Content based image retrieval system for Classification of breast region as mass or non masses using new shape invariant texture feature and SVM.

[1]Er. Sumit Chopra,[2]Dr. V . K . Banga
[1]Research scholar IKGPTU,[2] Principal , ACET, Amritsar
[1]CSE Department,
[1]K.C. College of Engineering and IT, Nawanshahr, India

Abstract— **Breast cancer is one of the deadliest problems that the women in European countries are facing. Several computer detection methodologies have been proposed which help experts to identify the suspicious region that are difficult to find with the naked eye, thus helping in diagnosis and detection of cancer. A methodology for classification of the regions extracted from mammogram to classify them as mass or non-mass is proposed. Mammographic Image Analysis Society (MIAS) database was used for image acquisition. Two new features were described for texture of the region of interest which has added advantage that even if the region of interest is not segmented properly, then the results will not be affected as texture is invariant to the shape of the region of interest extracted. Internal and external masks were used for analysis of texture. Support vector machine for used for classification of region as masses or non-masses. The results showed a large improvement as compared with the existing state of art techniques.**

IndexTerms— **Content Based Image Retrieval (CBIR) , Masses, Classification, Diversity tree, Cladogram**

I. INTRODUCTION

In recent times, the most common cancer found in women is breast cancer. Earlier it is diagnosed, more are the chances of recovery and more it enhances the treatment efficiency. While the correct diagnosis, saves time and unnecessary medical treatment, incorrect diagnosis leads to unnecessary cost and harassment of the patients due to unnecessary treatment.

One preventive method for the breast cancer is mammography. Women crossing the age of 40 are advised to go through routine mammography tests to check for breast cancer. The results of the mammography are checked by specialist or radiologist who interprets the results from the mammographic images. This step is a sensitive stage as the same mammogram can be interpreted differently by different specialists. Mammograms are obtained repeatedly as even small details can lead to incorrect diagnosis. One of the methods of being sure that cancer is present or not is by means of biopsy.

When other tests show that one might have breast cancer, one will probably need to have biopsy. Needing a breast biopsy doesn't necessarily means you have cancer. Most biopsy results are not cancer, but biopsy is the only way to find out. During a biopsy, a doctor will remove cells from the suspicious area so that they can be looked in the lab to see if cancer cells are present. If the doctor says that one don't need biopsy, but one still feels that there is problem with the breast, then one shouldn't hesitate in asking doctor about this or move to another doctor for second opinion.

For these reasons, there has been growing trend of research on image processing techniques to be used in mammograms with the objective of increasing diagnostic precision and providing second opinions to the experts. These techniques are further enhanced to develop Computer- Aided Detection/Computer- Aided diagnostic (CAD/CADx) systems.

Various algorithms have been developed for increasing the accuracy of the breast cancer detection using Computer Aided Detection system. The main problems with the existing systems are high rate of false positives, high rate of false negatives and reduced number of cases in evaluation, which provides a better conclusion. So, need of the hour is to design efficient CAD system to support breast mass calcification.

Generally, in most of the Content Based Image Retrieval system, the feature extraction stage is based on either shape like how round is the candidate under consideration or on texture, which describes aspects of the candidate based on grey level distribution. In our methodology, only texture feature is used for feature extraction. Using only texture, for the feature extraction process, has the added advantage that the segmentation results will not affect the output.

In biology, the term diversity is used to represent the different species of organism present in a community or area. A community is defined by set of species that occur in a certain location and at a certain time. Phologeny is a branch of Biology concerned with studying the evolutionary relationships between species, by verifying the relationships among them, in order to determine possible common ancestors. The edges of the trees denote the evolutionary relationships [1].

For describing the texture of the mass and non-mass two features based on texture known as Taxonomic diversity and distinction indexes are used. The first index considers the richness of the species and the taxonomic relationship between them, whereas the second represents the average taxonomic distance between the two individuals of a distinct species.

II. RELATED WORK

In recent times, researchers are working relentlessly on the studies for early detection of breast cancer by means of mammography using image processing and pattern detection techniques. In this section, a thorough literature survey of the related work have been given, which have strong relationship with the methodology used in the paper. Table 1 briefs the various related technologies.

