Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier

Ada[1], Rajneet Kaur[2]

[1]Student of masters of technology Computer Science, *Department of Computer Science and Engineering,*
Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.

[2]Assistant Professor, *Department of Computer Science and Engineering,*
Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.

ABSTRACT

The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. Classification is very important part of digital image analysis. It is a computational procedure that sort images into groups according to their similarities. In this paper Histogram Equalization is used for preprocessing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that we predict the survival rate of a patient by extracted features. Experimental analysis is made with dataset to evaluate the performance of the different classifiers. The performance is based on the correct and incorrect classification of the classifier. All experiments are conducted in WEKA data mining tool.

Keywords- Classification, Feature Extraction, CT-Scan images, Lung Cancer.

1.INTRODUCTION

In the modern age of computerized fully automated trend of living, the field of automated diagnostic systems plays an important and vital role. Automated diagnostic system designs in Medical Image processing are one such field where numerous systems are proposed and still many more under conceptual design due explosive growth of the technology today [1]. Lung cancer is considered to be the main cause of cancer death worldwide, and it is difficult to detect in its early stages because symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of cancer. More people die because of lung cancer than any other types of cancer such as breast, colon, and prostate cancers. There is significant evidence indicating that the early detection of lung cancer will decrease mortality rate [2]. There are many techniques to diagnose lung cancer, such as Chest Radiography (x-ray), computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology. However, most of these techniques are expensive and time consuming. In other words, most of these techniques are detecting the lung cancer in its advanced stages, where the patients' chance of survival is very low. Therefore, there is a great need for a new technology to diagnose the lung cancer in its early stages. Image processing and data mining techniques provide a good quality tool for improving the manual analysis [2].

2.METHODOLOGY

Our system has been fully implemented (in matlab) and tested with real CT scan images.

2.1 Images Collected

We have collected the 909 CT-Scan images of lung cancer from the private hospital. The digitized images are stored in the DIACOM format with a resolution of 8 bits per plane.

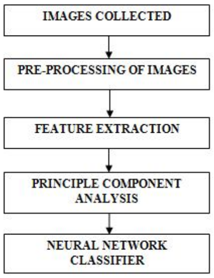


Figure1 Methodology of work

2.2 Pre-processing of Images

Most of the pre-processing is done with the help of MATLAB software. Each image sample is stored to a size of 512 X 512 pixels. Generally, the quality of image is affected by different artifacts due to non-uniform intensity, variations, motions, shift, and noise [3]. Thus, the pre-processing of image aims at selectively removing the redundancy present in scanned images without affecting the details which that play a key role in the diagnostic process. Hence, Histogram-Equalization becomes the important step in preprocessing. Therefore each image is preprocessed to improve its quality.

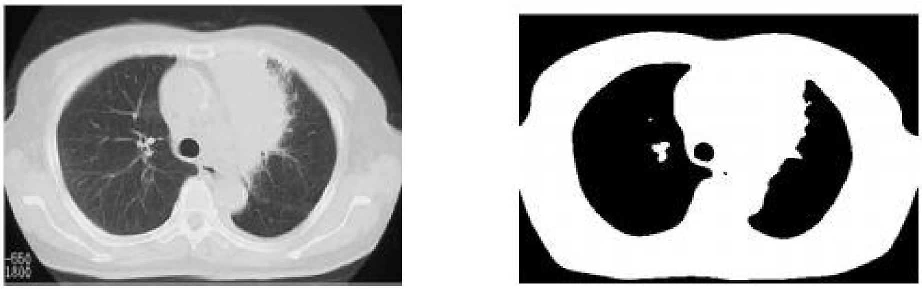

Input Image Output Image

Figure2 Shows the Histogram Equalization on CT scan image

2.3 Morphological Operators

Morphological operations are affecting the form, structure or shape of an object [12]. Applied on binary images (black & white images – Images with only 2 colors: black and white). They are used in pre or post processing (filtering, thinning, and pruning) or for getting a representation or description of the shape of objects/regions (boundaries, skeletons convex hulls).

