Lung Cancer Detection and Classification Using SVM

Prerana Prajapati, Vedika Hande, Aarti Ingale, Sanjeev Dwivedi

Student, Student, Student, Assistant Professor

B.E.Computer Engineering

Vidyalankar Institute of Technology, Mumbai, India

Abstract : Image processing techniques are widely used in several medical problems for image enhancement in the detection phase to support the early medical treatment. In this research, we aim to improve quality and accuracy of early detection of lung cancer through a combination of image processing techniques and machine learning. The Cancer Imaging Archive (TCIA) dataset has been used for training and testing purpose where DICOM is the primary format used for image storage .Classification is done using SVM (Support Vector Machine) classifier which identifies whether the CT image is cancerous and non-cancerous. Before training the classifier, we are performing image processing techniques on CT images such as converting the image into HU (Hounsfield Unit) scale to get the binary image of lungs followed by nodule segmentation where nodules are detected within the lungs. Further, features are extracted from CT images using GLCM (Gray Level Co-Occurrence Matrix) .The enhanced image is then given to the classifier to provide accurate results. As an enhancement, we have also performed Early Stage Detection for reducing the growing cancer burden. MATLAB image processing toolbox based has been used for the implementation on the CT scan images.

Keywords: Nodules Detection, Image Processing, Classification, Machine Learning

I. INTRODUCTION

Cancer is one of the most dangerous disease that causes death. Lung cancer has become one of the most common causes in both men and women. A large number of people die every year due to lung cancer .The disease has different stages where by it starts from the small tissue tissues and spreads throughout the different parts of the body. According to American Cancer Society, the cases of lung cancer increases very rapidly and almost 14% newly diagnosed cancers are a lung cancer and also the main cause of cancer death worldwide. The previous study of diagnosis showed that the most of the lung cancer patients belongs to the age of 60 years. The process of early detection of cancer plays an important role to prevent cancer cells from multiplying and spreading. Although CT scan imaging is best imaging technique that is reliable for lung cancer diagnosis because it can disclose every suspected and unsuspected lung cancer nodules in medical field. However, variance of intensity in CT images and anatomical structure misjudgments by doctors and radiologists might causes difficulty to interpret and identify the cancer from CT scan images. Therefore computer aided diagnosis can be helpful to assist radiologists and doctors to identify the cancerous cells accurately. Many computer aided techniques using image processing and machine learning has been researched and implemented. The main aim of this research is to evaluate the various computer-aided techniques, analyzing the current best technique and finding out their limitation and drawbacks and finally proposing the new model with improvements in the current best model.

II. PROBLEM STATEMENT

Lung cancer is the leading cause of cancer death and the second most diagnosed cancer in both men and women. An X-Ray image of our lungs may reveal an abnormal mass or nodule. As compared to X-Ray, CT scan can reveal small lesions in our lungs that might not be detected on an X-ray. However, it is still difficult for the doctors or radiologists to accurately detect even the small nodule within our lungs. In this study, we proposed different methods in analyzing lung cancer using image processing techniques along with machine learning algorithm. It is also possible to detect early stage of lung cancer that can improve the overall chances of survival of the patient. To classify the CT scan image whether it is cancerous or non-cancerous we have performed SVM classification algorithm. The main focus of our research is to provide accurate results by the means of machine learning algorithm

III. PROPOSED SYSTEM

Figure 1 shows a general description of lung cancer detection system that contains basic stages:

A. Collecting the dataset
B. Image Processing (Median Filter)
C. Converting Image to HU scale
D. Segmentation (Lungs and Nodules)
E. Feature Extraction
F. Classification
G. Stage Detection

