Performance Analysis of Machine Learning Classifiers for Brain Tumor MR Images

Lubna Farhi[1], Razia Zia[2], Zain Anwar Ali[3]

Abstract— **Brain cancer has remained one of the key causes of deaths in people of all ages. One way to survival amongst patients is to correctly diagnose cancer in its early stages. Recently machine learning has become a very important tool in medical image classification. Our approach is to examine and compare various machine learning classification algorithms that help in brain tumor classification of Magnetic Resonance (MR) images. We have compared Artificial Neural Network (ANN), K-nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers to determine the accuracy of each classifier and find the best amongst them for classification of cancerous and noncancerous brain MR images. We have used 86 MR images and extracted a large number of features for each image. Since the equal number of images, have been used thus there is no suspicion of results being biased. For our data set the most accurate results were provided by ANN. It was found that ANN provides better results for medium to large database of Brain MR Images.**

Index Terms— **ANN, SVM, NB, DT, KNN and GLCM.**

I. INTRODUCTION

A tumor is the term given to an uncontrollable and abnormal growth of cells in the body. In general, there are two types of tumors, benign or malignant. The first type is a benign tumor which is classified as such, due to its inability to spread to neighboring tissues or metastasize to other parts of the body. Hence it is classified as a non-cancerous type of tumor. However, a Benign tumor can turn Malignant (the second type of tumor), which is highly dangerous and can lead to death if untreated.

In order to detect such abnormalities within the body, various non-intrusive imaging techniques are used. These include computed tomography (CT), X-ray, Ultrasound and Magnetic Resonance Imaging (MRI). Amongst these techniques, MRI has shown to provide a higher resolution image as it combines magnetic fields and radio waves to capture an image of a body's interior. There are three major scans of MR images. T1, T2 and FLAIR weighted images. Functional images are T2 scan since they provide the highest contrast between different types of tissue. Currently, MR images are manually interpreted by a doctor or physician in order to identify the presence of a tumor which is time intensive and error prone.

Many techniques have been developed for brain tumor detection and classification which involves minimum human interactions. Some of these techniques are Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network and Support Vector Machine (SVM) among others.

Recently W. Chu *et al* [1], provides a new approach towards determining most efficient method for least square-SVM. Furthermore, Chang *et al* [2], [3] in 2003 presented that SVM classifier provides better results for tumor classification than ANN. The author has used 140 images of benign breast tumors and 110 samples of malignant breast tumors. However, due to the small number of features used for classification, results could not be thought of as reliable. EL-Sayed developed a hybrid methodology for classification of brain MR images [4]. In order to analyze medical images, especially brain images, and classify the type of tumor, the determination of tissue type i.e., abnormal or normal and tissue pathology is essential [5]. The technique consists of three main steps: feature extraction using DWT, dimensionality reduction using PCA and classification using ANN and k-NN. The classification technique provides accuracy of 95.6% and 98.6% for Artificial Neural Network and k-Nearest Neighbor respectively. The database used consists of 70 images in which 60 MRI images are of the abnormal human brain (i.e., cancerous) while 10 images show the normal human brain (non-cancerous). The results of the technique look promising there is a suspicion of ANN biasing due to the difference in the number of images for each class. The author has studied Gray Level Co-Occurrence Matrix (GLCM) as an absolute image quality metric [6].

The author has used six image sets each consisting of 5 images on various levels of compression. It was found that the most suitable value for the radius of the GLCM was one, while no definitive conclusion could be found for, the value of GLCM angle. It was also found that GLCM is a good technique to study images of different sizes, however, no conclusion could be found regarding the quality of the image. Shweta [7] has used GLCM feature extraction technique along with Artificial Neural network to classify brain cancer into four types: Pilocytic (grade I), Low Grade (grade II), Anaplastic (grade III) and Glioblastoma Multiforme (grade IV). Images are 256x256 gray level images having intensity between 0 and 255. The author has extracted seven GLCM based features from each image and has created GLCM matrices for four angles: 0º, 45º, 90º, 135º.