Various features can be used for mammographic images in order to classify the region as mass or non- masses. After the segmentation of the region of interest, the next step in mammographic images is to extract features from the region of interest. The mammogram images can be filtered using Gabour wavelets and directional features are extracted at different orientation and frequencies. Principal Component analysis can be used for reducing the dimensions of filtered and unfiltered high dimensional data [2] . Contourlet coefficients can be employed as a feature extractor to obtain the contourlet coefficients. The features can be selected using genetic algorithm which results in more compact and discriminative feature set [3]. The discriminating breast tissue patterns can be obtained by variants of Local Ternary Pattern and Local Phase Quantization . It shows very good results for distinguishing benign from malignant tissues [4].

In [5] , a Radial Basis Function Neural Network for mammograms classification based on Gray-level Co-occurrence Matrix (GLCM) texture based features was developed which turned out to be better than Back-propagation Neural Network in performing breast cancer classification. In [6], after segmentation of region of interest from mammographic images, the texture features were extracted from grey level co - occurrence matrix. Then classification of region as benign or malignant with the help of three classifiers namely adaboost, back propagation neural network and sparse representation classifiers.

In order to improve the results of the other texture methods of feature extraction, a new feature of texture variant, namely Law's Texture Energy Measure was used to improve the results of mammogram classification. Training data for the mammogram classification model is retrieved from Mammographic Image Analysis Society database [7]. In order to simplify the process of classification of mammograms into benign and malignant after segmentation of region of interest, an algorithm using simple image processing tasks of averaging and thresholding was developed. Max – Mean and Least- Variance technique was used for tumor detection [8].

A Content Based Image Retrieval system was designed in [9] which interpreted mammographic lesions based on medical characteristic specified in Breast Imaging Reporting and Data Systems (BIRADS) standard. A hierarchical similarity measure based on distance weighting function is used to maximize the effectiveness of each feature in a mammographic descriptor. A machine learning approach based on support vector machine and user relevance feedback was used to analyze the user's information need in order to retrieve target images more precisely. In [10], a CBIR system was developed which allows medical professionals to seek mass lesions which are pathologically similar to a given example. The shape and margin features of mass lesions are extracted to represent the characteristic of mammographic lesions and matched with the database images using hierarchical arrangement of mammographic features and a weighted distance measure.

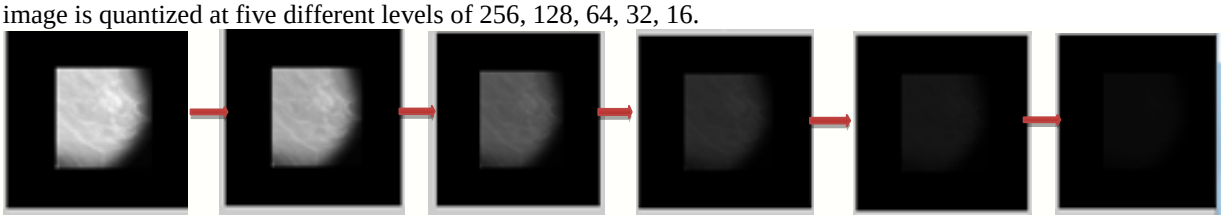
III. PROPOSED METHODOLOGY

The algorithm used in the classification of mammogram images into benign and malignant consists of three main steps which include image preprocessing, feature detection and Classification. The steps can be further subdivided into various stages given below:-

Image acquisition – The image used in the algorithm is from the MIAS database. While benchmarking an algorithm it is recommended to use a standard test dataset for comparison of results. The two most easily public database are Mammographic Image Analysis Society (MIAS) and Digital Database for Screening Mammograms (DDSM). In our algorithm MIAS database is used for image acquisition.

Image preprocessing - After image acquisition, the next step is to do preprocessing of the mammographic images so that the even fine details contribute to the feature extraction process. Logarithmic non – linear contrast enhancement process is used to improve the quality of the regions and a mean filter of 5 X 5 mask is used to eliminate small structure from the regions. Image preprocessing is an essential step for classification result as good enhancement technique will lead to good results whereas bad enhancement technique will lead to incorrect classification.

Image Quantization – The image acquired from the MIAS database is required to be converted from colored format to grey scale image. After converting the image from the grey scale to colored, the next step is to go for quantization of the image. Quantization is the process of converting a range of input values into smaller output values which closely approximates the original data. Quantization is a lossy compression technique achieved by compressing a range of values to a single quantum value. The image is quantized at five different levels of 256, 128, 64, 32, 16.



