International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.10 No.7, pp 23-31, 2017

Deep Learning Algorithm for Cancer Tissue Image Classification

E.Nagarajan*, Ch.Deekshitha, Ch.Uthpala

Department of Computer Science and Engineering, Sathyabama University, Chennai-119, India

Abstract : Application of artificial intelligence through the methods and algorithms of machine learning brought a drastic change in the field of developing and building intelligent machines. Deep learning techniques have been proven to be best ways to solve problems involving training and analysis of data in big quantities. Our aim is to apply the machine learning techniques/algorithms to train the histo-pathological image data of breast cancer tissue images for prediction of cancer intelligently. The machine learns from the data trained using machine learning algorithms and takes decisions over intelligent prediction of cancer.
Keywords : Image processing, SVM algorithm, Digital image.

Introduction:

The concepts of image processing are mainly employed to intensify the image quality and image data in order to extract the useful information from the digital images through different pixel-wise operations. We are here trying to do the same i.e. by applying the image transformation operations to process images and extract information prior to our necessity. Our aim is to predict cancerous tissue images by comparing them against a set of trained images that are processed through image processing techniques and then trained using machine learning algorithms. The best combination of image processing techniques and machine learning can help us in increasing the correctness ratio of getting the output of cancer image detection. For this purpose we are considering the digital histo-pathological images of cancer tissues.

Experimental:

Machine learning:

Deep learning being a branch of machine learning learns necessary feature -representations directly from data using the technique of non-linear processing layers. Deep learning models exceed the human-level potential of achieving accuracy. The concepts of deep learning are employed in order to manage the large set of image data (cancer histo-pathology) for training the data-base the helps the machine in learning the patterns and comparing them against the query images to detect the presence and type of cancer in it. Deep learning algorithms of image classification and image comparison are used this application. Unlike traditional machine learning methods the concept of deep learning tends to fuse all the steps involved in the whole process to wrap them into a single deep learning algorithm that can be employed to the application directly. This process is explained by comparing it to the traditional way of machine learning process in a pictorial way in the below figure.

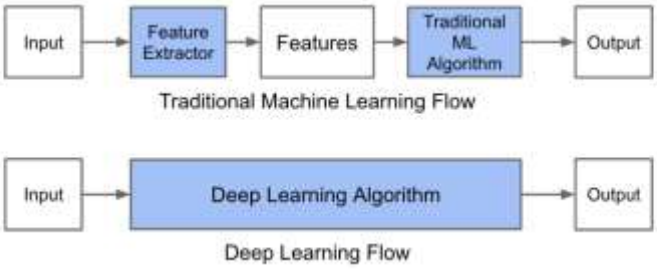

Fig.1 Machine learning and deep learning

Deep learning is applied to different types of problems like:

- Image classification
- Speech recognition
- Natural language processing

Related work:

In the work done by Chien-Chi Chen over the algorithms of image segmentation, he proposed two methods namely, "*Edge Based Segmentation"* and "*Region-Based Segmentation"*.[1]The image gradient vector is defined in the basic –edge detection using the below given formulae at a given point location (*x,y)*.

$$\nabla f = grad(f) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

The direction of edge is orthogonal to the direction of arbitrary point(x,y).Mehmet Sezgin and Bulent Sankur discussed different methods of thresholding in their work in "Survey over image thresholding techniques" in which they explained the thresholding based on histogram shapes[2]. In the histogram the distance between convex hulls is examined in the references 3 and 4-7. For the histogram h(g) the theoretic difference is calculated using the formula

|Hull(g)-p(g)|

Where, *hull(g)* is the convex hull of histogram *h(g)* and *p(g)* is the probability mass function.

Methodology (Results and discussion):

1. Pre-processing the image:

The main methodology for our work starts with pre-processing the query image which we try to diagnose to predict the existence of cancer in that image. For this we follow few steps of image processing that are explained below briefly.

Image resizing: the given input image is first resized by *scaling* the image for given number of times. The simple mat-lab code used to resize image is briefed here.[r c p]=size(im);im=imresize(im,[255 255]);

Once the image is resized it is then *gray-scaled* and *filtered*—F = medfilt2(image);

A gray scale image is produced in order to specify a single intensity value for red, green and blue color intensities in RGB space. All these color components are filled the same gray-color component as it is opposed to have three color components specifying each pixel of a color image. Once this processing is done the images are threshold for binarization. BW = imbinarize(I)

The above mat-code creates a binary image from image I by replacing all values above a globally determined threshold with 1s and setting all other values to 0s.