Since the images were not part of any recognized dataset, the results could not be verified. Similarly, the authors in [8] have used ANN to classify presence or absence of brain tumor. The authors have created five co-occurrence matrices in four

[1]Department of Electronic Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan. ifarhi@ssuet.edu.pk
[3]Department of Electronic Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan. raziaz@ssuet.edu.pk
[3]Department of Electronic Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan. zaali@ssuet.edu.pk

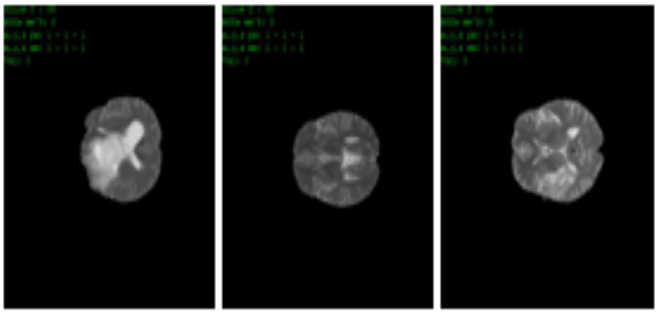
spatial directions using a dataset of 38 patients. Since the number of patients as well as the extracted features is very small thus the accuracy of the results could not be verified.

In this paper, we compare five techniques for brain tumor classification and analyze their performance. Each technique is evaluated by keeping in consideration its accuracy by using the probabilistic features, which contain the textural characteristics of image such as contrast, energy, entropy, etc., [9], [10].

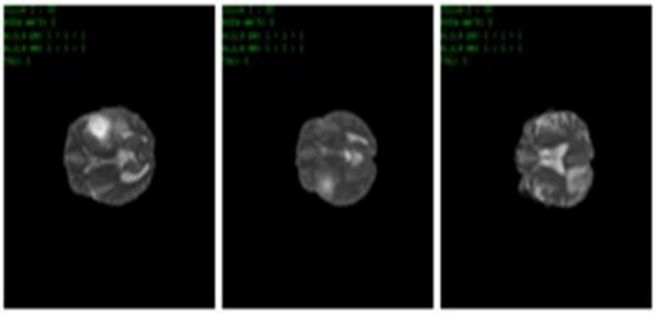
II. RESEARCH METHODOLOGY

A. Dataset

In this study, we used eighty six T2 weighted brain MRI images of high and low grade tumors collected from BRATS, 2012 dataset [11]. MRI brain tumor images are currently being examined, and graded by doctors based on their appearance under a microscope. Typically, tumors are characterized into four categories, also known as grades. The first grade, Grade I, is the first stage of brain cancer in which tumorous cells resemble normal brain cells and exhibit only minor growth. The second grade, called Grade II, is the stage of cancer when tumorous cells start to become malignant and lose their resemblance from ordinary brain cells. A Grade III tumor is the next stage of cancer where the cells are rapidly spreading (Anaplastic) and do not exhibit any property of a functioning brain cell. The last stage, also known as Grade IV, is identified as a mass of tumorous tissue that is growing uncontrollably. As a result, the first two stages, Grade I and II, are termed as low-grade tumors while Grade III and Grade IV are identified as high-grade tumors.



(a)



(b)

Fig. 1: (a) Low-Grade (Astrocytoma) (b) High-Grade (Glioblastoma) Tumor Images

B. Probabilistic Feature Extraction

The feature is defined as physiognomies of an object. Feature extraction is used to reduce the dimensions of a grayscale image while extracting all useful data from it. The extracted feature vector is used for classification.

In this study, 22 Probabilistic features were extracted from each image as defined in [12], [13], and [14].A Probabilistic Matrix of each image p (i, j), has the $(i, j)^{th}$ value of the image grid. The image is quantized to have a number of gray levels where μ_y, μ_x and σ_y, σ_x are the mean and standard deviations of the image matrix.

$$\mu_x = \Sigma i \, \Sigma j \;\; i.p(i,j) \quad (1)$$

$$\mu_y = \Sigma i \, \Sigma j \;\; j.p(i,j) \quad (2)$$

$$\sigma_x = \Sigma i \, \Sigma j \; (i- \mu_x)^2 . \; p(i,j) \quad (3)$$

$$\sigma_y = \Sigma i \, \Sigma j \; (j- \mu_y)^2 . \; p(i,j) \quad (4)$$

The extracted features are discussed in Table I

C. Classification Methods

Image classification is a process in which a set of distinct attributes is first extracted from an image and mapped to a predefined class. Once the classification is complete, a thematic map can be created that can be used to distinguish images with numerical values close to each other in the same class. Classification is divided into two categories: Supervised and unsupervised. Unsupervised classification, as the name suggests, requires no initial human input. This means that the images are not predefined by anyone, instead the algorithm defines the image clusters by finding similarities and relations between pixels in the images. However, supervised classification requires that the images are predefined by the user into distinct classes during the training phase. Then, the classifier uses the spectral signature of classes which it learns during the training phase to identify individual classes for new images.

