Kidney Tumor Segmentation and Classification on Abdominal CT Scans

Bansari Shah
Student of Computer Engineering
K J Somaiya COE
Vidyavihar

Charmi Sawla
Student of Computer Engineering
K J Somaiya COE
Vidyavihar

Shraddha Bhanushali
Student of Computer Engineering
K J Somaiya COE
Vidyavihar

Poonam Bhogale
Assistant Professor of Computer Engineering
K J Somaiya COE
Vidyavihar

ABSTRACT
This paper, deals with systematic study of simple segmentation and classification algorithms for kidney tumor using Computed Tomography images. Tumors are of different types having different characteristics and also have different treatment. It becomes very important to detect the tumor and classify it at the early stage so that appropriate treatment can be planned. This CT scans are visually examined by the physician for detection and diagnosis of kidney tumor. However this method lacks accuracy and detection of size of the tumor. So to overcome this, a computer aided segmentation technique has been proposed, which extracts the tumor part from the kidney, further on which feature extraction method is performed for extracting certain features and the type of tumor i.e. malignant or benign is displayed by using simple classifiers .

General Terms
Algorithms, Kidney Tumor, Computed Tomography scans, Process.

Keywords
Pre-processing, Fuzzy C-means, Grey Level Co-occurrence Matrix, K Nearest Neighbour classifier, Support Vector Machine classifier

1. INTRODUCTION
Human body consists of myriad number of cells[5]. For a body to remain healthy, cells grow and divide in orderly fashion. When cell growth becomes uncontrollable the extra mass of cell transforms into tumor. CT scans and MRI are used for identification of tumor[5]. The goal of the system is to detect tumor by incorporating image processing, pattern analysis, and computer vision techniques for enhancement, segmentation and classification for kidney diagnosis, as shown in Fig 1.. This system can be used by radiologists and healthcare specialists[1,2,9,10]. The system is expected to improve the sensitivity, specificity, and efficiency of kidney tumor screening. Tumor segmentation is done by fuzzy c-means algorithm , extraction of features is implemented by grey level co-occurrence matrix method (GLCM) and finally classification of tumor if it is benign or malignant is obtained by support vector machine tool (SVM) and k nearest neighbour classifier (KNN).

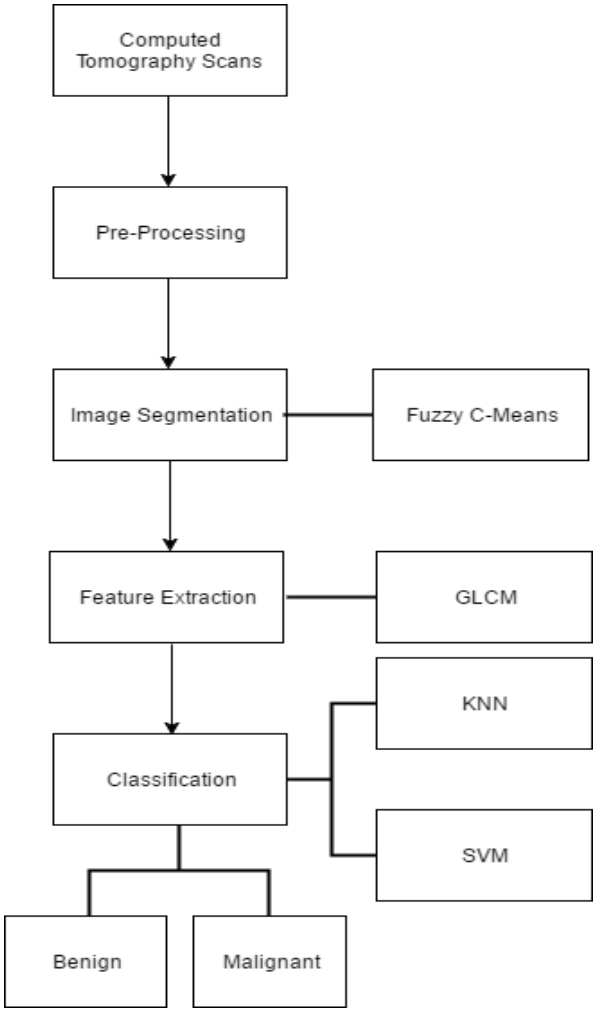

Fig 1: Block Diagram Of Proposed System

2. SEGMENTATION
2.1 Fuzzy C-Means Clustering Algorithm
A. Fuzzy Clustering

The fuzzy logic is used to process data by giving the partial membership value to each pixel in the image. The membership value of the fuzzy set ranges between 0 and 1. Fuzzy clustering is a multi valued logic system that uses

intermediate values i.e., member of one fuzzy set can also be member of another fuzzy set while in the same image. There are no discontinuous or sudden transitions between full membership and non-membership functions. The membership function defines the fuzziness of an image and the information contained in the image. The primary features involved in characterization using a membership function are: core, support, and boundary. The core is completely a member of the fuzzy set. The support is non membership value of the set and boundary is the partial membership value, having its value between 0 and 1. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center, more is its membership towards the particular cluster center. Hence, addition of membership of each and every data point must be equal to one.

