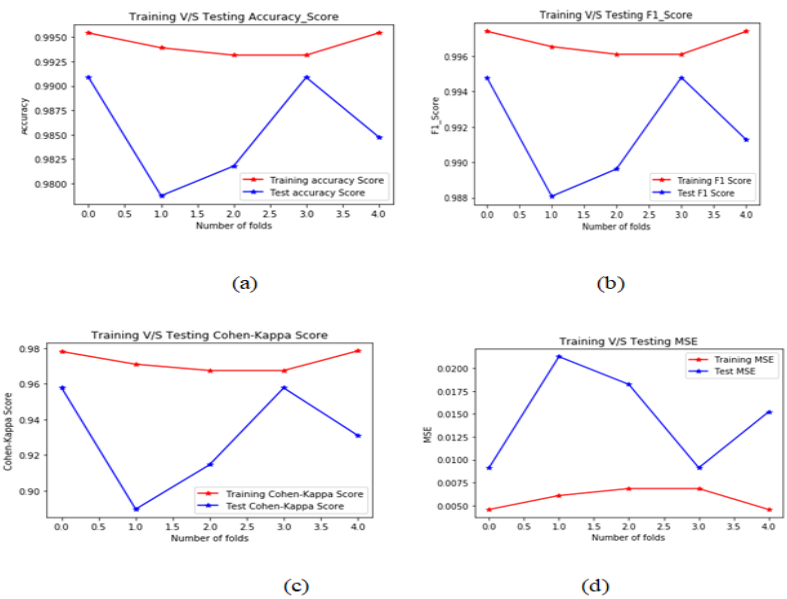


Figure 4. The score obtained during each fold of the proposed classifier

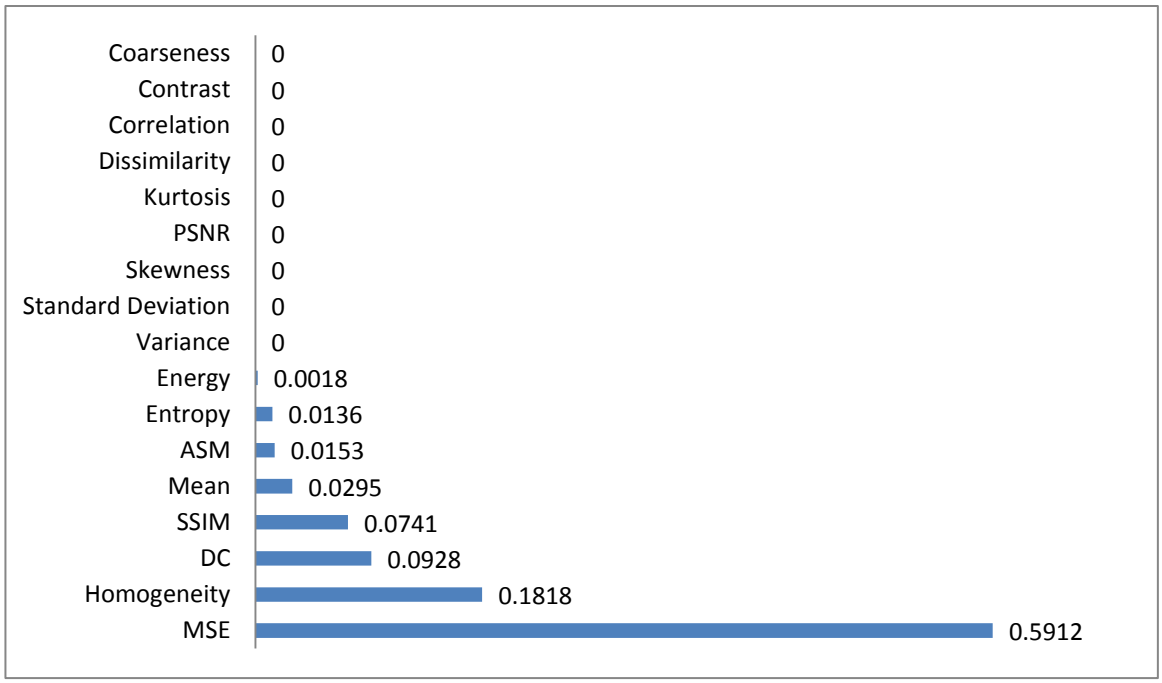


Figure 5.Feature Importance Ranking obtained from Cross-Validated NGBoost Classifier

5. Conclusion and Enhancement in the Future

The use of ensemble ML techniques was utilized in this paper that identifies patients with brain abnormalities. This research has been carried out to point out the feasibility of using ML techniques that predicts brain tumors with promising efficiency and the lowest error rate. The results are encouraging and far better than other classifiers. The proposed classifier model detects brain tumors of those who are suffering from this disease. It is to be mentioned here that the final decision is based on the radiologist. The results in terms of performance matrices are better than other classifiers. However, this research can even be extended by detecting exact types of brain tumors while analyzing past medical records of patients. In other words, the impact of other diseases and chronic effects on brain tumor occurrence can also provide insight into this field.

References

[1]. W. B. Pope and M. W. Itagaki, "Characterizing Brain Tumor Research: The Role of the National Institutes of Health SUMMARY," pp. 605–609, 2010, doi: 10.3174/ajnr.A1904.

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

[3]. T. Duan et al., "NGBoost: Natural Gradient Boosting for Probabilistic Prediction," 2019. arXiv:1910.03225, doi:10.1007/978-3-642-41136-6_5

[4]. 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.

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

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

[7]. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees Extremely randomized trees," no. January 2014, 2006, doi: 10.1007/s10994-006-6226-1.