Common Morphological Operations [11]-

- Shrinking the foreground (“erosion”)
- Expanding the foreground (“dilation”)
- Removing holes in the foreground (“closing”)
- Removing stray foreground pixels in background(“opening”)
- Finding the outline of the foreground
- Finding the skeleton of the foreground

2.4 Feature Extraction

Image features Extraction stage is an important stage that uses algorithms and techniques to detect and isolate various desired portions or shapes (features) of a given image. To predict the probability of lung cancer presence, the following two methods are used: binarization and GLCM, both methods are based on facts that strongly related to lung anatomy and information of lung CT imaging.

2.4.1 GLCM (Grey Level Co-occurrence Method) [5]

The GLCM is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in an image. Firstly we create gray-level co-occurrence matrix from image using *graycomatrix* function in MATLAB. Then we normalize the GLCM using the following formula

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}}$$

Where i is the row number and j is the column number. From this we calculate texture measures from the GLCM. The following features [5] are extracted using this method-

- Contrast
- Energy
- Entropy

- Homogeneity
- Maximum Probability
- Correlation
- Cluster shade
- Cluster Prominence
- Dissimilarity
- Autocorrelation
- Sum variance
- Sum Entropy
- Difference Variance
- Difference Entropy
- Information Measure

2.4.2 Binarization Approach [4]

Binarization approach has been applied for detection of cancer. In this we extract the **number of white pixels** and check them against some threshold to check the normal and abnormal lungs. If the number of the white pixels of a new image is less that the threshold, then it indicates that the image is normal, otherwise, if the number of the white pixels is greater than the threshold, it indicates that the image in abnormal.

Combining Binarization and GLCM approaches together will lead us to take a decision whether the case is normal or abnormal.

2.5 PCA (Principle Component Analysis) [6]

PCA is to standardize the data in image. Real-world data sets usually exhibit relationships among their variables. These relationships are often linear, or at least approximately so, making them amenable to common analysis techniques. One such technique is principal component analysis ("PCA"), which rotates the original data to new coordinates, making the data as "flat" as possible.

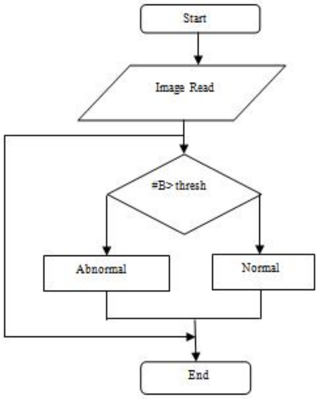


Figure3 Binarization check method [4]

The features extracted are passed through the PCA data mining for better classification. The following steps takes place in PCA:-

i. Calculate the mean and standard deviation of the features in the image using MATLAB.

ii. Subtract the sample mean from each observation, then dividing by the sample standard deviation. This centers and scales the data.

iii. Calculating the coefficients of the principal components and their respective variances is done by finding the Eigen functions of the sample covariance matrix.

iv. The matrix contains the coefficients for the principal components. The diagonal elements store the variance of the respective principal components. We can extract the diagonal.

v. The maximum variance in data results in maximum information content which is required for better classification.

2.6 Neural Network Classifier

Supervised feed-forward back-propagation neural network ensemble used as a classifier tool. As discussed previously, neural network differs in various ways from traditional classifiers like Bayesian and k – nearest neighbor classifiers. One of the main differences is linearity of data. Traditional classifiers like Bayesian and k – nearest neighbor requires linear data to work correctly. But neural network works as well for non-linear data because it is simulated on the observation of biological neurons and network of neurons. Wide range of input data for training makes neural network to work with higher accuracy, in other words a small set of data or large set of similar data makes system to be biased .Thus neural network classifier requires a large set of data for training and also long time to train to reach the stable state. But once the network is trained it works as fast as biological neural networks by propagating signals as fast as electrical signals [7].