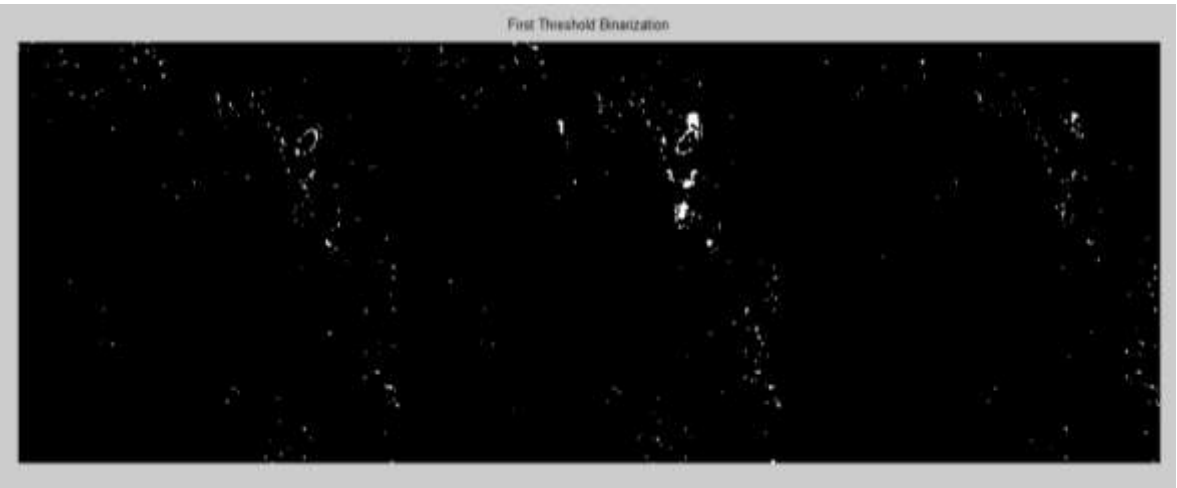

Fig2 Binarised tissue image

As soon as the image is binarized it is set to fuzzy C-Means clustering using the following mat-code for two levels of clustering namely level '0' and level '1'.

[bwfim0,level0]=fuzzy_c_means_cluster(fim,0);

[bwfim1,level1]=fuzzy_c_means_cluster(fim,1);

These levels of clustering also help in separating the fore-ground and back-ground of the digital tissue image. After the images are clustered they are segmented using the fuzzy segmentation techniques.

imhist(I); imhist(I,n); imhist(X,map)

The above mat-code is used to calculate the histogram for the intensity image' I' and displays a plot of the histogram.

2. **Fuzzy C-Means clustering and segmentation:**

The fuzzy C-Means segmentation algorithm has been used to segment the tissue image as it is found to be the best way of segmenting the fine outlines and boarders of the digital tissue image[8]. The fuzzy clustering is described by the fuzzy matrix M with 'n' rows and 'c' columns. The objective function of the fuzzy algorithm is to simplify the following equation.

$j_m \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^{m} d_{ij}$

$d_{ij} = |x_i - y_j|$

Where, $d_{ij}$is the Euclidean distance from the object $I_X$to the cluster center Y_J and J_m is the object function of the fuzzy algorithm.

Once the images are segmented using the fuzzy C-Means segmentation algorithm the images are processed to extract the features from the histogram of the image generated which will be used to compare against the features of the trained data-set of features.

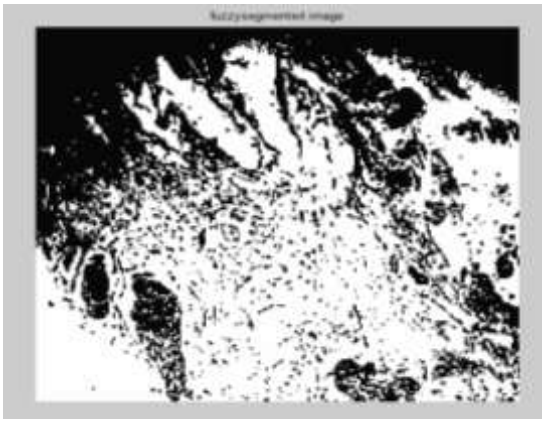

Fig.3 Tissue image segmented using fuzzy logic

3. HOG- Histogram Oriented Gradient:

Histogram : The graphical representation of pixel intensity of a given image. The intensity level is gradually changes based on three parameters like, Angle, magnitude and gradient features.