In this research, we compare the accuracy of the classifier by using following algorithms:

i. Artificial Neural Networks:

An artificial neural network is a type of machine learning model that is based on the biological nervous system. The network comprises of a set of nodes, called neurons, and connections between those nodes. The different connections between neurons determine the flow of data and the global behavior of the network. There are two stages of operations of ANN. The first stage is the training phase where the network learns the input data patterns and the corresponding outputs. While in the second stage, known as the testing phase, it predicts the output of an unknown dataset based on what it has learned in the previous stage.

$$y = f\left(\sum_{j=1}^{d} w_j x_j + w_0\right) \equiv f\left(\sum_{j=0}^{d} w_j x_j\right) \quad (5)$$

Back Propagation algorithm is one of the most efficient methodologies used to train ANN. It is used to feed-forward multilayer networks that consist of continuous differentiable activation functions and neurons. The artificial neural network contains three types of layers: input, output and hidden layers.

Table I: Extracted Features

FEATURE	DESCRIPTION
Energy	Energy is the measure of intensity uniformity for an image.
Entropy	This feature defines the measure of disorder in an image.
Homogeneity	The inverse different moment is called homogeneity of an image.
Contrast	Contrast shows the difference between the brightest and the lowest intensity of set of adjacent pixels in an image.
Correlation	Correlation defines how much gray tones are linearly dependent on each other.
Cluster Prominence	It defines the asymmetry of an image.
Cluster Shade	Cluster-shade is similar to cluster prominence in that it also defines the lack of symmetry in the image.
Variance	Variance of an image is the computation of how much gray level values vary from their mean.
Autocorrelation	It shows the similarity between gray tones as a function of time lag.
Dissimilarity	Dissimilarity defines the dependence or independence of pixel intensities on each other.
Maximum Probability	Maximum probability defines the maximum intensity observed in the image.
Sum, Average	$\sum_{i=2}^{2Ng} i\, g_{x+y}(i)$ $g_{x+y}(k) = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} g(i, j)$ $k=i+j$
Sum Variance	$\sum_{i=2}^{2Ng} (i - \text{SUM ENT})^2\, g_{x+y}(i)$
Sum Entropy	$-\sum_{i=2}^{2Ng} i\, g_{x+y}(i) \log\{ g_{x+y}(i) \}$
Difference Variance	Variance of g_{x-y} $g_{x-y}(k)= \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} g(i,j)$ $k=\lvert i-j \rvert$
Information Measures of Correlation	$\frac{HXY - HXY1}{\max\{HX, HY\}}$ $(1-\exp [-2.0(HXY2-HXY)]^{1/2})$
Difference Entropy	$-\sum_{i=0}^{Ng-1} i\, g_{x-y}(i) \log\{ g_{x-y}(i) \}$

The activation function of the artificial neuron is the addition of all inputs x_i multiplied by the relevant weights w_{ij} of the neurons.

$$Aj(x, w) = \sum_{i=0}^{n} xi\ wji \quad (6)$$

Let us define *y* as k- dimensional vector of real numbers such that $y \in R^K$ and $h_\Theta(x)$ also k- dimensional vector, so, $h_\Theta(x)_i$ refers to the i^{th} value in that vector. $h_\Theta(x) \in R^K$ $(h_\Theta(x))_i = i^{th}$ output. If m shows the number of training data and λ defines a regularization parameter, then the cost function of J (θ) layered artificial neural network, having *k* distinct classes and *sl* neurons in each layer, such that the cost function is defined by the following equation whose output is a k dimensional vector.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\theta(x^{(i)}))_k + (1-y_k^{(i)}) \log(1-(h_\theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m} + \sum_{l=1}^{L-1}\sum_{i=1}^{sl}\sum_{j=1}^{sl+1}(\theta_{ji}^{(l)})^2 \quad (7)$$