(a) (b) (c) (d) (e) (f)

Fig. 1 Image quantization at five different grey levels of 256, 128 , 64, 32, 16 . Fig. 1(a) Original image 1(b) Image quantization at 256. 1(c) Image quantization at 128 1(d) Image quantization at 64. Fig. 1(e) Image quantization at 32 Fig. 1(e) Image quantization at 16.

Image thresholding - After quantization next step is thresholding of image. Thresholding of an image is a process of converting a grey level image into binary image by turning all the pixels below an optimal threshold to zero and all the pixels above the optimal threshold to one. If g_k denotes the intensity of the pixel in an image and Po_r represent occurrence probability of intensity g_k for an image. The probability of occurrence of the grey value g_k in the figure is represented in Eq. (1).

$$po_r(g_k) = \frac{m_k}{m} \quad k = 0,1,2,.....,L-1 \qquad (1)$$

The input image is transformed into an enhanced version by going through the transformation function given in the equation (ii) Where L contains possible intensities in an image and m_k contains the pixels number with intensity g_k. Thus, transformed image is formed by portraying every grey level g_k in the input image into an equivalent intensity level o_k in the transformed image as given in Eq. (2).

$$o_k = T(r_k) = \sum_{j=0}^{k} po_r(g_j) = \sum_{j=0}^{k} \frac{m_j}{m} \quad k = 0,1,2,.....,L-1 \qquad (2)$$

Image segmentation – After thresholding the next step is to partition an image into set of non overlapping but when sub parts are combined they form an original image. The purpose of the segmentation is to decompose an image into parts that are meaningful for a particular application. The segmentation will output the image part that is to be checked for cancer. After the quantization step, the region of interest is divided into layers to perform a local analysis, which is useful because these areas of the layer may supply information that can distinguish benign and malignant masses. The pixels at the center will give the origin of the mass and the boundary pixels indicates how a mass has grown. Thus, we look for texture feature by means of polygenetic tree in each layer. Internal and external masks are used separately for finding these layers. The approach is discussed in the next section.

Computation of variables using internal and external mask - This approach is used to find diversity patterns in the areas close to the border of the regions and in the inner areas. The regions are generated through masks, which are binary images. The first internal mask was determined by binarization of the quantized ROI. By reducing the scale with respect to the first one while maintaining the center of mass, the successive internal masks are generated. A value of 20% for the diminution of the scale, as verified by various tests and five image masks with this scaling proportion provided best results. With this approach, the total number of variables generated is 25. Five variables for the internal mask computation and five variables are generated for each quantization level which totals to 25. The external mask is determined by finding the difference between successive masks. For example, the first external mask is found by subtracting the first internal mask from the second internal mask and so on. The number of variable generated using this approach will be 20. In total there will be four masks and each for each mask there will be five quantization level which will total to 20 variables.

Diversity tree - Using the concepts of Biology, an evolutionary relationship between species is represented in terms of hierarchical trees called Diversity trees. In these trees, the leaves represent the species and internal nodes represent the common ancestors to the species. It is possible to make an evolutionary connection between species being species. These trees show the inferred evolutionary relationship between among various biological species or other entities – their phylonogy based upon similarities and differences in their physical or genetic characteristic. The graphical representation used to describe the relationship between the ancestors species is known as inclined cladogram.

Phologenteic trees allow the extraction of indexes that connect diversity, richness and parenthood between species. Fig. 2 represents an example of ape's phylogenetic tree represented by an inclined cladogram, where one can see that a chimpanzee has a higher phylogenetic proximity to human as compared to Siamang and Gibbon. In the inclined cladogram, the edges depict the phylogenetic distance between two species, the leaf nodes are the species being analyzed, the internal nodes correspond to some common ancestor.