B. Mathematical representation

Algorithmic steps for Fuzzy c-means clustering [2]

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the fuzzy membership 'µij' using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3) Compute the fuzzy centers 'vj' using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

4) Repeat step 2) and 3) until the minimum 'J' value is achieved or ||U(k+1) - U(k)|| < β.
where,
'k' is the iteration step.
'β' is the termination criterion between [0, 1].
'U = (µij)n*c' is the fuzzy membership matrix.
'J' is the objective function.

C. Screenshot for Fuzzy C-Means Clustering

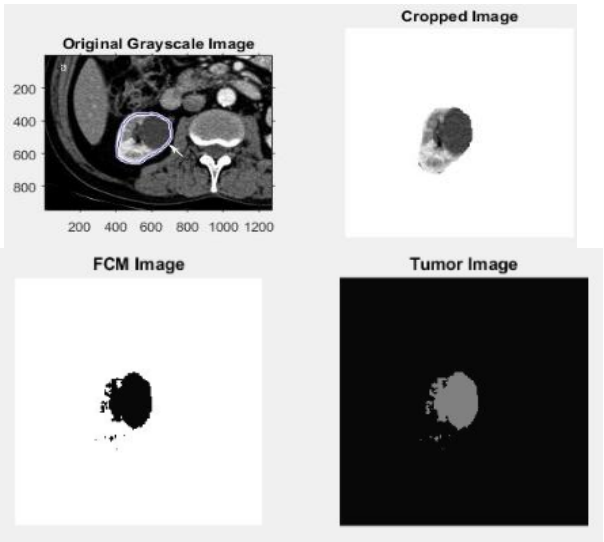

Fig 2: Output image for segmentation using FCM.

Fig.2 is the image of tumor obtained after cropping kidney from abdominal scan and applying Fuzzy C-Means Clustering to it. The Fuzzy C-Mean algorithm clusters the image according to some characteristics.

3. FEATURE EXTRACTION

3.1 Grey Level Co-occurrence Matrix

A.GLCM

GLCM is a widely used method for medical image analysis, classification. This method gives us information about relative position of two pixels with respect to each other. The GLCM is then created by counting the number of occurrences of pixel pairs at a certain distance. To compute the GLCM matrix for an image f (i, j), a distance vector d=(x, y) is defined. The (i,j)th element of the GLCM matrix P is defined as the probability that grey levels i and j occur at distance d and angle θ, then extracting texture features from GLCM matrix P. Four angles(0,45,90,135)and four distances(1,2,3,4) can be used to calculate the co-occurrence matrix as shown in Fig 3.

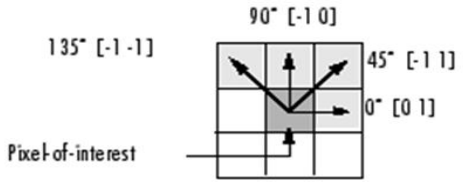

Fig 3:Calculation of co-occurrence matrix in GLCM.

B. Expressions of GLCM descriptors are: [7]

1) Correlation defined by Eq. (1)

$$= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j) X (i X j) - (\mu_X X \mu_y) / \sigma_X \sigma_y \quad ...(1)$$

2) Contrast defined by Eq. (2)

$$= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j)(i - j)^2 \quad ...(2)$$

3) Energy defined by Eq. (3)

$$= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j)^2 \quad ...(3)$$

4) Entropy defined by Eq. (4)

$$= - \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i, j) \log(P(i, j)) \quad ...(4)$$

5) Homogeneity defined by Eq. (5)

$$= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{p(i, j)}{1 + |i - j|} \quad ...(5)$$

6) Peak Sound To Noise Ratio(PSNR)

The PSNR value calculatesthe peak signal-to-noise ratio, in decibels, between two images. This ratio is then used as a quality measurement between the original image and a compressed image. The higher the PSNR value, better is the quality of the compressed or reconstructed image.

The Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are the two error metrics that are used to differentiate amongst two image compression qualities. The

MSE value represents the cumulative squared error between the compressed and the original image, and the PSNR value represents measure of the peak error. The lower the value of MSE, lower is the error.

To compute the PSNR, the block first calculates the mean-squared error using the following Eq.(6):

$$MSE=_{M,N}[I_1(m,n)-I_2(m,n)]^{\wedge}2/M*N \quad ...(6)$$

In the previous equation, M and N are the number of rows and columns in the input images, respectively. Then the block computes the PSNR using the following Eq. (7):

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \quad ...(7)$$

4. CLASSIFICATION

4.1 K-Nearest Neighbour

A. KNN

The k-nearest neighbor is a semi-supervised learning algorithm. It requires training data and a predefined k value to find the k nearest data based on distance computation. If k data have different classes, the algorithm predicts class of the unknown data to be the same as the majority class.

