https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*

A Genetic Algorithm Based Feature Selection for Classification of Brain MRI Scan Images Using Random Forest Classifier

Dr. S. Mary Joans[1], J. Sandhiya[2]

[1]Professor and Head, Department of ECE, Velammal Engineering College, Chennai, Tamil Nadu, India
[2]M.E applied Electronics, Department of ECE, Velammal Engineering College, Chennai, Tamil Nadu, India

Abstract— *A brain tumour is a mass of tissue that is formed by a gradual addition of anomalous cells and it is important to classify brain tumours from the magnetic resonance imaging (MRI) for treatment. Magnetic Resonance Imaging is a useful imaging technique that is widely used by physicians to investigate different pathologies. After a long clinical research, it is proved to be harmless. Improvement in computing power has introduced Computer Aided Diagnosis (CAD) which can efficiently work in an automated environment. Diagnosis or classification accuracy of such a CAD system is associated with the selection of features. This paper proposes an enhanced brain MRI image classifier targeting two main objectives, the first is to achieve maximum classification accuracy and second is to minimize the number of features for classification. Feature selection is performed using Genetic Algorithm (GA) while classifiers used are Random forest Classifier.*

Keywords— Feature selection, Brain MRI, Genetic algorithm, Classifier.

I. INTRODUCTION

A brain tumour is an intracranial solid neoplasm which is characterised as an abnormal growth of cells within the brain or the central spinal canal. It is important to find out tumour from MRI images but it is somewhat time-consuming and difficult task sometimes when performed manually by medical experts. Huge amount of time was spent by radiologist and doctors for detection of tumour and classifying it from other brain tissues. However, exact labelling of brain tumours is a time-consuming task, and considerable variation is observed between doctors. Subsequently, over the recent decade, from various research results it is being observed that it is very time consuming method but it will get faster if we use image processing techniques. Primary brain tumours do not multiply to other body parts and can be malignant or benign and secondary brain tumours are all the time malignant. Malignant tumour is more risky and life threatening than benign tumour. The benign tumour is easier to identify than the malignant tumour. Also the first stage tumour may be malignant of benign but after first stage it will change to dangerous malignant tumour which is life frightening.

Different brain tumour detection algorithms have been developed over the last few years. Normally, the automatic classification problem is very tough and it is yet to be fully and satisfactorily solved. The main aim of this system is to make an automated system for detecting and identifying the tumour from normal MRI.

Magnetic resonance is at present one of the basic and most widely used techniques in medicine. The method commonly finds application in identifying suspected pathologies of the brain; in this case, the different signal intensities characteristic of the healthy and pathological tissues enable us to localize probable tumours.

The MRI brain tumour image classification is becoming increasingly essential in the medical arena. Physical classification of magnetic resonance (MR) brain tumour images is a difficult and time-consuming task. Medical Image Processing has developed to detect as well as analyse various disorders. The medical images facts are acquired from Bio-medical imaging procedures like Computed Tomography scan, Magnetic Resonance Imaging scan and Mammogram scan, which indicates the presence or absence of the lesion. The most vital challenge in brain MRI analysis has problems such as noise, intensity non-uniformity (INU), partial volume effect, shape complexity and natural tissue intensity variations. Inclusion of a priori medical knowledge is necessary for robust and exact analysis under such conditions.

The classification of MRI brain image data as normal and abnormal are vital to analysis for the normal patient and to consider only those who have the chance of having abnormalities. Diagnosis of abnormalities can be done automatic with more accuracy in feature selection and classification of disease. The supervised learning technique such as Random forest classifier is used for classification as it gives better accuracy and performance than other classifiers.

https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*

II. RELATED WORK

The MR human brain images are classified into its explicit category using supervised techniques like artificial neural networks, support vector machine, and unsupervised techniques like self-organization map (SOM), fuzzy c-means, via the feature set as discrimination function. Other supervised classification techniques, such as k-nearest neighbours (k-NN) also cluster pixels based on their similarities in every feature [3]. Classification of MR images both as normal or abnormal can be done using both supervised and unsupervised techniques [2].