The architecture of the neural network consists of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables [8].

Steps Performed in Neural Network Classifier:-

- Create feed-forward back propagation network.
- Train neural network with the training samples and the group defined for it.
- The input image extracted PCA standardized data as the test samples, simulate the neural network to check whether the particular selected input sample has cancer or not.
- From the results of network and the samples trained in network classification rate is calculated using some mathematical formulas.

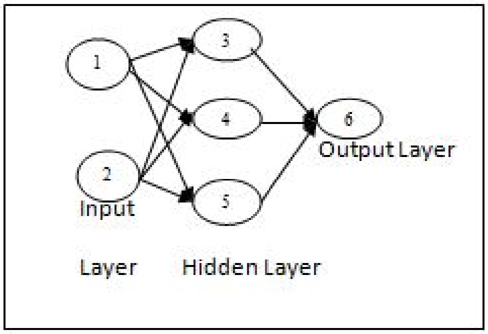


Figure4 A neural network with one hidden layer[8]

2.7 Survival rate and year

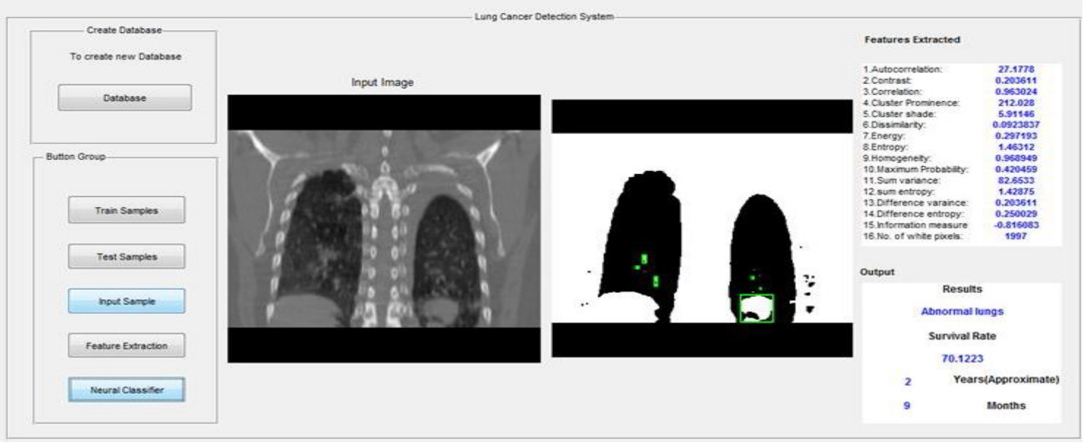
Lung cancer survival rates are a measure of how many people remain alive with lung cancer after a certain amount of time. For example, a 5-year survival rate of 40% for a condition would mean that 40% of people, or 40 out of 100 people, would be alive after 5 years. When talking about lung cancer, physicians often use the term median survival as well. Median survival is the amount of time at which 50% of people with a condition will have died, and 50% are still alive [10].

Fine contrast feature demonstrated a substantial degree of concordance with PET tumor staging (stage I, contrast <3.2591; stage II, $3.2591\leq$ contrast ≤ 4.2632; stage III, $4.2632<$ contrast ≤ 4.9345; stage IV: contrast > 4.9345.Furthermore a fine contrast above 4.2632 predicted tumors above stage II [9].

3.RESULTS AND COMPARISON

Lung CT-Scans has been collected and after image processing total 16 features have been extracted from the images and then classification is done on the images to check the normal or abnormal state of the patient. After that we check the survival rate and year of the patient on the basis of features extracted.