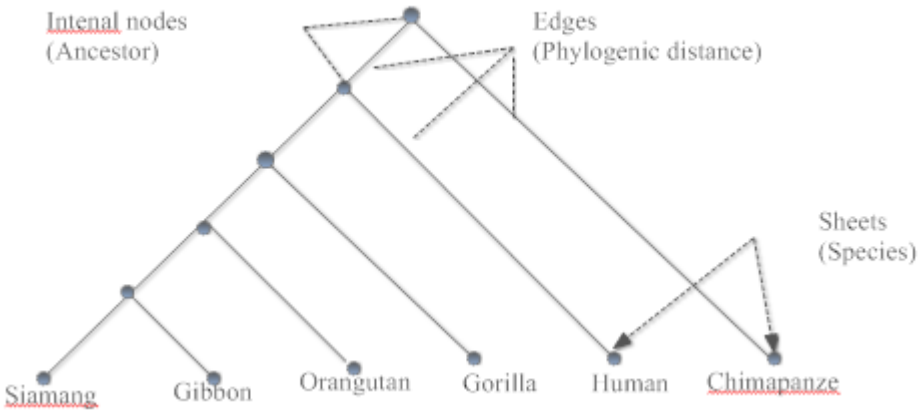
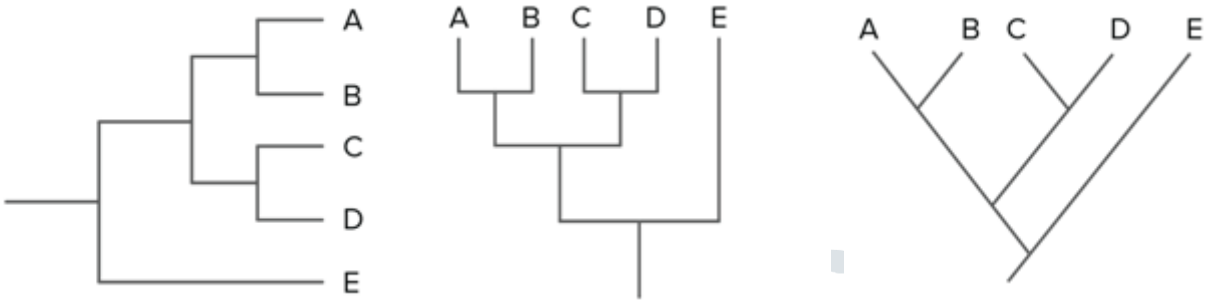


Fig. 2 Diversity or Phylogenetic tree Source [11]

The difference between two randomly chosen organism in phylogeny existing within a community is described by two parameters namely taxonomic diversity (Δ) and taxonomic distinctness indexes (Δ*) [12] A phylogenetic tree is a diagram that represents evolutionary relationships among organisms. Phylogenetic trees are hypotheses, not definitive facts. The pattern of branching in a phylogentic tree reflects how species or other groups evolved from a series of common ancestors. In trees, two species are more related if they have a more recent common ancestor and less related if they have a less recent common ancestor. Phylogenetic trees are drawn in many different formats. Some are blocky, like the trees shown in Fig. 3(a). Others use diagonal lines as shown in Fig. 3(b). We may also find trees of either kind oriented vertically or horizontally on their sides as shown in Fig. 3(c).Rotating a tree about its branch doesn't change the information it carries. Most modern systems of classification are based on evolutionary relationships among organisms- that is, on the organism's phylogeny. Classification systems based on phylogeny organize species or other groups in ways that reflect our understanding of how they evolved from their common ancestors.



(a) (b) (c)

Fig. 3 Different formats of Diversity Trees

The difference between two randomly chosen organisms in a phylogeny existing within a community is described by the taxonomic diversity (Δ) and taxonomic distinctness (Δ*) indexes. These indexes considers the three main factors namely the number of species, the number of individuals and the connectivity structure of the species i.e. the number of edges. These two indexes are used to differentiate between masses and non masses.

The taxonomic diversity index (Δ) considers the abundance of the species and the taxonomic relationship between them. Thus, its value expresses the mean taxonomic distance between any two individuals, randomly picked from the sample.[12]. The parameter is defined as follows

$$\Delta = \frac{\sum\sum_{i<j} w_{ij} x_i x_j}{[n(n-1)/2]} \qquad (3)$$

where $x_i (i = 0, ..., s)$ is the abundance of the ith species, $x_j (j = 0, ..., s)$ is the abundance of the jth , s represents the number of species, n is the total number of individuals and w_{ij} is the distance from species i to species j in the taxonomic classification .

The taxonomic distinctness index (Δ^*) , in turn, represents mean taxonomic distance between two individuals that belong to different species [12]. This index is defined by

$$\Delta^* = \frac{\sum\sum_{i<j} w_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j} \qquad (4)$$

where $x_i (i = 0, ..., s)$ is the abundance of the ith species, $x_j (j = 0, ..., s)$ is the abundance of the jth , s represents the number of species and w_{ij} is the distance from species i to species j in the taxonomic classification .