Komal et al., [2] proposed an automation system, that performs binary classification to detect the presence of brain tumour. The dataset constitutes 212 brain MR images. It takes MR scan brain images as input, performs pre-processing, extracts texture features from processed image and classification is performed using machine learning algorithms such as Multi-Layer Perceptron (MLP) and Naive Bayes. It has been concluded with an accuracy of 93.6% and 91.6% respectively.

Namitha Agarwal et al., [4] proposed a technique where first and second order statistical features are used for classification of images. In this paper, investigations have been performed to evaluate texture based features and wavelet-based features with commonly used classifiers for the classification of Alzheimer's disease based on T2-weighted MRI brain image. It has been concluded that the first and second order statistical features are considerably better than wavelet based features in terms of all performance measures such as sensitivity, specificity, accuracy, training and testing time of classifiers.

Joshi et al., [17] proposed the brain tumour recognition and its type classification system using MR images. From the images, the lesion region is segmented and then texture features of that section are extracted using Gray Level Co-occurrence Matrix (GLCM) like energy, contrast, correlation and homogeneity [4]. For classification, neuro-fuzzy classifier is used. Gladis Pushpa et al., [19] proposed a technique that combines the intensity, texture and shape based features and classifies the lesion region as white matter, CSF, Gray matter, normal and abnormal area using SVM. Linear Discriminant Analysis (LDA) and Principle Component Analysis (PCA) are used to ease the number of features in classification.

Evangelia et al., [13] performed a binary classification to explore the use of pattern classification methods for distinguishing primary gliomas from metastases, and high grade tumour (type3 and type4) from low grade (type2). This scheme has a series of steps including ROI definition, feature extraction, feature selection and classification. The extracted features consist of tumour shape and intensity characteristics as well as rotation invariant texture features. Feature subset selection is performed using Support Vector Machines (SVMs) with recursive feature elimination.

In our research work, the classification method has been used to classify brain tumour using different levels of statistical feature extraction methods. For classification, the supervised machine learning algorithm–Random forest classifier has been employed. From the analysis, the suitable feature set that discriminates a tumour with improved performance has been identified. Accuracy, specificity and sensitivity measures have been used to analyze the result of tumour region.

Classification is an automated process that intends to order every information/ data/instance in specific class, in light of the data portrayed by its features. However, without previous knowledge, useful features cannot be determined for classification. So initially it requires an introduction of large number of features for classification of a particular dataset. Introducing a large number of features may include irrelevant and redundant features which are not helpful for classification and this can even lessen the performance of a classifier due to large search space known as "the curse of dimensionality" [2]. This problem can be subsided by selecting just relevant features for grouping. By omitting irrelevant and unnecessary features, feature selection reduces the training time and minimizing the feature set, thus improving the performance of classifier [3, 4]. During the analysis of tissue in MRI by radiologists, image texture plays a pre-dominant role. In fact texture (in) homogeneity is one of the most common individual MRI features used for tumour diagnosis [5]. Studies have shown that texture information can improve accuracy of classification and produce comparable/preferable results to radiologists when used for machine classification of MRI tissue [6]. Different families of texture calculation methods are being used for MRI analysis and it has been shown that combination of texture features from different families can lead to better classification performance [7].

III. METHODS AND METHODOLOGY

The main purpose of this paper is to identify the tumour and to do the detailed diagnosis of that tumour which will be used in treating the cancer patient. The detailed description about the proposed system is given below.

A. Pre-processing

In the image processing the gray scale image is processed by using different methods like brightness, Filtering and thresholding. Brightness makes the image by which white objects are distinguished from gray and light items from dark objects. Hence by changing the brightness of the image the tumour recognition in the MRI image is easier.