Angle is defined as in image in two directional views namely X and Y directions.

X represents in $\mathbf{I_x}$

Y represents in$\mathbf{I_y}$

Angle is =a(tan(I_X/I_Y));

Gradient : The samples of pixels varied in two types same intensity level and different values of intensity. Represents, sqrt(I_X^2 + I_Y^2).

Magnitude : Saturation of contrast level among different region in image.. It can be obtain through angle and magnitude values. For that we take 8 x 8 patch and analyze the values. Get back the overall magnitude features.

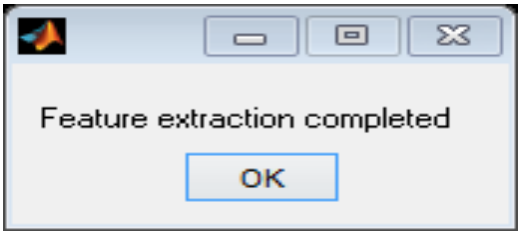

Fig 4 pop-up showing that histogram features are extracted

Once the histogram features are extracted from the digital tissue image they are classified using the 'SVM' classification algorithm to compare against the trained data-set of digital tissue images of ant memory size.

4. Hyper-plane:

A hyper plane is plotted in a graph when the extracted features are to be separated in the 'SVM' classifier to use it for further classification[9]. Its explanation and plotting are explained below briefly.

Once the features are extracted and the histogram is plotted basing two features namely, magnitude and gradient and they are separately plotted in a hyper-plane graph that us shown below. The features that are represented in the below graph are classified like this: i) blue star- histogram of gradient ii)red circle- histogram of magnitude.

In the below figure it is clearly shown that the hyper plane is drawn by considering the nearest distance of features gradient and magnitude. The plane 'A' is drawn by considering the nearest distance from the feature 'Histogram of magnitude' and 'B' is drawn by considering the nearest distance from the feature 'Histogram of gradient'[10]. Now the original hyper plane 'C' is drawn by taking the equal distances from both the planes A and B that clearly separates the image features using 'SVM' classification.

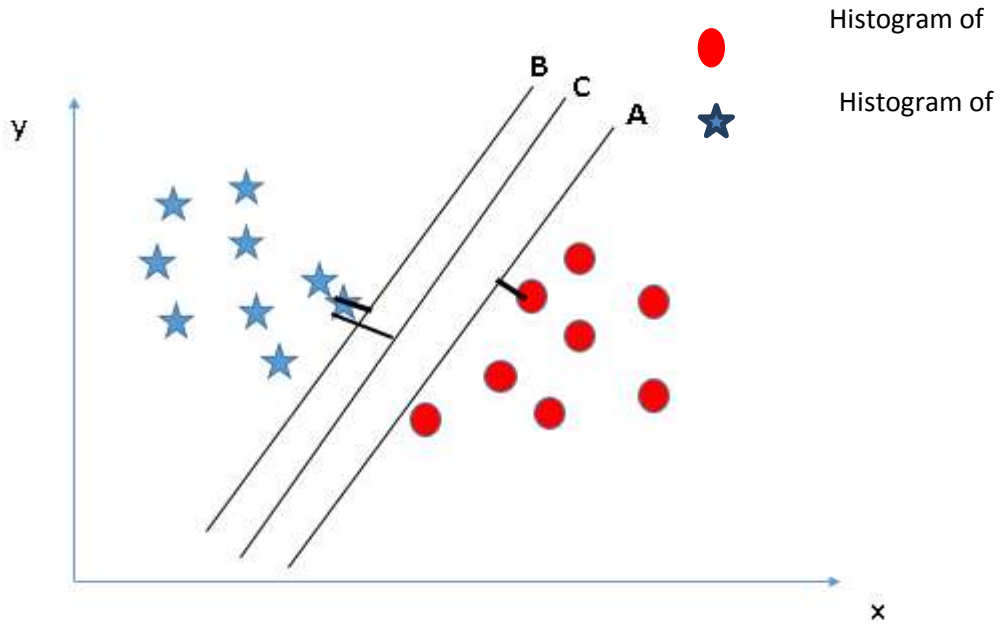

Fig 5 Hyper plane plotting

Cancer Histo-pathology:

The cancer histo-pathology unveils a large image data that is very much helpful in learning the image features so as to train the image data-base that is used in detecting the presence and type of cancer using the machine learning algorithms. A sample image of cancer histo-pathology is given in the following figure.