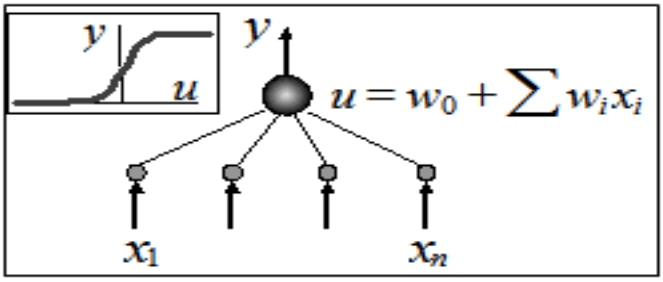


Fig. 2: Single Artificial Neuron

Information which is sent to input layer may be numerical, the input layer forwards it to the hidden layer. Every neuron that exists in the hidden layer has a set activation value and information from the input layer is used to activate these neurons. The several hidden layers are set to achieve minimum error at the output layer. The learning method used in Back Propagation is supervised learning which means that in the training stage the inputs and outputs of the network are provided and the error between expected and actual result is computed. This algorithm essentially reduces the error between the two results by increasing or decreasing the number of layers.

ii. Decision Tree:

In Decision tree the datasets are repeatedly partitioned into smaller and more uniform datasets. These subsets are again divided into further subsets through various variables and features.

The data, is divided depending upon the maximum reduction in deviance over all splits of all the nodes, to select the succeeding split. If a node is divided into nodes u and y, and then the deviance reduction is calculated by:

$$D = D_s - D_u - D_y \quad (8)$$

Where,

$$D_i = -2\, \Sigma_k\, n_{ik} \log p_{ik} \quad (9)$$

nik is the number of cases related to class k in any node i and pik represents the probability distribution of class k in the node i.

The three steps involved in the classification of a data using Decision tree classifier are; partitioning of nodes, determining terminal nodes, allocation of class labels to the terminal nodes. The key features of Decision tree are; very less computational load and it do not require extensive network training. That is the reason that it is best suited for non-parametric data.

iii. K-Nearest Neighbor:

KNN is a nonparametric classifier which classifies an object into a particular class depending on its degree of likeness to that class [15]. The input dataset is alienated into k classes during the training stage of the KNN classifier, with individual class comprising of inputs belonging to its class. The

Euclidean distance is calculated between all inputs and then they are partitioned into k classes. Each input consists of m features. The Euclidean distance between two inputs xi and xl is found by:

$$d(xi, xl) = \sqrt{(xi_1 - xl_1)^2 + (xi_2 - xl_2)^2 + .. + (xi_m - xl_m)^2} \quad (10)$$

During the testing stage, the Euclidean distance is calculated between arbitrary or unknown input xa and all points in the KNN classifier. The smallest distance between this input xa and a class tells its association to that particular class.

iv. Naïve Bayes:

Naïve Bayes classifier is probabilistic classifiers. It applies Bayes' theorem to find the probability of a sample to belong to a particular class [16]. It assumes that all features of the samples in a specific class are independent of each other, given the context of the class. If x represents an input sample with m features such that:

$$x(n) = \{x_1, x_2, \ldots\ldots x_m\} \quad (11)$$

Then, according to Bayes' theorem, for a class C P (x|C) is the posterior probability of C conditioned on x, P(x) is the prior probability of x and P(C) is the prior probability of C. The probability of x belonging to class C is given by:

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} \quad (12)$$

After calculating probability for each of the *n* classes the Bayes classifier categorizes image *x* to a class C_i if:

$$P(C_i|X) > P(C_j|X) \quad (13)$$

If $\lambda(\alpha_i|C_j)$ is considered to be the loss encountered for taking action α_i, when the true class is C_j, the conditional risk for the Bayesian classifier for, *m* classes are expressed as:

$$R(\alpha_i|x) = \sum_{j=1}^{m} \lambda(\alpha_i|C_j)\, P(C_j|x) \quad (14)$$

If we consider zero-one loss function, then the conditional risk for the Bayesian classifier can be simplified as:

$$R(\alpha_i|x) = 1 - P(C_j|x) \quad (15)$$

v. Support Vector Machine:

SVM belongs to the class of non-parametric classifier just like k-NN. A hyper-plane or a set of hyper-planes in a high dimensional space are built between two classes in order to differentiate them. Usually it is used in binary classification, however it can be modified to classify data into more than two classes using 1-vs-1 or 1-vs-all technique.