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x1, x2, ..., xmx, and my-by-n data matrix Y, which is treated as my(1-by-n) row vectors y1, y2,..., ymy, the various distances between the vectors xs and yt are defined as follows:

1. Euclidean Distance

The Euclidean distance is a measure to find distance between two points [8], defined by Eq. (8)

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)' \quad ...(8)$$

2.Hamming Distance Hamming distance, which is the percentage of coordinates that differ [8], can be defined by Eq.(9)

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj})}{n}\right) \quad ...(9)$$

3.Cosine Distance The Cosine distance is computed from one minus the cosine of the included angle between points [8], defined by Eq. (10)

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right) \quad ...(10)$$

4. City Block Distance[8]

The city block distance between two points is the summation of the absolute difference of Cartesian coordinates, defined by Eq. (11)

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \quad ...(11)$$

5. Correlation Distance

Distance based on correlation is a measure of statistical dependence between two vectors [8], defined by Eq. (12)

$$d_{st} = \left(1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}}\right)$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}$$

...(12)

B. Empirical Study Methodology In this section is presented, a study framework using k-nearest neighbor algorithm with various distance metrics. The framework is shown in Fig4.

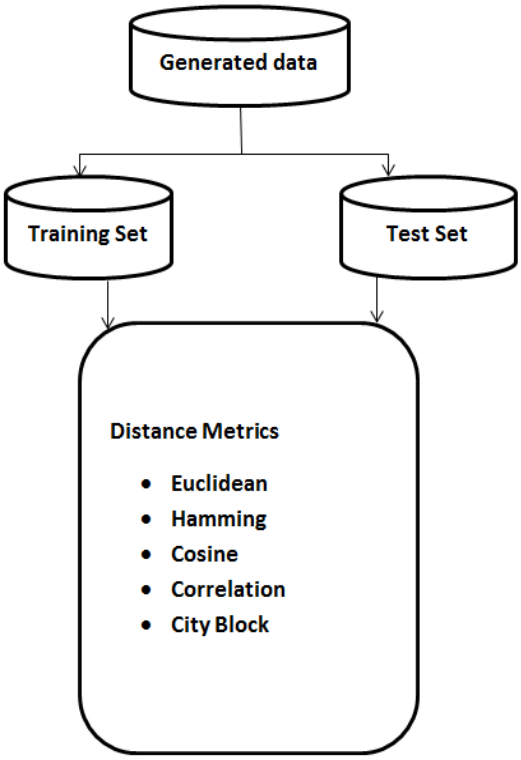

Fig 4:The framework of our empirical study.

From Fig. 4 the detail of each step can be explained as follows:

Step 1: Generate a binary data set with several different distribution and different amount of data in each class. Step 2: Use data from step 1 for data classification by applying the k-nearest neighbor algorithm with various distance metrics to compute the k-nearest data points for making classification. Step 3: Analyze the results and conclude about the performance of classification using various distance metrics, as seen for classifying tumor as malignant or benign in Fig 5.

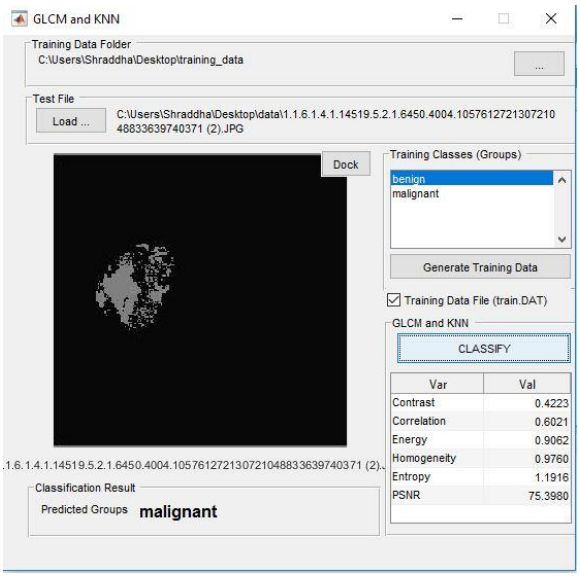

Fig 5:Classification using KNN

4.2 Support Vector Machine (SVM)

Support vector machines (**SVMs**) are a type of supervised learning models along with associated learning algorithms that analyze data and recognize various patterns, used for classification analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes, malignant and benign forms the output, making it a non-probabilistic binary linear classifier. Now that there are set of training examples at hand, each marked as belonging to one of two categories, an SVM training algorithm constructs a model that assigns new examples into one category or the other [8]. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Newer examples are then plotted into it and then predicted to belong to a category based on which side of the gap they fall on.