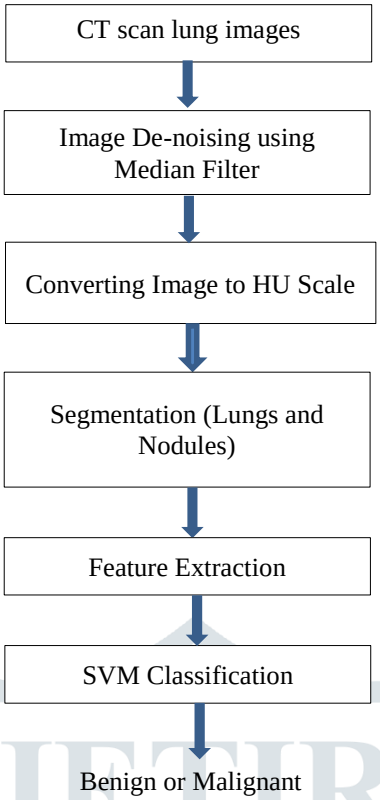

Fig-1:Flowchart of Proposed System

3.1 Collecting Dataset

The Lung Cancer Image Dataset [8] has been chosen from TCIA (The Cancer Imaging Archive) which is an online CT scan image dataset publicly available for the researchers in the field of digital image processing. Images in the dataset are in the DICOM format with size 512*512 pixel. In TCIA dataset, we have used 20% data for testing and remaining 80% of dataset is used for training the SVM classifier.

3.2. Image Processing

Smoothing is an image processing technique used in order to reduce noise in an image to produce clearer image. Most of the techniques are based on low pass linear filters. It is mostly based on the averaging technique of the input image. To perform a smoothing operation we will apply a filter to our image. The most common type of filters is the median filter which is used in our proposed system.

3.3. Converting image to HU Scale

Hounsfield Unit is the unit of measurement used in CT scan images which is a measure of radiodensity. CT scanners are carefully calibrated to accurately measures this.HU scale is a linear transformation of the original linear attenuation coefficient measurement into one in which the radiodensity of distilled water at standard pressure and temperature (STP) is defined as zero HU, while the radiodensity of air at STP is defines as -1000 HU. The HU value is therefore given by

$$HU = 1000 \times \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}}$$

Fig 2. Hounsfield Formula

TABLE 1
Hounsfield Unit

Substance	HU
Air	-1000
Lung	-500
Fat	-100 to -50
Water	0
CSF	15
Kidney	30
Blood	+30 to +45
Muscle	+10 to +40
Grey matter	+37 to +45
White matter	+20 to +30
Liver	+40 to +60
Soft Tissue, Contrast	+100 to +200
Bone	+700(cancerous bone) to +3000(cortical bone)

By default however, the returned values are not in this unit. But this can be fixed. Some scanners have cylindrical scanning bounds, but the output image is square. The pixels that fall outside of these bounds get the fixed value -2000. The first step is setting these values to 0, which currently corresponds to air. Next we convert our image back to HU units, by multiplying with the rescale slope and adding the intercept.

CT_IMAGE_HU=(CT_IMAGE*RESCALE_SLOPE)+INTERCEPT

3.4. Segmentation

Segmentation is a process which splits the image into its constituent regions or objects. Segmentation is usually used to trace objects and borders such as lines, curves, etc. in images. The main objective of segmentation is to simplify and change the representation of the image into something that is more significant and easier to examine.

In our proposed system, Segmentation involves two parts which are Lung Segmentation and Nodules Segmentation.Lung Segmentation is implemented through HU Scale where once the image is converted into HU Scale, we get segmented binary image of lungs. As a precautionary measure, we have also used active contour method to trace the lungs part properly so as to improve the accuracy. Further in Nodules Segmentation, nodules are detected within the binary image by using Morphological Operations such as dilation and erosion. Dilation and Erosion are often used in combination for specific image preprocessing applications such as filing holes or removing small objects. As a result of which, we get segmented nodule image given to feature extraction to perform textual analysis on the image.