https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*

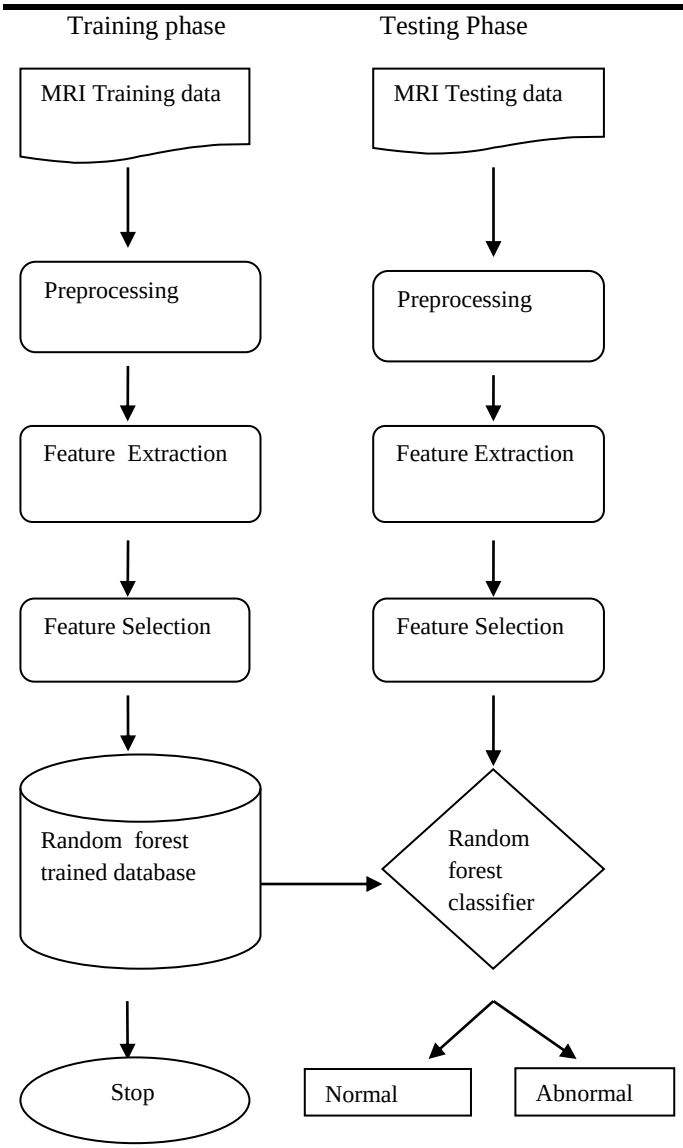


Fig.1: Flow diagram of proposed method

Filters can smooth, sharpen, transform, and remove unwanted noise from an image so that we can pull out the information needed to sharpen edges, counting the edges of any holes inside a particle, and create contrast between the elements and the background.

Preprocessing of MR brain image is the first step in our projected technique. Preprocessing includes image filtering for further processing. Our goal behind performing preprocessing is improvement of the image quality to get more surety and ease in spotting the lesion. Steps involved for preprocessing are as follows:

1) Input Image is converted to gray scale.

2) A 3x3 median filter is applied on MR brain image in order to eliminate the noise.

Median filter must only change the intensity of corrupted pixels on the damaged image in order to protect the local details of the image it has to be done. However, for fixed-valued impulse noise (i.e. salt-and-pepper noise) it is very difficult to detect the tainted pixels from the image correctly.

B. Feature Extraction Using GLCM

The Gray Level Coocurrence Matrix (GLCM) method is a way of extracting second order statistical texture features.

A GLCM is a matrix in which the number of rows and columns is equal to the number of gray levels, G, in the image. The matrix element P (i, j | Δx, Δy) is the relative frequency with the two pixels, separated by a pixel distance (Δx, Δy), occur within the given neighborhood, one with intensity 'i' and the other with intensity 'j'. The matrix element P (i, j | d, θ) contains the second order statistical probability values for changes between gray levels 'i' and 'j' at a specific displacement distance d and at a particular angle (θ). Using a huge number of intensity levels G implies storing a lot of impermanent data, i.e. a G × G matrix for every combination of (Δx, Δy) or (d, θ). Because to the high dimensionality, the GLCM matrix are extremely sensitive to the size of the texture samples on which they are approximate. Therefore, the number of gray levels is often reduced.

Gray Level Co-Occurrence Matrix (GLCM) has proved to be an accepted statistical method of extracting texture features from images. According to the gray level co-occurrence matrix, Haralick contains thirteen texture features measured from the probability matrix to extract the quality of texture statistics of brain MR images. Some of the important features are, Angular Second Moment or energy, Entropy, Correlation, and the Inverse Difference Moment.

1. ***Angular Second Moment***

Angular Second Moment is also called as Uniformity or Energy. It is the sum of the squares of values in the GLCM. Angular Second Moment measures the image homogeneity. Angular Second Moment has high value when image has very good homogeneity / when pixels are very similar.

$$\text{ASM} = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} Pij^2$$

Where i, j are the spatial coordinates of the function p (i, j), Ng is the gray tone.