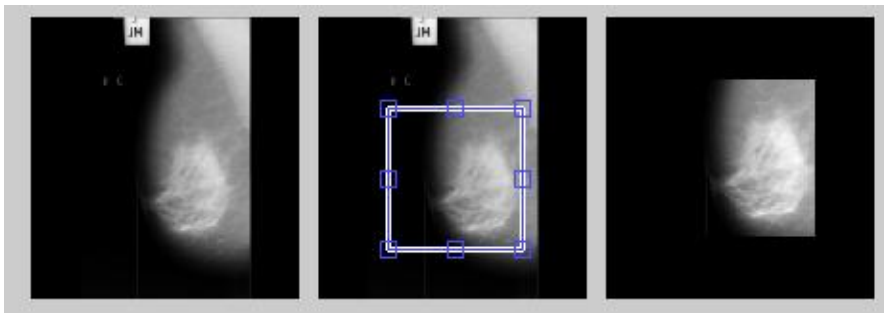
There are many architectures in the literature that represent the species through trees, such as the architecture of a rooted tree in the shape of an inclined cladogram [13].The term community in biology represents the region of interest for mammographic images (ROI).Species in biology represents the maximum number of grey levels in ROI. Richness of species represent the number of pixels for a specific gray level value found in the ROI. Individuals represent the number of pixels of a particular species contained in the ROI. Relative abundance represents the number of pixels found in the ROI, which have the same gray level value (species). Fig. 4 represents the results of the proposed methodology. Fig. 4 (a) represents the test image. 4(b) represents the breast area and 4(c) represents the extracted region. 4(d) represents the extraction of ROI using internal and external mask .

Support Vector Machines (SVM) :- In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Suppose a set X of I retrieved mammograms, X, which have been indicated by the user as relevant and irrelevant are given as x = {xi| I = 1, 2, …….l}, where xi is th ith mammogram in x. Let y = {yi| i= 1,2,3,…..l: }, where yi is either 1 or -1 which is the class label with respect to x, with yi=1 indicating that xi has been specified by the user as relevant and yi=-1 as irrelevant. The set of returned mammograms x can be optimally separated by the hyper plane [14].

$$w.x - b = 0 \qquad (5)$$

Where w is a normal vector perpendicular to the hyper plane while b is the displacement of hyper- plane from the original along w. The hyper plane that optimally separates the positive and negative images can be obtained by finding the smallest possible w.

As the data set x are often not linearly separable. SVM maps the input data into a higher dimensional space through an underlying nonlinear mapping function and then finds an optimal hyper plane in the feature space [15].



(a) (b) (c)

(d)

Fig. 4 Results of proposed methodology (a) Test image (b) Breast area (c) Breast area extraction (d) Segmentation of ROI using internal and external masks

	Recall and Precision values		
S.No	***Test Plan***	***Precision***	***Recall***
1	Training/Test -> (90/10)	0.99	0.98
2	Training/Test -> (80/20)	0.98	0.99
3	Training/Test -> (70/30)	0.99	0.98
4	Training/Test -> (60/40)	0.98	0.98
5	Training/Test -> (50/50)	0.99	0.98
6	Training/Test -> (40/60)	0.97	0.96
7	Training/Test -> (30/70)	0.83	0.853
8	Training/Test -> (20/80)	0.83	0.80
9	Training/Test -> (10/90)	0.72	0.70

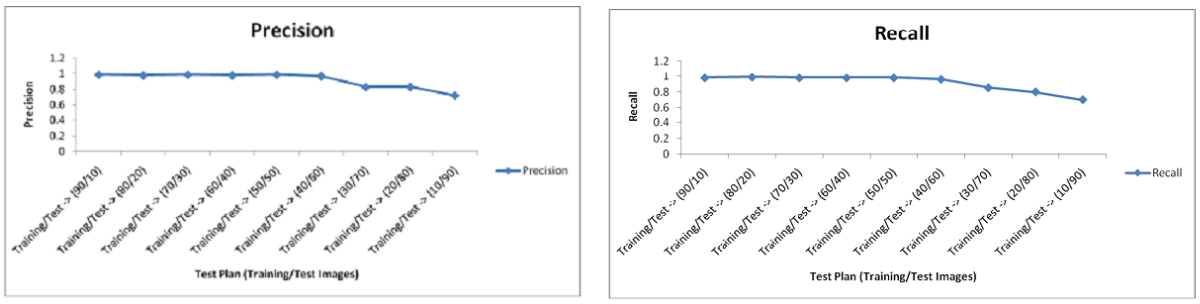


Fig. 5 Precision and Recall Graph using different set of training and testing images