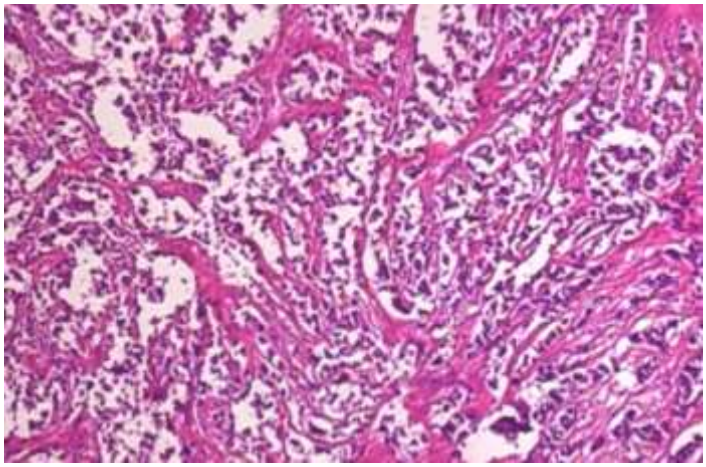

Fig.6A sample tissue image (histo-pathology)

5. SVM Classifier:

A SVM classifier learns a separating hyper-plane between two classes which maximizes the `margin' - the distance between the hyper-plane and the nearest data point of each class[11]. The appeal of SVMs is twofold.

This type of classifier mainly doesn't need any complicated or tuned arguments and secondly it unveils a great propensity to generalize a small training collection. They are mainly manageable of getting trained to high-dimensional spacing. Appendix gives a short description of these classifiers. The only parameters needed to tune a SVM are the `capacity' and the choice of kernel. Error tolerance and generalization capability of SVM during the trained samples' classification are controlled by the SVM 'capacity'.

A well generalization of testing samples is not possible even for a high capacity SVM which could correctly classify all training samples. An SVM classifier constructed and tuned for mainly training the samples is also unable of generalizing system presented testing(query) samples. Conversely, an accurately sufficient data (samples) cannot be produced by a low capacity SVM classifier. An incorrect classification of testing and training samples could possibly detain the performance of SVM in such cases.

For tuning the SVM parameters an extra evaluation data set is available in different experimental conditions. Tuning the SVM parameters can be made easy in such conditions. In case if the evaluation data-set is not available tuning gets harder and the need for alternates comes to. One such alternate is to performing cross-validation, which is computationally very expensive. Another alternative is to avoiding the search of optimal kernels and use one that is sufficiently general. In our case we use the Gaussian kernel for all classifiers which guarantees that the Gaussian distance is in a valid range. In terms of capacity we choose a value that is close to the absolute maximum kernel distance, in our case 1.0. The numerical stability of the SVM algorithm is guaranteed by this choice providing accurately sufficient generalization.

Besides, the false negative values are controlled by setting an additional tuning step to this process. We find and fix a threshold value through our implementation which guarantees that the values returned by SVM that are bigger than this are not false negatives[12].

The possibility of a query protein belonging to any particular structural class can be determined by testing it with the same SVM created for that class. The distance between test feature vector and margin are represented by a "score" produced by the SVM classifier. As the score is bigger the vector is further away from the margin and the classifier is more confident and trust0worthy on its output.

If the score is below the threshold set above, we classify the vector as belonging to that particular class. Otherwise, it is classified as not belonging to the class.

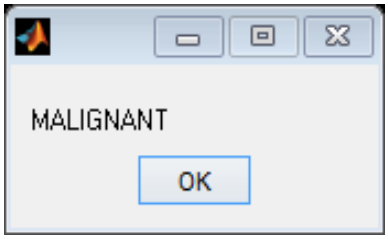

Fig.7 Tumor classification

Algorithm: SVM Classification
dataset benign (0,255,0), malignant (255,0,0);
for (int i = 0; i <image_rows; ++i)
for (int j = 0; j <image_cols; ++j)
{Mat query.img = (Mat_<float>(1,2) <<i,j);
floatmatch = SVM.predict(query.img);
if (match == 1)
image.at<dataset>(j, i) = benign;
else
if (match == -1)
image.at<dataset>(j, i) = malignant;}

The work flow of the entire process is explained in the below given simple block diagram. It not only explains the work flow but also the flow followed for training the image data-base. The algorithms used for training and classification are also specified in it.