2. Entropy

Entropy is the amount of information of the image that is needed for the image compression. Entropy process the loss of information or message in a transmitted signal and also measures the image information.

$$\text{ENTROPY} = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} -P_{ij} * \log P_{ij}$$

3. ***Correlation***

Correlation measures the linear dependency of grey levels of neighbourhood pixels. This is often used to calculate strain, deformation, strain, displacement and optical flow,

https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*

but it is widely applied in many areas of science and engineering. One of the very common application is for measuring the activity of an optical mouse.

Correlation = $\frac{\sum_{i=0}^{Ng-1}\sum_{j=0}^{Ng-1}(i,j)P(i,j)-\mu_x\,\mu_y}{\sigma_x\sigma_y}$

4. *Inverse Difference Moment*

Inverse Difference Moment (IDM) is called as the local homogeneity. It is high when local gray level value is uniform and inverse of GLCM is high.

$$\text{IDM} = \frac{\sum_{i=0}^{Ng-1}\sum_{j=0}^{Ng-1}P_{ij}}{1+(i-j)^2}$$

IDM weight value is the inverse of the Contrast weight.

Co occurrence matrix is often formed by using two offsets i-e distance (d = 1, 2, 3…) and angle (h = 0, 45, 90, and 135).

To avoid direction dependency, Haralik also suggested using the angular mean $M_T(d)$, variance $V^2_T(d)$ and range $R_T(d)$ of GLCM.

$M_T(d) = \frac{1}{N_\emptyset}\sum_\emptyset T(d,\emptyset)$

$R_T(d)$ = Max[T(d,∅)]-Min[T(d,∅)]

$V_T^2(d) = \frac{1}{N_\emptyset}\sum_\emptyset[T(d,\emptyset) - M_T(d)^2$

Where $N_\emptyset$ represents the number of angular measurements and T(d,∅) are scalar texture measures.

C. Feature Selection:

The main assumption when using feature selection is that there are a lot of redundant or irrelevant features which sometimes reduces the classification accuracy [13]. Features are evaluated using a fitness function, thus selecting the best rated features among the feature set. A feature selection is an operator fs which maps from the m dimensional (input) space to n dimensional (output) space given in mapping.

$$f_s : R^{r*m} \rightarrow R^{r*n}$$

Where m ≥ n and m, n ∈ Z^+, R^{r*m} matrix containing original feature set having r instances; R^{r*n} is a reduced feature set containing r instances in subset selection.

1. *Discrete Binary Genetic Algorithm*:

Genetic algorithms (GA) is the general adaptive optimization search methodology based on the direct correspondence to Darwinian's natural selection and genetics in biological systems. It has been proved to be a promising alternative to conventional heuristic techniques. It is based on the Darwinian standard of 'survival of the fittest', Genetic Algorithm works with a set of candidate solutions called a population and obtains the optimal solution after a series of iterative computations.

GA evaluates each individual's fitness, i.e. quality of the solution, through a fitness function. The fitter chromosomes have higher probability to be kept in the next generation or to be selected into the recombination pool using the tournament selection techniques. If the the chromosome or the fittest individual in a population cannot meet the requirement, successive populations will be reproduced to provide more alternating solutions. The crossover and mutation functions are the main operators that randomly convert the chromosomes and finally impact their fitness value. The evolution will not stop until acceptable results are obtained. Associated with the characteristics of exploitation and exploration search, Genetic Algorithm can deal with large search spaces efficiently, and hence has fewer chance to get local optimal solution than other algorithms.

GA is a heuristic process of natural selection which is inspired from the procedures of evolution in nature. This algorithm uses the Darwin's theory "Survival of fittest" motivated by inheritance, mutation, selection, and crossover. In comparative terminology to human genetics, gene represent feature, chromosome are bit strings and allele is the feature value [14]. From algorithm perspective, population of individuals represented by chromosomes are the arrangement of binary strings in which each bit (gene) represents a specific feature within a Chromosome (bit strings). Chromosomes are evaluated using Objective function (fitness function) which ranks individual chrome by its numerical value (fitness) within a population.

2. Process of GA:

Step 1 (Generation begins): A random population (a_{11} a_{12}..... a_{mn}) matrix p of size n x m is generated using population size n and number of features m.