3.1 For abnormal lungs



Features Extracted

1.Autocorrelation:	27.1778
2.Contrast:	0.203611
3.Correlation:	0.963024
4.Cluster Prominence:	212.028
5.Cluster shade:	5.91146
6.Dissimilarity:	0.0923837
7.Energy:	0.297193
8.Entropy:	1.46312
9.Homogeneity:	0.968949
10.Maximum Probability:	0.420459
11.Sum variance:	82.6533
12.sum entropy:	1.42875
13.Difference varaince:	0.203611
14.Difference entropy:	0.250029
15.Information measure	-0.816083
16.No. of white pixels:	1997

Output

Results

Abnormal lungs

Survival Rate

70.1223

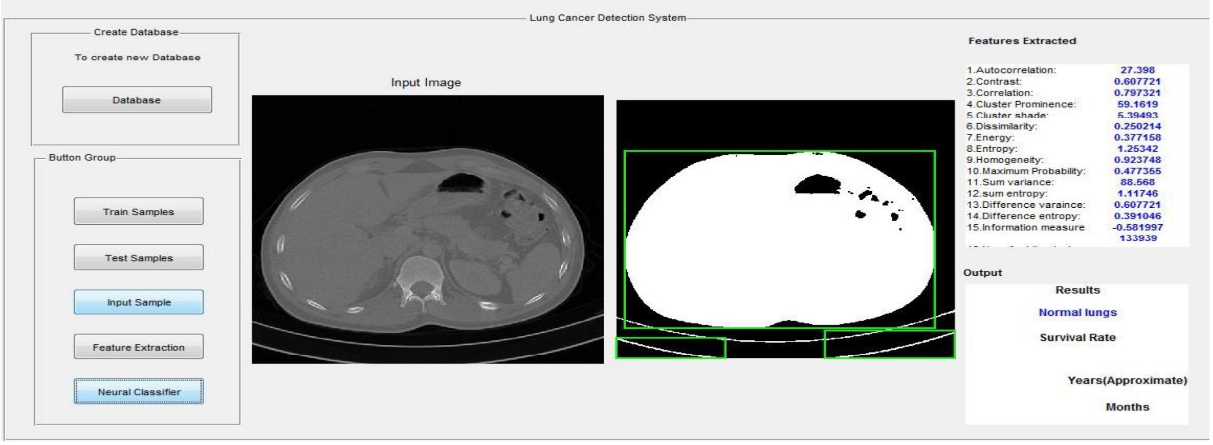
2 Years(Approximate)

9 Months

Figure5 Result of Abnormal Lungs

In this firstly we load the database and then training and test sets. After that we can input the image between 1 to 909 and calculate the values of features extracted. Now in next step classification is done with the help of neural network which tells the state of the patient whether it is normal or abnormal. Along with this on the basis of features we can also calculate the survival rate and year of the abnormal patient.

3.2 For Normal Lungs



Features Extracted

1.Autocorrelation:	27.398
2.Contrast:	0.607721
3.Correlation:	0.797321
4.Cluster Prominence:	59.1619
5.Cluster shade:	5.39493
6.Dissimilarity:	0.250214
7.Energy:	0.377158
8.Entropy:	1.25342
9.Homogeneity:	0.923748
10.Maximum Probability:	0.477355
11.Sum variance:	88.568
12.sum entropy:	1.11746
13.Difference varaince:	0.607721
14.Difference entropy:	0.391046
15.Information measure	-0.581997
	133939

Output

Results

Normal lungs

Survival Rate

Years(Approximate)

Months

Figure6 Result of Normal Lungs

For normal lungs we cannot calculate the survival rate and year of the patient.

3.3 Comparisons

A database file of 909 tuples and 16 attributes has been made in ASCII in CSV format, then conversion of this file to CSV file is done. Then Convert this file into ARFF which are readable in Weka. The generated ARFF file is opened in Weka and then calculate the learning rate of supervised classification using tp rate and fp rate. Along with this attribute removal is also done. Some of the attributes are removed as they contribute nothing in classification process, have been implemented and results have been compared on different parameters. Each of the method is tested at the Percentage split set to 30% on dataset.