SVM provides good generalization capability along with reduction in computational complexity as well as removal of over fitting of data. However, the training time of SVM is large compared to other learning algorithms and optimal parameters are difficult to determine when there is non-linearly separable data. SVM is capable of using linear, polynomial or Sigmoid Kernel functions for decision function which makes it very versatile. Let a sample *x* having *m* attribute i.e. $x=(x_1, x_2, \ldots, x_m)$ and each pattern x_j belongs to class $y_j \in \{-1, +1\}$. For *n* patterns the training set is defined by $T = \{(x_1, y_1), (x_2, y_2), (x_n, y_n)\}$ and S is a dot product space, where patterns *x* are fixed, $(x_1, x_2, \ldots, x_n \in S)$. The hyper-plane in space *S* is written as:

$$\{x \in S | w.x + b = 0\}\ ,\ w \in S,\ b \in R$$

And

$$w \,.\, x = \sum_{i=1}^{n} w_i x_i \quad (16)$$

The logistic regression for a classifier is:

$$J(\theta) = \min_{\theta} \frac{1}{m}\left[\sum_{i=1}^{m} y(i) cost1(\theta^T x) + \big(1 - y(i)\big)(cost0(\theta^T x)) + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta j^2\right] \quad (17)$$

For binary SVM, we have two logistic regression at y=0 and y=1 defined as cost0 (θT x) and cost1 (θT x) respectively.

So, the cost function for binary SVM is defined as:

$$J(\theta) = \min_{\theta} \frac{1}{m}\left[\sum_{i=0}^{1} y(i) cost1(\theta^T x) + \big(1 - y(i)\big)(cost0(\theta^T x)) + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta j^2\right] \quad (18)$$

For multiclass SVM the logistic regression in the classifier is equal to the number of classes.

D. Performance Evaluation Parameters

Performance of classifiers can be found by following evaluation parameters:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (19)$$

$$\text{Sensitivity} = \frac{TP}{TP+FP} \quad (20)$$

$$\text{Specificity} = \frac{TN}{TN+FN} \quad (21)$$

Table II: Summary of Evaluation Parameters

		Predicted Class	
		Cancer Positive (+)	Cancer Negative (-)
Actual Class	Cancer Positive (+)	TP (++)	FN (+-)
	Cancer Negative (-)	FP (-+)	TN (- -)

Where, FP = false positive, and FN = false negative, TP = true positive, TN = true negative; +, - are two labels, of classes. TP can be said to occur when classification results are positive with the presence of clinically proven deviation while TN is when the result is negative in the absence of deviation. Furthermore, FP proves that the result is positive with the absence of clinically proven deviation, while FN is when the result is negative in the presence of clinically proven deviation.

III. IMPLIMENTATION

Classification consists of the following realization steps

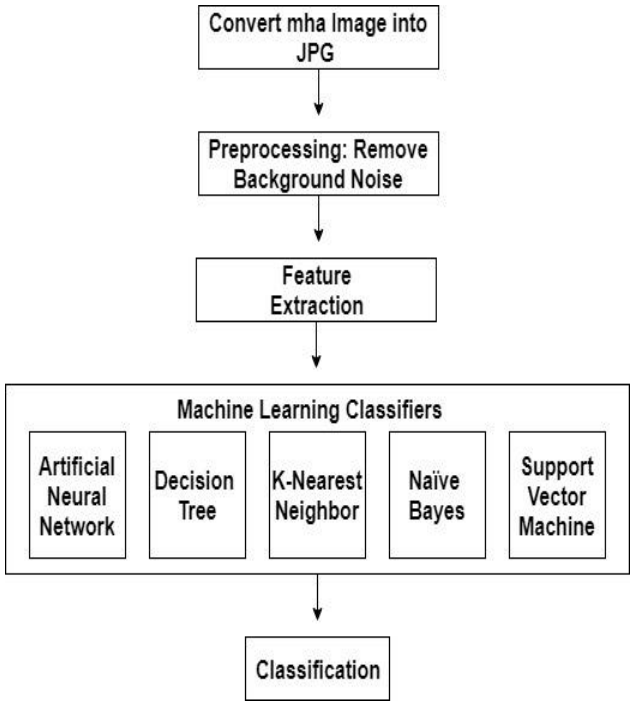


Fig. 3: Flow Chart

IV. RESULT AND ANALYSIS

Brain tumor MRI eighty six T2 weighted images of low grade (class I and II) and high grade (class III and IV) are used for comparative study [17]. We compute 22 feature of each slice therefore input vector is 86x22. The performance measured and outcomes of five classifiers are shown in Table IV.