[8]. J. N. Rich et al., "Statement brain tumours," Nat. Rev. Clin. Oncol., vol. 16, no. August, 2019, doi: 10.1038/s41571-019-0177-5.

[9]. X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. U. Zhang, "A Survey on Deep Learning based Brain-Computer Interface: Recent Advances and New Frontiers," vol. 1, no. 1, 2018, doi: 10.1145/1122445.1122456.

[10]. J. Amin, M. Sharif, M. Yasmin, and S. L. Fernandes, "Big data analysis for brain tumor detection : Deep convolutional neural networks," Futur. Gener. Comput. Syst., 2018, doi: 10.1016/j.future.2018.04.065.

[11]. J. Juan-albarracín, E. Fuster-garcia, and J. V Manjón, "Automated Glioblastoma Segmentation Based on a Multiparametric Structured Unsupervised Classification," pp. 1–20, 2015, doi: 10.1371/journal.pone.0125143.

[12]. M. Soltaninejad et al., "Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI," 2016, doi: 10.1007/s11548-016-1483-3.

[13]. V. V. Priya, "An Efficient Segmentation Approach for Brain Tumor Detection in MRI," vol. 9, no. May, 2016, doi: 10.17485/ijst/2016/v9i19/90440.

[14]. D. N. Louis et al., "The 2016 World Health Organization Classification of Tumors of the Central Nervous System : a summary," ActaNeuropathol., vol. 131, no. 6, pp. 803–820, 2016, doi: 10.1007/s00401-016-1545-1.

[15]. P. Afshar, K. N. Plataniotis, and A. Mohammadi, "CAPSULE NETWORKS FOR BRAIN TUMOR CLASSIFICATION BASED ON MRI IMAGES AND COARSE TUMOR BOUNDARIES," pp. 1368–1372, 2019. arXiv:1811.00597,DOI: 10.1109/ICASSP.2019.8683759, doi:10.1007/978-3-642-41136-6_5Corpus ID: 7122892

[16]. J. Amin, M. Sharif, M. Raza, and M. Yasmin, "Detection of Brain Tumor based on Features Fusion and Machine Learning," J. Ambient Intell. Humaniz. Comput. 2018, doi: 10.1007/s12652-018-1092-9.

[17]. K. Usman and K. Rajpoot, "Brain tumor classification from multi-modality MRI using wavelets and machine learning," Pattern Anal. Appl., 2017, doi: 10.1007/s10044-017-0597-8.

[18]. P. Mlynarski, A. Criminisi, and N. Ayache, "Deep Learning with Mixed Supervision for Brain Tumor," pp. 1–23. 10.1117/1.JMI.6.3.034002, DOI:10.1117/1.JMI.6.3.034002

[19]. S. Vijh, S. Sharma, and P. Gaurav, "Brain Tumor Segmentation Using OTSU Embedded Adaptive Particle Swarm Optimization Method and Convolutional Neural Network." Springer International Publishing. doi.org/10.1007/978-3-030-25797-2_8

[20]. H. Mohsen, E. A. El-dahshan, E. M. El-horbaty, and A. M. Salem, "Classification using deep learning neural networks for brain tumors," Futur. Comput. Informatics J., vol. 3, no. 1, pp. 68–71, 2018, doi: 10.1016/j.fcij.2017.12.001.

[21]. A. Veeraraghavan, S. Member, and A. K. Roy-chowdhury, "Matching Shape Sequences in Video with Applications in Human Movement Analysis," no. January, 2006, doi: 10.1109/TPAMI.2005.246.

[22]. G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," no. 1995, 1998. 10.1016/j.media.2017.07.005, doi:10.1016/j.media.2017.07.005.

[23]. Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep Learning for Brain MRI Segmentation : State of the Art and Future Directions," pp. 449–459, 2017, doi: 10.1007/s10278-017-9983-4.

[24]. J. Cheng et al., "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition," no. October, 2015, doi: 10.1371/journal.pone.0140381.

[25]. 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.

[26]. JakeshBohaju, "Brain Tumor." Kaggle, doi: 10.34740/KAGGLE/DSV/955413.

[27]. H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[28]. S. M. Vieira, U. Kaymak, and J. M. C. Sousa, "Cohen's kappa coeffcient 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.

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

[30]. B. H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," in IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993-2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

Author's Biography

Prof. Shawni Dutta is associated with the Bhawanipur Education Society College, Kolkata, India. She is teaching in the Department of Computer Science. She has published a large number of papers in Computer Science and Engineering in Indian and International Journals. Her research areas include Machine Learning, Medical Data Analysis, Malaria, and Recommended System. She published a book on "COVID-19- Validation of Disease by Machine Learning Approach", LAP, Lambert Academic Publishing, Germany. She delivered Key Note speech in "Growth of COVID-19 in the World using Machine Learning Approach," Virtual Summit on Saturday, July 11th, 2020, organized by Sciinov Group, Europe.

Prof. (Dr.) Samir Kumar Bandyopadhyay is presently associated with The Bhawanipur Education Society College, Kolkata, India, as Academic Advisor and Professor of the University of Calcutta. He did B.E., M. Tech., MBBS (Cal), M.D. (CAL), MRCP (U.K.), FRCS (U.K.), Ph.D. (Cal), Ph.D. (UCL, U.K.), Ph.D. (MIT, USA). He did his M.D. in Radiology. He is a visiting Professor and Resource Person of different Universities and institutes of South Korea, USA, Australia, Thailand, Singapore, and Sri Lanka. He published papers of about 800 and supervised a Ph.D. of 100 Scholars. He published 20 books in Computer Science and Engineering. He was the former Vice-Chancellor of WBUT, State University, India.