3.3.1 Performance Parameters

The comparison of various classification algorithms is done on the basis of following performance parameters:

a) **TP Rate:** The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall. In the confusion matrix, this is the diagonal element divided by the sum over the relevant row.

b) **FP Rate**: The False Positive (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. In the confusion matrix, this is the column sum of class x minus the diagonal element, divided by the rows sums of all other classes.

c) Correctly Classified Instances: Correctly classified instances are used in order to find out which algorithm correctly classifies maximum number of instances.

d) Incorrectly Classified Instances: Incorrectly classified instances are used in order to find out which algorithm incorrectly classifies maximum number of instances.

e) Root Mean Square Error: The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed.

f) Mean Absolute Error: The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

3.3.2 Training rate of supervised classification

a) Using TP Rate

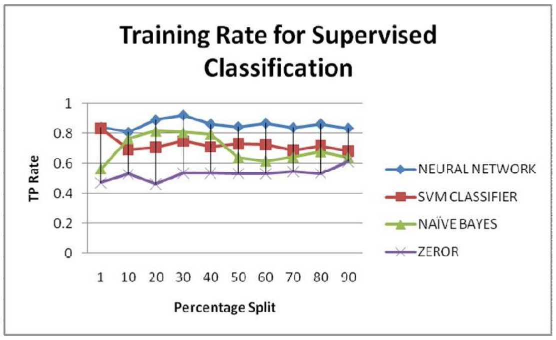


Figure7 Training Rate for Supervised Classification

In this graph we compare the TP rate of different classifiers with the percentage split in which we start the split from 1% and goes to 90 % that means 1% is training set and rest 90% is testing set and we identify that at 30% we get the maximum value of TP rate so training rate is 30% in this case.

b) Using FP Rate

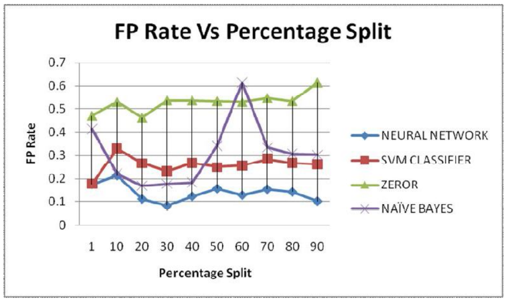


Figure8 FP Vs Percentage Split

In case of classification after 30 % near about 90 % the parameters goes below the value specified at 30% of percentage split. This is because of the over training. So we can say that 30 % is the required training in case of supervised classification.

3.3.3 Minimal feature set

Information gain algorithm is applied and attributes are arranged in the order of significance. Number of White Pixels attributes shows the highest priority. Learning rate of four algorithms has been obtained by starting the training from 1% till 90%. It has been observed at 30% of percentage slit in training classifier gives the maximum value. Classification algorithms have been implemented on the 909 Lung CT-Scan images dataset using the percentage split of 30% during training. Algorithms which are compared against different parameters are ZeroR, Naïve Bayes, Neural

Network and SVM algorithm. Values of different parameters have been compared in order to evaluate which algorithm gives better results in case of Lung CT-Scan Images.