Step 2 (Tournament): This phase selects the best-fit individuals for reproduction. Two chromosomes (parent chromes) with the highest fitness will take part in cross over.

Step 3 (Cross Over): Analogous to biological crossover, it is the exchange of bits within the selected parents to produce offspring. Number of bits b selected from parent P_n computed using , where parameter: 0<k<1 is a crossover probability.

$$b = K * P_n$$

Step 4 (Mutation): Mutation refers to the change (growth) in the genome of chromosome, flipping of bit strings (genes) of chromosome.

Step 5 (Fitness Evaluation): Analogous to "survival of fittest", chromosomes with a certain level of fitness will survive for next generation while the others whose fitness is less than the threshold value will be discarded.

https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*

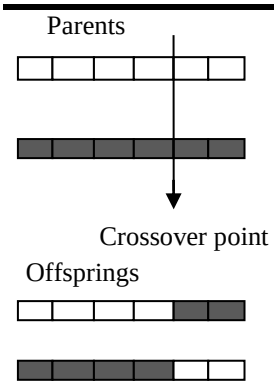


Fig.2: Illustration of crossover

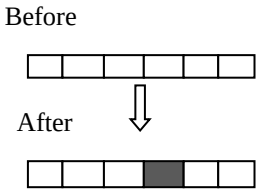


Fig.3: Illustration of Mutation.

D. Classification

Classification is the method of identification, discrimination of objects or patterns on the basis of their attributes. It is done using supervised learning. In this type of machine learning, machine classifies objects on the basis of previous knowledge. The system is trained using some attributes (features) along with their label, these attributes are used by classifier to guess the unknown objects.

1.Random Forest Classifier:

The random forest classification is an ensemble method that can be thought of as a form of nearest neighbour predictor.

The random forest starts with a standard machine learning technique called a "decision tree" which, in ensemble terms, corresponds to a weak learner. In a decision tree, an input is given at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets.

For some number of trees *T*, the system is trained as,

1. Sample *N* cases at random with replacement to generate a subset of the data . The subset must be about 66% of the total set.

2. At each node, For a few number *m*, predictor variables are selected at random from all the predictor variables. The predictor variable that gives the best split, according to some objective function, is used to do a binary split on that node.

3. At the subsequent or next node, choose other m variables at random from all predictor variables and repeat the same.

When a new input is entered into the system, it runs down all of the trees. The result may either be an average or the weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

IV. EXPERIMENTAL RESULTS

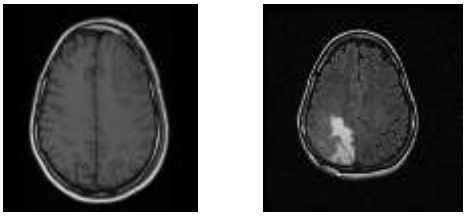
***A. Dataset*:**

Project is carried out on Intel(R), Core(TM) i5-4530 s CPU @ 2.30 GHz, with 4.00 GB of RAM. MATLAB 8.1.0 (R2013a) is used for simulation. A set of 20 images of size (256 _ 256) with a format of Portable Network Graphics are taken from Harvard Medical school website http://www.med.harvard.edu/AANLIB/ among which 10 are normal and remaining 10 are abnormal scans. The abnormal scans consist of three diseases viz. glioma, visual agnosia and meningioma.

The dataset of axial Magnetic Resonance Imaging (MRI), are collected from the subjects of various brain tumour types.

The brain tumour types considered in our system are Normal and Abnormal as shown in figure.

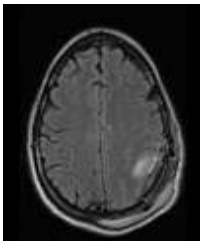
The images collected from different patients are grouped into two sets for using it during training and testing stages of the system.