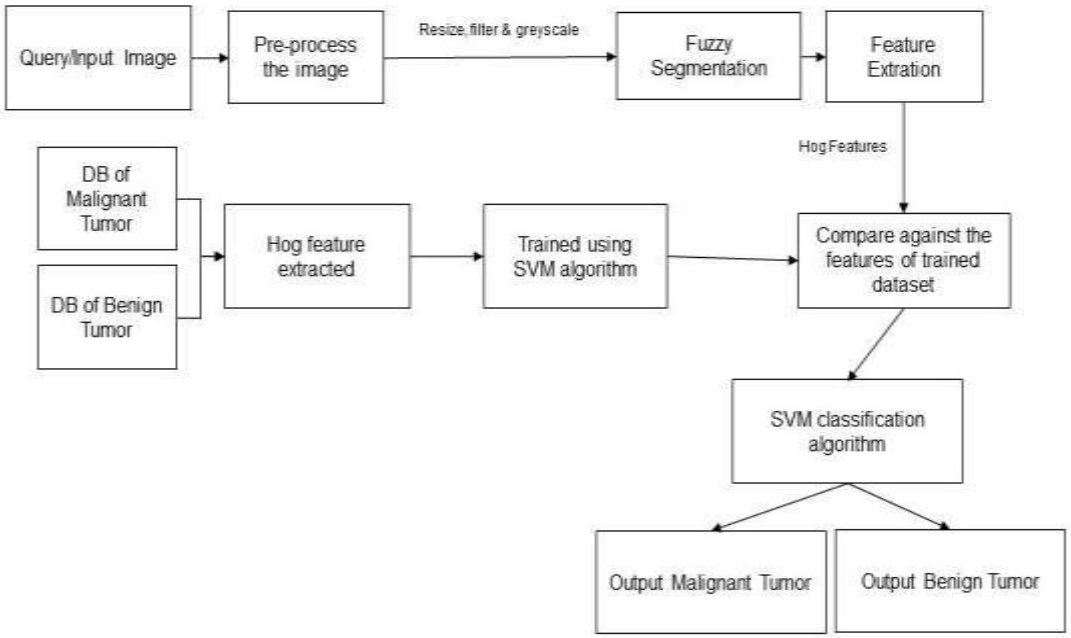

Fig.8 work flow diagram

Training the data-set:

The images are pre-trained before they are compared against the query image to detect cancer using the features extracted. The training process starts by first filtering the noise and proper images and saving them separately in some file and extracted from there for further processing. The proper images are trained using the SVM Training algorithm using the matlab code.

function [svm_struct, svIndex] = svmtrain(training, groupnames, varargin)

SVMSTRUCT = SVMTRAIN(TRAINING, Y) trains a support vector machine (SVM) classifier on data taken from two groups. Y is a column vector that contains the knownclass labels for TRAINING. Each element of Y specifies the group thecorresponding row of TRAINING belongs to. TRAINING and Y must have thesame number of rows. SVMSTRUCT contains information about the trainedclassifier, including the support vectors, which are used by SVMCLASSIFYfor classification. SVMTRAIN treats NaNs, empty strings or 'undefined' values as missing values and ignores the corresponding rows inTRAINING and Y.

The images thus trained are extracted as labels and stored in some mat-file. While we try to compare the data-base image features to query image the SVM classification algorithm these labels are called by setting the path of target mat-file where they are stored to classify the type of cancer.

