Automation of Radio Diagnosis for the Identification of Cancer Using Support Vector Machine

Ganesh N
Information Science and Engineering
R V College of Engineering
Bangalore, India
ganeshn1411@gmail.com

Prof. G S Mamatha
Associate Professor, Information Science and Engineering
R V College of Engineering
Bangalore, India
mamatha.niranjan@gmail.com

Abstract—**Radio Diagnosis employs the technique of analyzing the images obtained through the process of scanning a patient's body and identifying the disease. The paper aims at providing an automated mechanism for this analysis using the technique of Support Vector Machine (SVM). The DICOM images obtained from the scan serves as the input for the classification process. The paper uses the Hounsfield coefficient unit of each tissue to classify the image. The classified values are then compared with that of a normal human body. A deviation of a set of values in a particular position indicates a tumor. The Hounsfield Coefficient Value of the particular set of deviated cells identifies the type of cancer.**

Index Terms—**Classification, DICOM images, Hounsfield Coefficient, Radio Diagnosis, Support Vector Machine**

I. Introduction

Radiology is field of medicine that uses imaging for both diagnosis and treatment of diseases within the human body [1]. As this field involves the use of very sophisticated equipment, it is one of the costliest fields of medical science. This limits the scope of Radiology in remote places especially in developing and the underdeveloped world. The very fact that the number of people needing radiology is not met with the number of existing radiologists and is not reachable in remote areas for military applications serves as the purpose of this project. Radiology integrates technology and clinical medicine. The ability to produce images of the human body using various techniques has changed the practice of medicine. Radio Diagnosis is the diagnosis of x-rays or in a broader sense, diagnostic imaging, that includes CT, ultrasound, and magnetic resonance. Radiologist has the skills in performing and interpreting diagnostic imaging tests that involve the use of various imaging equipment.

About two thirds of the world's population has little or no access to radiological services. In South Africa, comparatively one of the better staffed countries in Africa, most of the hospitals in the public sector have never had a radiologist [2]. Taking a peek at statistics of India, a country with a billion people has around 5,500 radiologists, a severely imbalanced ratio of 1:196,000 (the corresponding ratio in the US is being 1:11,000) [3]. These figures show that there is acute need for radiologists or at least an alternative for them. The paper aims at providing the latter.

SVM is one of the most popular classification algorithms used worldwide. SVM and some of its variations are extensively used in the classification of medical images. Yun Jiang et.al, [4] uses SVM coupled with Rough Set Theory (RST) called improved support vector Machine (ISVM) to classify digital mammography to achieve 96.56% accuracy. Chi-Hoon Lee et.al, [5] proposes a method for segmentation of Brain tumors using Random Fields and SVM. El-Naqa I et.al, [6] uses SVM for the detection of micro calcification (MC) clusters produced in digital mammograms and observes SVM outperforming other well-known classification algorithms with an accuracy of 94%. Yong Fan et.al, [7-8] uses SVM for medical images to rank computed features from the extracted regions and classification using the best set of features with an accuracy rate of 91.8.

The rest of the paper is organized as follows. Section 2 introduces DICOM images which is the input for the system. It describes Hounsfield coefficient obtained from these images. Section 3 describes the experimental work carried out to demonstrate the project along with the methodology adopted. Section 4 elaborates on the results achieved and the inference obtained. Section 5 includes the conclusion with a note on the future work to be carried out.

II. Hounsfield Coefficient

All medical images are obtained in the form of DICOM images. In medical imaging, DICOM (Digital Imaging and Communications in Medicine) is a standard for handling, storing, printing, and transmitting information [9]. Unlike normal image formats, DICOM images have the capability to capture data from a 3D perspective and provide multi-dimensional images. This is very helpful for medical imaging to capture data from various perspectives during the process of scanning.

Every tissue within the DICOM image is differentiated using grayscale color codes using the Hounsfield coefficient value. The Hounsfield Coefficient/Unit (HU) scale is a linear transformation of the linear attenuation coefficient measurement into one in which the radio density of distilled water (at standard pressure and temperature) is defined as zero HU, while the radio density of air at STP is defined as -1000 HU [10].

$$HU = \frac{(\mu_x - \mu_{water})}{(\mu_{water} - \mu_{air})} \times 1000 \qquad (1)$$

III. Experimental Work

An experiment was conducted to classify the images for the identification of tumor and understand the efficiency of the recognition. The experiment was conducted on 3 data sets containing scanned brain images with 2 containing tumors and the normal image as the training data set.

Fig. 1 shows the block diagram of the entire process starting from data accusation from the DICOM images up till the results of the classification algorithm.