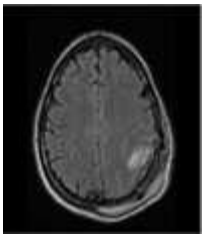
(a) (b)

Fig.4. (a) Normal image (b) Abnormal image

B. Performance and Analysis of Proposed System

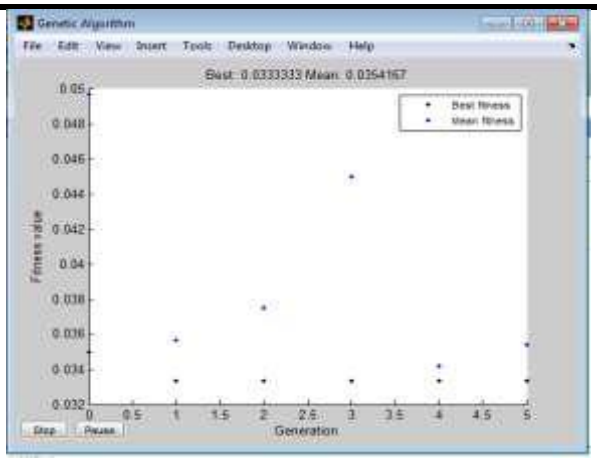


(a)

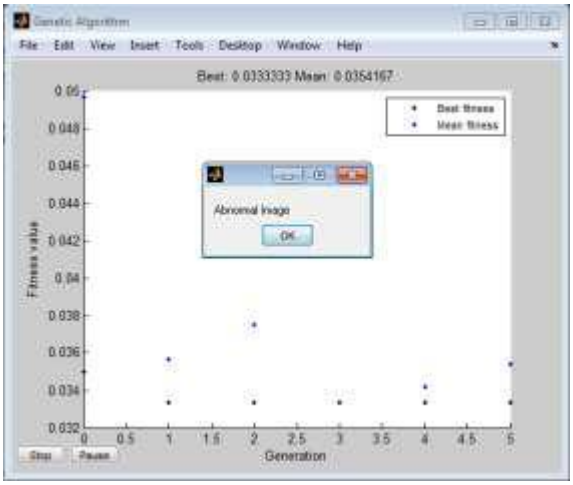


(b)

https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*



(c)



(d)

Fig.5(a): Input Image (b). Noise Removal (c).Feature Selection by Genetic Algorithm (d). Classifier Output

V. CONCLUSION AND DISCUSSION

Image processing has become a very important task in today's world. Today applications of image processing can be initiated in number of areas like medical, remote sensing, electronics and so on. If we focus on medical applications, image classification is widely used for diagnosis purpose.

Combining feature extraction and feature selection for classification has enhanced the accuracy of the classifier. Thus feature selection process has also enhanced the classification results. It is clearly seen, that without feature selection the classifier performance is weak, when compared with classification after feature selection.

Random Forest classifier has improved classification accuracy using least number of features.

The proposed work can be extended to classify different abnormalities in brain, such as, Alzheimer's disease, visual agnosia, Glioma with tumour, Herpes encephalitis with a tumour, bronchogenic carcinoma and Multiple sceloris with a tumour by reducing computational cost and further increase in mean-accuracy for the classification of Human Brain MRI scans.

Table.2: Performance of KNN classification

KNN Classification			
Image	**Accuracy**	**Specificity**	**Senstivity**
Normal and Abnormal	93.34	90	95

Table.2: Performance of proposed system.

Random Forest Classification			
Image	**Accuracy**	**Specificity**	**Sensitivity**
Normal image	99	90	98
Abnormal image	90	95	98
Total	95	92	98