a) Using TP Rate

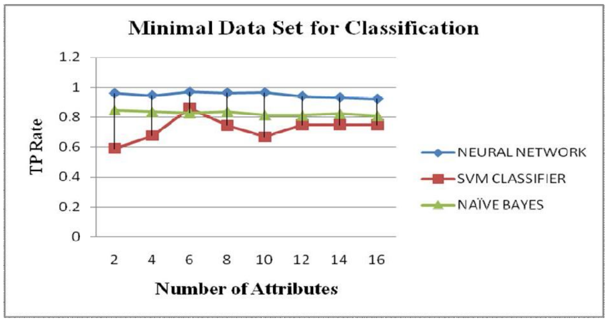


Figure9 Minimal feature set using TP Rate

b) Using FP Rate

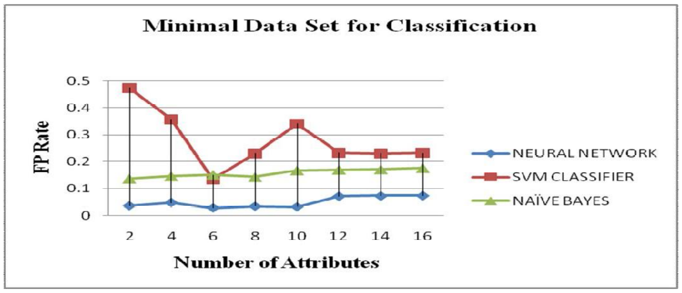


Figure10 Minimal feature set using FP Rate

c) Mean Absolute Error

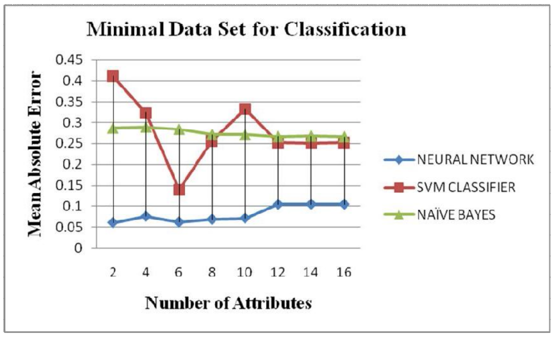


Figure11 Minimal feature set using Mean Absolute Error

d) Root Mean Square Error

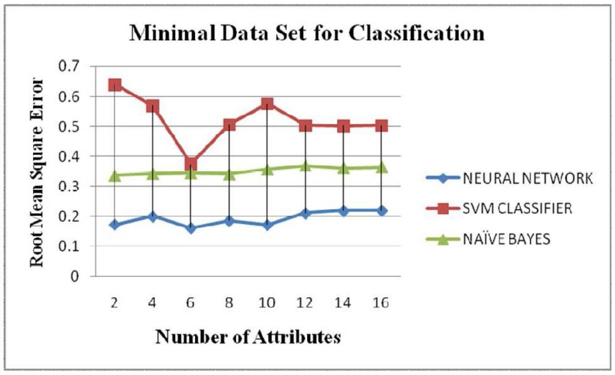


Figure12 Minimal feature set using Root Mean Square Error

Now we are come to know from minimal feature set that only 12 attributes are important for the work. After 12 features all algorithms give constant value to the other features. So we compare the different classifiers on these 12 features.

3.3.4 Comparison of Classifiers at 12 Features

In order to find out which algorithm correctly classifies maximum number of instances. 30% of attributes have been taken in training phase and rest in the testing phase. Obtained values have been plotted below

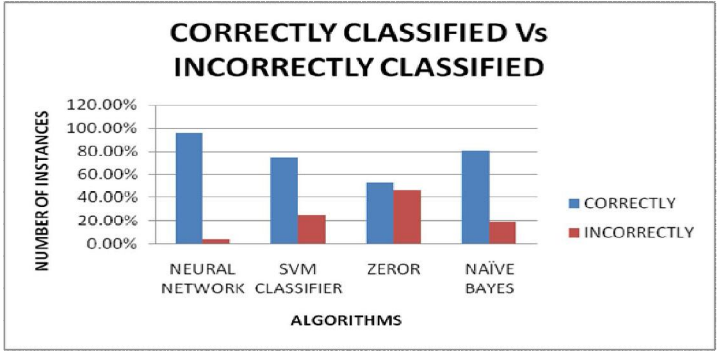


Figure13 Correctly Classified Vs Incorrectly Classified Instances

It has been observed from the plotted vales that Neural Network correctly classified 96.4% of instances whereas ZeroR classifies 53.30%, Naïve Bayes classifies 63.44%, SVM classifies 72.69% of instances. Neural Network gives maximum number of correctly classified instances and ZeroR gives minimum number of correctly classified instances.