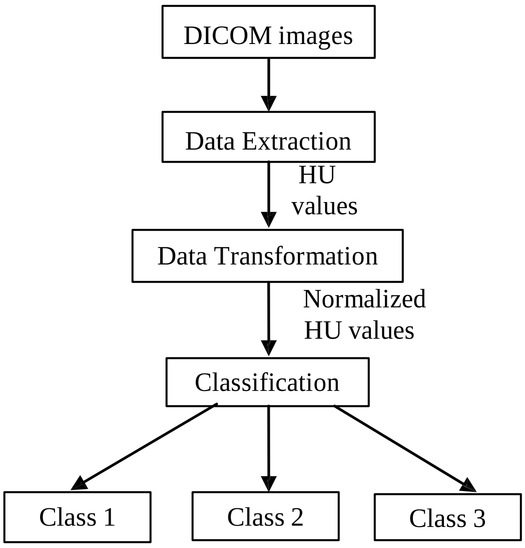

Fig.1. Block diagram of the system

A. Preprocessing

Prior to obtaining the data from the images, some preprocessing needs to be performed. This involves determination of standard images that must be stored in the database to be compared with the input images. The images of a healthy human body to be considered as the standard images are to be grouped into various categories based on age group as shown in Table 1. There are totally 10 different age groups to be considered. Along with age group, gender and origin are other factors to be considered as human body structures differ based on these parameters. In total, a set of 60-70 standard images need to be stored in the database for effective analysis of a varied set of input images. For each of the 10 age groups considered the standard image is framed using the mode of that particular group.

$$\text{Mode } (x - y) = \text{most repeated value in the range} \qquad (2)$$

Table I. Different age groups and mode standard images

Age Group	**Mode**
Below 5	2
5-10	7
10-18	15
18-25	21
25-35	28
35-45	41
45-55	49
55-65	58
65-75	71
75 and above	80

The mode values given in the table indicate the most common age of a disease in the particular age group. Consideration of mode value rather than mean or median comes with the advantage that the most common age will match most of the ages of the input images, thus, leading to more efficient results.

B. Data Extraction

The first step in the system is acquisition of data. DICOM images obtained from the scan are used as input for the system. DICOM images are coupled with some information regarding the patient to clearly separate one image form the other. This avoids the possibility of mismatch of images among the patient records. Once these images are obtained, the Hounsfield coefficient values of every pixel is obtained and stored in a CSV file. The CSV file contains Hounsfield coefficient values of different pixels against its position across 3 dimensions x, y and z.

C. Data Transformation

Once the data is acquired from DICOM images, the values need to be transformed to match the standard image before classification. The dimensions of the appropriate standard image are determined and the image is reconstructed according to theses dimensions. The relative positions of different tissues within the image are used to match to the standard image. These form the set of normalized Hounsfield coefficient values which can be further passed to the classifier.

D. Training

The machine needs to be trained in order to effectively classify various HU values into different classes. The machine is trained for 100 training set reading of a normal human image before any test of classification is carried out. The data obtained from the DICOM images which are stored in a CSV file is processed for data transformation by reducing dimensions. This is done by calculation the mean and standard deviations for each set of readings. The reduced dimensions for the data lower the memory cost and the algorithm becomes faster. To ensure that a good data set is acquired, the process of training will be repeated for the same object multiple times with the same user as well as different users.

E. Classification

Classification is the process of grouping similar data into a common group and differentiating the dissimilar ones. Supervised learning algorithms obtain a function or a mapping between the input and the output that can predict an output for new inputs. One of the most popular machine learning algorithms used in supervised learning is SVM (Support Vector Machine). SVM, unlike most classification algorithms maps the data into a multidimensional space. This data is further separated with the help of hyper planes separate the dissimilar values [11].

The 3 dimensions representing the position of each point derives a 3 dimensional space for SVM as shown in Fig 2. Each HU value is plotted as a point in this 3-D space. If the different classes of points are linearly separable, linear kernel is used, else Radial Basis Function kernels is used to map samples into another dimensional space. The Maximum marginal hyper-plane is found that differentiates between the two classes and the test data is plotted next and compared to which class it belongs.

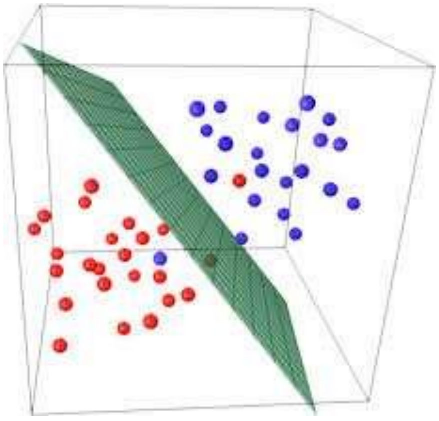

Fig.2. 3-Dimensional SVM plot

IV. Results

The experiment was conducted in order to test the efficiency of the classification algorithm to identify the presence of a tumor. The experiment included 3 sets of DICOM images of the brain. The machine was trained with a data set containing an image of a healthy human body and datasets of both, images that did not contain tumor and those that did were considered for classification. The classification algorithm is analyzed in terms of accuracy, error rate, sensitivity and specificity as described below where TP (True Positive) refers to the correctly classified positive tuples, TN (True Negative) refers to the correctly classified negative tuples, FP (False Positive) refers to the incorrectly classified positive tuples, and FN (False Negative) refers to the incorrectly classified negative tuples.