IV. CONCLUSION

A Content Based image retrieval has been developed for classification of mammographic images as cancerous or non-cancerous using Support Vector Machine. This system helps the radiologists to identify and diagnose the cancer. The features used in the

CBIR system are based on texture which means that even if the region of interest is not segmented properly, the results won't be affected as it is invariant to the region of interest When compared with the existing state of art techniques, it shows improved results.

REFERENCES

[1] F.S.S.D.Oliveria,A.O.D.C.Filho,A.C.Silva,A.C.D.Paiva,M.Gattass, "Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM," Elsevier Computers in Biology and Medicine, vol.57, pp.42- 53,2015.

[2] I.Buciu,A.Gascadi ," Directional features for automatic tumor classification of mammogram images," Biomedical Signal Processing and Control, vol.6,no.4, pp.370-378,October 2011.

[3] F.Moayedi,Z.Azimifar,R.Boostani,S.Katebi, "Contourlet- based mammography mass classification using the SVM family," Elsevier Computes in Biology and Medicine,vol.40,no.4, pp. 373-380, 2010.

[4] L.Nanni,S.Brahnam,A.Lumini, "A very high performing system to discriminate tissues in mammograms as malignant and beingn," Elsevier Expert System With Applications,vol.39,no.2, pp.1968-1971, 2012.

[5] M.Pratiwi,Alexander,J. Harefa,S.Nanda, "Mammogram Classification using Gray- level Co- occurrence Matrix and Radial Basis Function Neural Network,"Elsevier Procedia Computer Science in International Conference on Computer Science and Computational Intelligence (ICCSCI) ,vol.59, pp.83-91,2015.

[6] K.Vaidehi,T.S.Subashini, "Automatic Characterization of Benign and Malignant masses in Mammography,"Elsevier Procedia Computer Science in International Conference on Information and Communication Technologies ,vol. 46, pp. 1762-1769, 2015.

[7] A.S.Setiawan,Elysia,J.Wesley,Y.Purnama, "Mammogram Classification using Law's Texture Energy Measure and Neural Network," Elsevier Procedia Computer Science in International Conference on Computer Science and Computational Intelligence,vol. 59, pp. 92-97, 2015.

[8] A.K.Singh,B.Gupta, "A novel approach for Breast Cancer Detection and Segmentation in Mammograms," Elsevier Procedia Computer Science in Eleventh International Multi- Conference on Information Processing ,vol. 54, pp. 676- 682, 2015.

[9] C.H.Wei,Y.Li,P.J.Huang, "Mammogram retrieval through machine learning within BI-RADS standards," Elsevier Journal of Biomedical Informatics ,vol.44, no.4, pp.607-614,August 2011.

[10] C.H.Wei,S.Y.Chen,X.Liu, "Mammogram retrieval on similar mass lesions,"Elsevier Computer Methods and Programs in Biomedicine,vol. 106,no.3, pp.234-248, 2012 .

[11] A.D.Baxevanis,B.F.F.Ouellette, " Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Methods of Biochemical Analysis," Wiley,New York, NY,USA,2004.

[12] M.W.Pienkowski,A.R.Watkinson,G. Kerby,K.R.Clarke,R.M.Warwick, "A taxonomic distinctness index and its statistical properties," Journal of Applied Ecology,vol.35,no.4, pp. 523-531, 1998.

[13] H.A.S.Moura,G.V.R.Vaina, "Anace: Phylogenetic Trees Drawing Web Service," BIOTECHNO : The third International Conference on Bioinformatics, bio computational System and Biotechnologies, pp. 73-77, 2011

[14] C.J.Burges, "A tutorial on support vector machines for support vector machines," Data mining and knowledge discovery , vol.2, no.2, pp.121-167, 1998.

[15] B.Scholkopf,S.Mika,C.J.C.Burges,P. Knirsch,K.R.Muller, K.,G.Ratsch,A.J.Samola, "Input space versus features space in kernel – based methods," IEEE Transactions on Neural Network ,vol.10,no.5, pp. 1000-1017, 1999.