More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. A clear segregation of the two types of data points in Fig 6 is seen, showing the hyperplane generated using SVM.

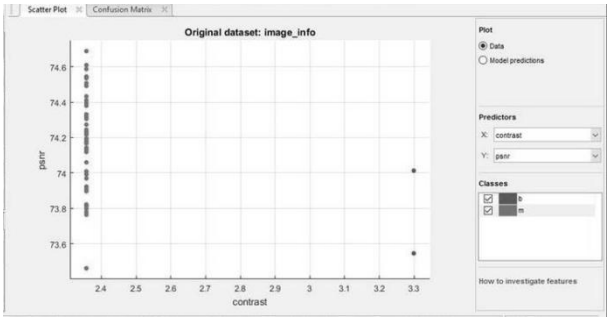

Fig 6: Scatter Plot of SVM.

5. CONFUSION MATRIX

The output of the computing done via SVM is mapped via the confusion matrix. A confusion matrix consist of information about actual and predicted classes done by a classification system. Performance of such systems is evaluated using the data in the matrix. The following table in Fig 7 shows the confusion matrix for a two class classifier.Confusion Matrix helps in detecting the accuracy of datasets.

The data in the confusion matrix as shown in Fig 7 have the following meaning in context.

a is the number of correct predictions that an instance is negative,

b is the number of incorrect predictions that an instance is positive,

c is the number of incorrect of predictions that an instance negative, and

d is the number of correct predictions that an instance is positive.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Fig 7:Confusion Matrix

6. CONCLUSION

Various segmentation and classification of kidney and kidney tumor methods for CT image have been discussed in this paper. This performance study reveals that semi automated segmentation method reduces errors occurring while doing manual segmentation. Experimental results were obtained by using matlab software. In this paper improvement in the accuracy of results are obtained by incorporating noise removal steps while doing fuzzy c means clustering. Final stage in tumor detection is to decide whether it is cancerous or non-cancerous. This classification is done by designing an artificial neural network using KNN, and also implemented using SVM algorithm. The accuracy is mapped using the confusion matrix, and the outcomes are compared. Improvement in algorithm is done by training the network with more image features and by doing rigorous testing.

7. REFERENCES

[1] Anton Bardera, Jaume Rigau, Imma Boada, Miquel Feixas, and Mateu Sbert, " Image Segmentation Using Information Bottleneck Method ",Page Number 1601-1612, IEEE Transactions on Image Processing, Vol. 18, No. 7, July 2009.

[2] Mr.Rohit S. Kabade and Dr. M. S. Gaikwad, "Segmentation of Brain Tumour and Its Area Calculation in Brain MR Images using K-Mean Clustering and Fuzzy CMean Algorithm " IJCSET Vol. 4 No. 05 May 2013

[3] J.selvakumar ,A.Lakshmi and T.Arivoli."Brain Tumor Segmentation and Its Area Calculation in Brain MR Images using K-Mean Clustering and Fuzzy C-Mean Algorithm"IEEE- ICAESM -2012 March 30, 31, 2012

[4]Robert.L.Cannon;Jitendra.V.Dave;James.C.Bezdek"Efficient Implementation of the Fuzzy c-Means Clustering Algorithms"IEEE TRANSACTFIONS ON PATTERN

ANALYSIS AND MACHINE INTELLIGENCE. VOL. PAMI-8, NO. 2, MARCH 1986

[5]Kuldeep Pawar , Jayamala.K.Patil "A Systematic Study of Segmentation Methods For Detection of Kidney Tumor Using Computed Tomography Images."International Journal of Engineering Sciences & Research Technology Nov., 2012

[6] Mredhula.L and Dr.M.A.Dorairangaswamy "Detection And Classification of Tumors in CT Images" IJCSE Vol. 6 No.2 Apr-May 2015

[7]U Akilandeswari, Nithya Rajendran and B Santhi, "Review on Feature Extraction Methods in Pattern Classification", European Journal of Scientific Research, Vol.71 No.2 (2012).

[8] Kittipong Chomboon* , Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak Kerdprasop, Nittaya Kerdprasop, "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm ", International Conference on Industrial Application Engineering 2015

[9]Jian Wu, Feng Ye, Jian-Lin Ma, Xiao-Ping Sun, Jing Xu, Zhi-Ming, " The Segmentation and Visualization of Human Organs Based on Adaptive Region Growing Method" ,Page Number-439-443, IEEE 8th International Conference on Computer and Information Technology Work shops978-0-7695-3242-4/08,IEEE,2008.

[10] Rafael C. Gonzalez, Richard E.Woods, " Digital Image processing", published by Pearson Education, Inc. 2002.