Table 3 shows the results obtained through the process of classification using SVM. As we can see, the accuracy achieved is quite high. In the case of training data set accuracy is 94.7% and for the two Validation data sets considered, accuracy is approximately 90% which is comparatively higher than other standard classification algorithms.

$$Error\ Rate = \frac{FP + FN}{P + N} \tag{2}$$

$$Sensitivity = \frac{TP}{P} \tag{3}$$

$$Specicficity = \frac{TN}{N} \tag{4}$$

$$Accuracy = Sensitivity\frac{P}{(P + N)} + Specificity\frac{N}{(P + N)} \tag{5}$$

Table II. Result Analysis

	Training Data Set	**Test Data Set I**	**Test Data Set II**
Sensitivity (%)	94	91	88
False Negative (%)	6	9	12
Specificity (%)	96	93	93
False Positive (%)	4	7	7
Accuracy (%)	94.7	91.6	89.7
Error Rate (%)	5.3	8.4	10.3

V. Conclusion and Future Work

The paper proposes a technique that can automate the diagnosis of tumor cells from scanned images. As there is a need for an alternative to radiologists due the scarce number of radiologists and the huge demand to the field, an automated system to diagnose the images after scanning, can not only help in radiology reach out to the masses who currently have no access to radiological treatments but also supplement the doctors with their diagnosis. The paper uses 3 datasets for the process of classification as shown in Table 3. The results obtained through classification using SVM show superior performance with high sensitivity, specificity and accuracy values. These values indicate an opportunity that can revolutionize the way of Radio Diagnosis is performed. The paper currently considers only 3 data sets of brain images as the input which will be expanded to the complete human body in the future course. A major future scope would be to connect these automated systems to each other across the Internet. The system could be implemented in client-server architecture. While the centralized server could act as the database containing various standard images, the client machine could just receive the data of the particular patient and send it to the server for processing. This could further enhance processing by comparing inputs from various clients. This could connect the unconnected group of standalone machines over the Internet.

Acknowledgment

We would like to thank K R Hospital, Mysore and Nimhans, Bangalore for their inputs with regard to DICOM images. I am thankful to Dr. Cauvery N K, HOD, Information Science and Engineering for providing the lab facilities and other resources required to complete this project. I am grateful towards Dr. B.S. Satyanarayana, Principal, RVCE for constant encouragement and facilitates extended in completion of this work.

REFERENCES

[1] Harjit Singh, Janet Neutze, Radiology Fundamentals: Introduction to Imaging and Technology, 4th ed., Springer New York, 2012, pp. 7-10.

[2] Sarah Jersild, “Radiologist sightings drop around the world”, 2003 Special Edition, Diagnostic Imaging, July 2003, pp. 11-14.

[3] Phaneendra K Yalavarthy, “Medical Image Informatics: Challenges”, Proc. of Indo-US Workshop on Large Data Analytics and Intelligent Services, 2011, pp. 4-7.

[4] Yun Jiang, Zhanhuai Li, Longbo Zhang, Peng Sun, “An Improved SVM Classifier for Medical Image Classification”, Proc. of RSEISP International Conference, Warsaw, Poland, 2007, pp 764-773

[5] Chi-Hoon Lee, Mark Schmidt, Albert Murtha, Aalo Bistritz, Joerg Sander, and Russell Greiner, “Segmenting Brain Tumors with Conditional Random Fields and Support Vector Machines”; Proc. of First CVBIA International Workshop, Beijing, China, 2005, pp. 469-478.

[6] El-Naqa I, Yongyi Yang, Wernick M N, Galatsanos N P, “A support vector machine approach for detection of micro calcifications”, IEEE Transaction on Medical Imaging, Vol. 21 (12), 2002, pp 1552 – 1563

[7] Yong Fan, Dinggang Shen, Christos Davatzikos; Classification of “Structural Images via High-Dimensional Image Warping, Robust Feature Extraction, and SVM”; Proc. of 8th International Conference on Computer Science and Information technology, Palm Springs, CA, USA, 2005, pp. 1-8.

[8] Vanitha L, Venmathi A R, “Classification of Medical Images Using Support Vector Machine”, Proc. of International Conference on Information and Network Technology IPCSIT, Vol.4, 2011, pp. 63-67.

[9] Oleg S. Pianykh, “Digital Imaging and Communications in Medicine (DICOM)”, 2nd ed., Springer Berlin Heidelberg, 2012, pp. 3-5.

[10] Dr. Roberto Molteni, physicist, “From CT Numbers to Hounsfield Units in Cone Beam Volumetric Imaging: the effect of artifacts”, Proc. of 62nd International conference on Medical research AAOMR, 2011, pp. 6-10.

[11] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concepts and Techniques”, 2nd ed., Elsevier, 2012, pp. 337-344.