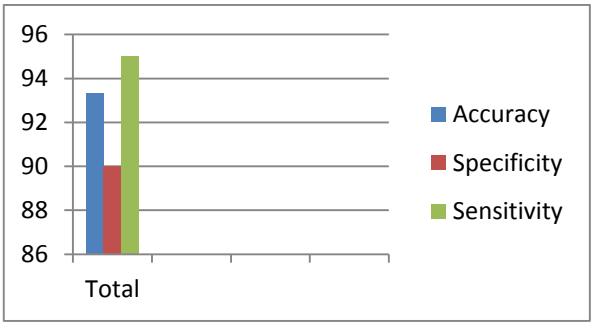


Fig.6: Performance of KNN classifier

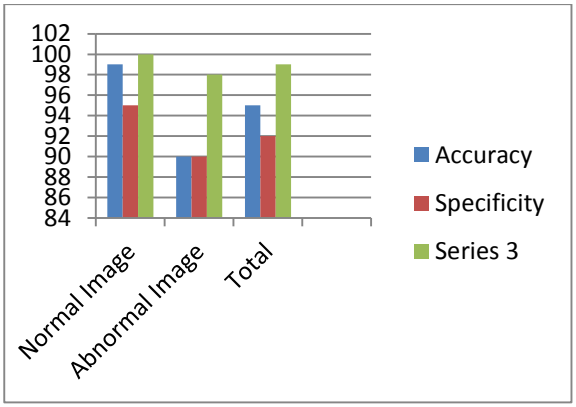


Fig.7: Performance of proposed system

REFERENCES

[1] Atiq ur, R. et al.: Hybrid feature selection and tumor identification in brain MRI using swarm intelligence. In: 11th International Conference on Frontiers of Information Technology (2013).

[2] Komal Sharma, Akwinder Kaura and Shruti Gujral, 2014, "Brain Tumor Detection based on Machine Learning Algorithms", International Journal of Computer Applications, vol. 103, no.1, pp. 7-11.

https://dx.doi.org/10.22161/ijaers.4.5.20 *ISSN: 2349-6495(P) | 2456-1908(O)*

[3] Walaa Hussein Ibrahim, Ahmed Abdel Rhman Ahmed Osman and Yusra Ibrahim Mohamed, 2013, "MRI Brain Image Classification using Neural Networks" ,IEEE International Conference On Computing, Electrical and Electronics Engineering, ICCEEE, pp. 253-258.

[4] Namita Aggarwal and Agrawal R K, 2012, "First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images", Journal of Signal and Information Processing, vol. 3, no. 2, pp. 146-153.

[5] Unler, A., Murat, A.: A discrete particle swarm optimization method for feature selection in binary classification problems. Eur. J. Oper. Res. 206(3), 528–539 (2010).

[6] De Schepper, A., Vanhoenacker, F., Parizel, P., Gielen, J. (eds.): Imaging of Soft Tissue Tumors, 3rd edn. Springer, Berlin (2005).

[7] Atiq ur, R. et al.: Hybrid feature selection and tumor identification in brain MRI using swarm intelligence. In: 11th International Conference on Frontiers of Information Technology (2013).

[8] García, M.A., Puig, D.: Improving Texture Pattern Recognition by Integration of Multiple Texture Feature Extraction Methods, pp. 7–10 (2002).

[9] Sidra et al.: Improved tissue segmentation algorithm using modified Gustafson-kessel Clustering for brain MRI (2014).

[10] Sivanandam, S.N., Deepa, S.N.: Introduction to Genetic Algorithms. Springer, Heidelberg (2008).

[11] Yao, M.: Research on learning evidence improvement for KNN based classification algorithm. Int. J. Database Theory Appl. 7(1), 103–110 (2014).

[12] Juntu, J., De Schepper, A.M., Van Dyck, P., VanDyck, D., Gielen, J., Parizel, P.M., Sijbers, J.: Classification of Soft Tissue Tumors By Machine Learning Algorithms (2011).

[13] Evangelia I. Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R. Melhem and Christos Davatzikos, 2009, "Classification of Brain Tumor Type and Grade using MRI Texture and Shape in a Machine Learning Scheme", Magnetic Resonance in Medicine, vol. 62, no. 6, pp. 1609–1618.

[14] Fritz et al.: Statistical Texture Measures Computed from Gray Level Concurrence Matrices, 5 November 2008.

[15] Prasad, B.: Speech, Audio, Image and Biomedical Signal Processing Using Neural Networks. Springer, Berlin (2008). 356 p.

[16] Ferńandez-Delgado, M., et al.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15, 3133–3181 (2014).

[17] Dipali M Joshi, Rana N K and Misra V M, 2010, "Classification of Brain Cancer Using Artificial Neural Network", IEEE International Conference on Electronic Computer Technology , ICECT, pp. 112-116.

[18] Sidra et al.: Improved tissue segmentation algorithm using modified Gustafson-kessel Clustering for brain MRI (2014).

[19] Gladis Pushpa Rathi V P and Palani S, 2012, "Brain Tumor MRI Image Classification with Feature Selection and Extraction using Linear Discriminant Analysis", International Journal of Information Sciences and Techniques, vol.2, no.4, pp. 131-146.

[20] Quora: https://www.quora.com/What-are-theadvantages-of-different-classification-algorithms.