The Fig.4 represents the comparison between evaluation parameters. The performance evaluation graph shows that the Artificial Neural Network gives the best performance after 8 iterations. The training data accuracy decreases after epoch 8, while validation and test accuracy has a slight increase. Due to the large slope in training data accuracy after epoch 8, the overall accuracy decreases. Thus the best performance is obtained after 8 iterations.

Table III: Accuracy and Confusion Matrix

	Accuracy %	Confusion matrix TP FN FP TN
ANN	87.4	35 14 8 29
DT	84.2	37 6 5 38
K-NN	79.1	43 0 18 25
Naïve Bayes	54.6	16 27 12 31
SVM	51.1	10 33 9 34

Table IV: Comparison of Evaluation Parameters

	Sensi-tivity	Speci-fictiy	Preci-sion	Error rate
ANN	0.91520	0.9901	0.9710	0.021046
DT	0.93333	0.9591	0.96552	0.055046
KNN	0.9987	0.57143	0.74074	0.22018
Naïve Bayes	0.48333	0.7551	0.70732	0.3945
SVM	0.61667	0.67347	0.69811	0.3578

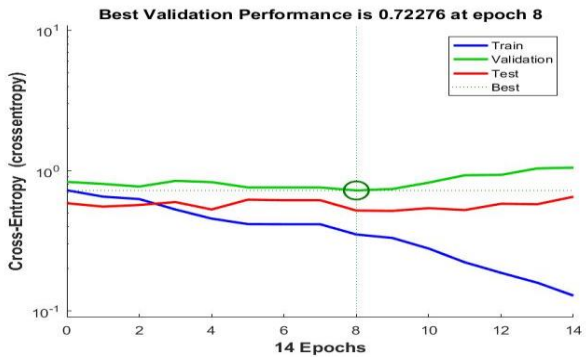


Fig. 4: ANN Performance Curve

Region of the convergence curve (ROC) for training and testing data shows the overall performance according to the accuracy of ANN classifier.

V. CONCLUSION

In this paper, we compare various machines learning classification method. The accuracy of each of the classification techniques is analyzed by their Confusion Matrices and evaluation parameters. Simulated results show that artificial neural network classifier has better accuracy of up to 87.4%, for binary brain tumor identification for low grade and high grade tumors. It has been noted, that accuracy varies with the size of the dataset and true features extraction. We also found that, increase in the number of images without an increase in true features extracted; correspond to a decrease in accuracy.