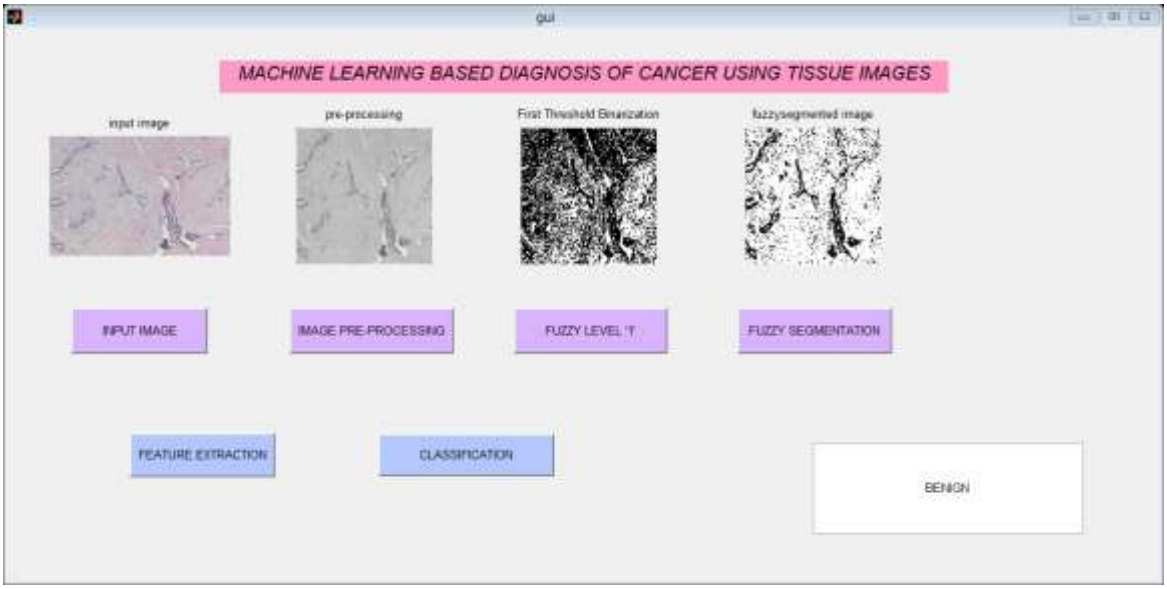

Fig.9 GUI developed showing the process of tissue classification

The training and classification algorithms make use of labels extracted from image features, but not the images directly. The features considered here are '*magnitude and gradient*'of the image histogram generated.

Conclusion:

In this paper we have attempted in applying deep learning concepts which was quite successful in image classification. Support vector Machines was applied in classifying the benign and malignant tumors using the cancer tissue images. Deep learning concept is applied to extract the various features in every hidden layer to improvise on the performance of the network. The performance of classification has improved considerably when deep learning algorithm is applied. In future convolution neural networks(CNN) and Restricted Boltzman's Neural Networks(RBN) for further improvising on the prediction of tumor using the tissue images.

References:

1. Chien-Chi Chen –"An introduction to image segmentation", Graduate Institute of Communication Engineering-National Taiwan University, Taipei, Taiwan, ROC.
2. Mehmet Sezgin , Bulent Sankur Tubı˙tak Marmara Research Center- Information Technologies Research Institute Gebze, Kocaeli Turkey, "Survey over image thresholding techniques and quantitative performance evaluation" published in Journal of Electronic Imaging 13(1), 146–165 (January 2004).
3. J. S. Weszka and A. Rosenfeld, ''Threshold evaluation techniques,''IEEE Trans. Syst. Man Cybern. SMC-8, 627–629 (1978).
4. A. Rosenfeld and P. De la Torre, ''Histogram concavity analysis as an aid in threshold selection,'' IEEE Trans. Syst. Man Cybern. SMC-13, 231–235 (1983).
5. J. Weszka and A. Rosenfeld, ''Histogram modification for threshold selection,'' IEEE Trans. Syst. Man Cybern. SMC-9, 38–52 (1979).
6. L. Halada and G. A. Osokov, ''Histogram concavity analysis by qua-icurvature,'' Comp. Artif. Intell. 6, 523–533 (1987).
7. S. C. Sahasrabudhe and K. S. D. Gupta, ''A valley-seeking threshold selection technique,'' Comput. Vis. Image Underst. 56, 55–65 (1992).
8. Mahesh Yambal , Hitesh Gupta Patel College of science & technology, Ratibadh, Bhopal , Head of the department (CSE), Patel college of science & technology, Ratibadh, Bhopal-Image Segmentation using Fuzzy C Means Clustering: A survey--International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013.

9. Udo von Toussaint-'General Hyper-plane Prior Distributions Based on Geometric Invariances for Bayesian Multivariate Linear Regression' –published in "Entropy" ISSN: 1099-4300-Entropy 2015, 17, 3898-3912; doi:10.3390/e17063898.
10. O'Hagan, A. Kendall's Advanced Theory of Statistics, Bayesian Inference, 1st ed.; Arnold Publishers: New York, NY, USA, 1994; Volume 2B.
11. Support Vector Machines for Machine Learning by Jason Brownlee on April 20, 2016 in Machine Learning Algorithms.
12. Support Vector Machines for Classification and Regression by Steve R. Gunn Technical Report Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science -10 May 1998.

*****