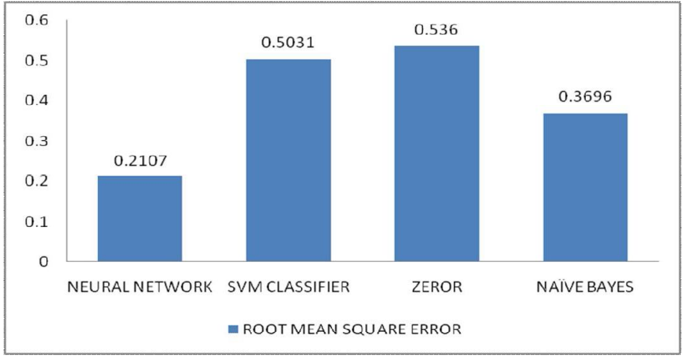


Figure14 Root Mean Square Error Comparison

It has been observed from the plotted values that Neural Network classifier gives minimum value and other Classifier gives maximum value of Root Mean Square Error.

4.CONCLUSION

A database file consisting of 909 images with 12 lists of attributes is prepared by using minimal feature set approach and it used for comparison in next phase. Neural Network Algorithm is implemented using open source and its performance is compared to other classification algorithms. It shows the best results with highest TP Rate and lowest FP Rate and in case of correctly classification, it gives the 96.04% result as compare to other classifiers.

5.FUTURE SCOPE

- The million order dataset can be selected and image classification can be done on larger dataset. With increased size of dataset various issues such as uploading data, managing feature set, increased execution time of classification algorithms etc. can be considered.
- More image features can be extracted for better classification. Various combinations of preexisting features can be used to correctly classify medical data.
- The researchers can put their emphasis by implementing ANT COLONY in the combination of the NEURAL NETWORK to check out over the accuracy of the survival rate of the patients.

References

[1.] Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, “ Artificial Neural Network based Cancer Cell Classification (ANN – C3)”, Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.

[2.] Almas Pathan, Bairu.K.saptalkar, “Detection and Classification of Lung Cancer Using Artificial Neural Network”, International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue:1.

[3.] Dr. S.A.PATIL, M. B. Kuchanur, ” Lung Cancer Classification Using Image Processing,” International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[4.] Mokhled S. AL-TARAWNEH, “Lung Cancer Detection Using Image Processing Techniques,” Leonardo Electronic Journal of Practices and Technologies Issue 20, January-June 2012, p. 147-158.

[5.] Fritz Albregtsen, “Statistical Texture Measures Computed from Gray Level Coocurrence Matrices,” International Journal of Computer Applications, November 5, 2008.

[6.] Taranpreet Singh Ruprah, “Face Recognition Based on PCA Algorithm,” Special Issue of International Journal of Computer Science & Informatics (IJCSI), 2231–5292, Vol.- II, Issue-1, 2.

[7.] ZAKARIA SULIMAN ZUBI1, REMA ASHEIBANI SAAD, “Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer”, Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, LIBYA, 2011.

[8.] Balaji Ganeshan, Sandra Abaleke, Rupert C.D. Young, Christopher R. Chatwin, Kenneth A. Miles, “Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage,” Cancer Imaging , v.10(1): 137–143, 2010 July 6.

[9.] Lynne Eldridge MD. (2013, March 22). Lung Cancer Survival Rates by Type and Stage [Online]. Available: http://lungcancer.about.com/od/whatislungcancer/a/lungcancersurvivalrates.htm.

[10.] Morphological Operators, CS/BIOEN 4640: Image Processing Basics, February 23, 2012.

[11.] Image Processing – Laboratory 7: Morphological operations on binary images, Technical University of Cluj-Napoca, Computer Science Department.