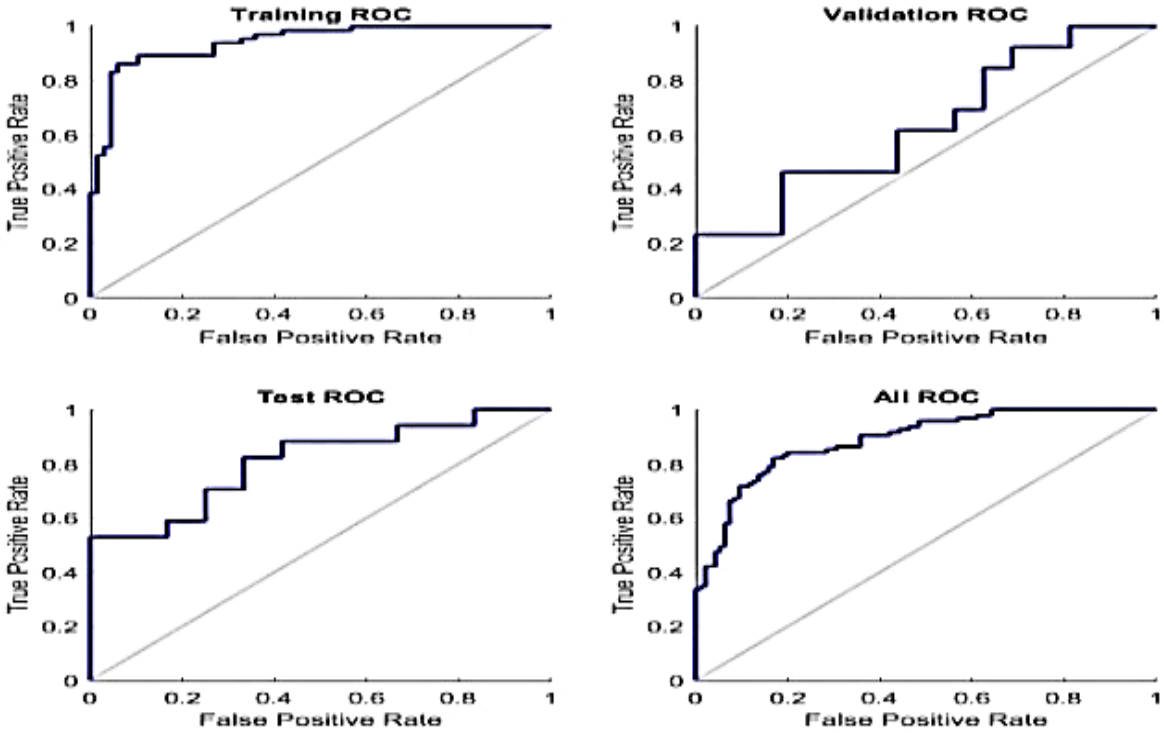


Fig. 5: ANN ROC Curve

Therefore, it was concluded that the number of dataset images in each class, feature extraction method and the selected classification model, play a big role in identifying the type of tumor with better accuracy of brain MRI images.

REFERENCES

[1] Chu, W., Ong, C. J., & Keerthi, S. S. (2005). An improved conjugate gradient scheme for the solution of least squares SVM. IEEE Transactions on Neural Networks, 16 (2), 498-501.

[2] Chang, R. F., Wu, W. J., Moon, W. K., Chou, Y. H., & Chen, D. R. (2003). Support vector machines for diagnosis of breast tumors on US images. Academic Radiology, 10 (2), 189-197.

[3] Chang, R. F., Wu, W. J., Moon, W. K., & Chen, D. R. (2003). Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis. Ultrasound in medicine & biology, 29 (5), 679-686.

[4] El-Dahshan, E. A., Salem, A. B. M., & Younis, T. H. (2009). A hybrid technique for automatic MRI brain images classification. Studia Univ. Babes-Bolyai, Informatica, 54 (1), 55-67.

[5] Vidyarthi, A., & Mittal, N. (2014). Comparative study for brain tumor classification on MR/CT images. In Proceedings of the Third International Conference on Soft Computing for Problem Solving (pp. 889-897). Springer, New Delhi.

[6] Ekenel, H. K., Fischer, M., Gao, H., Kilgour, K., Marcos, J. S., & Stiefelhagen, R. (2007, November). Universität Karlsruhe (TH) at TRECVID 2007. In TRECVID.

[7] Jain, S. (2013). Brain cancer classification using GLCM based feature extraction in artificial neural network. International Journal of Computer Science & Engineering Technology, 4 (7), 966-970.

[8] Kathalkar, A. A., Kawitkar, R. S., & Chopade, A. (2013). Artificial neural network based brain cancer analysis and classification. International Journal of Computer Applications (0975–8887) Volume, 66.

[9] Kamavisdar, P., Saluja, S., & Agrawal, S. (2013). A survey on image classification approaches and techniques. International Journal of Advanced Research in Computer and Communication Engineering, 2 (1), 1005-1009.

[10] Farhi, L., & Yusuf, A. (2017, February). Comparison of brain tumor MRI classification methods using probabilistic features. In IASTED International Conference, Innsbruck, Austria Biomedical Engineering.

[11] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2 (2), 121-167.

[12] Haralick, R. M., & Shanmugam, K. (1973). Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, (6), 610-621.

[13] Soh, L. K., & Tsatsoulis, C. (1999). Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. CSE Journal Articles, 47.

[14] Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of gray level quantization. Canadian Journal of remote sensing, 28 (1), 45-62.

[15] Zhang, M. L., & Zhou, Z. H. (2005, July). A k-nearest neighbor based algorithm for multi-label classification. In Granular Computing, 2005 IEEE International Conference on (vol. 2, pp. 718-721). IEEE..

[16] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (vol. 3, No. 22, pp. 41-46). New York: IBM.

[17] Bratz.com. (n.d.). Brain Tumor MR Images. URL: http//bratz.com