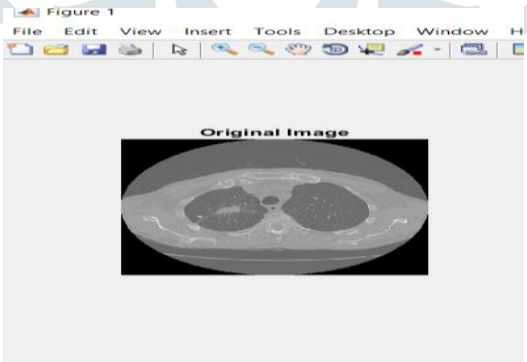

Fig 3. Original Image

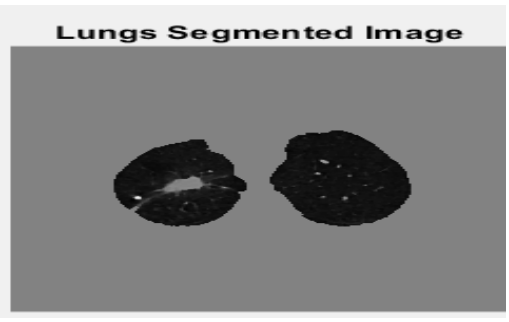

Fig 4. Lungs Segmented Image

Fig 5. Nodule Segmented Image

3.5. Feature Extraction

Feature Extraction stage is a crucial stage for the Computer Aided Diagnosis (CAD) system. It uses different methods and algorithms for feature extraction from the segmented image. The extracted image can be classified as either cancerous or non-cancerous using texture properties. We have used HARALICK GLCM (Gray Level Co-Occurrence Matrix) for texture feature extraction from CT scan images. The GCLM is a tabulation of how often different combinations of pixel brightness values (gray levels) occur in an image. Firstly we create gray-level co-occurrence matrix from image in MATLAB.

Some of the features are extracted using this method are:

- Contrast
- Entropy
- Energy
- Homogeneity
- Correlation

Then these features are used for tumor classification.

3.6. Classification

This stage classifies the detected nodule as malignant of benign. Support Vector Machine (SVM) is used as a classifier. SVM is supervised machine learning algorithm which defines the function that classifies data into two classes. In our proposed system, we have defined two classes as cancerous or non-cancerous. SVM is a binary classification method that takes as input labeled data from two classes and outputs a model file for classifying unknown or known data into one of two classes.

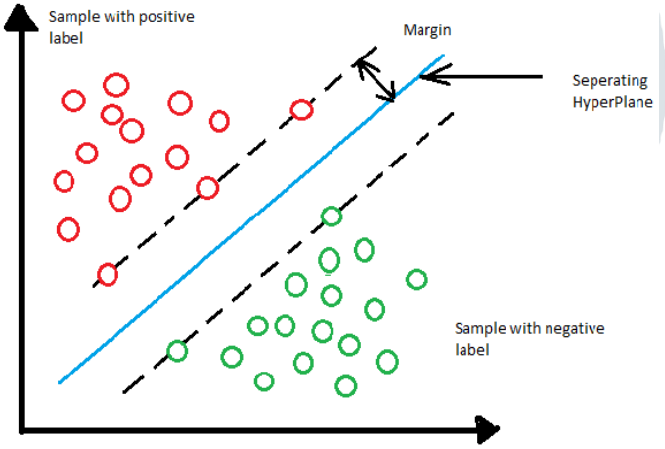

Fig 6. SVM Classifier

Training an SVM involves feeding known data to the SVM along with previously known decision values, thus forming training Set. It is from the training set that an SVM gets its intelligence to classify unknown data.

TABLE 2
TUMOR CLASSIFICATION USING SVM

Image	Output Value	Classification
Image_1	0	Non-Cancerous
Image_2	1	Cancerous
Image_3	0	Non-Cancerous
Image_4	1	Cancerous
Image_5	1	Cancerous
Image_6	1	Cancerous
Image_7	1	Cancerous
Image_8	1	Cancerous
Image_9	0	Non-Cancerous
Image_10	1	Cancerous

3.7. Stage Detection

Early Detection represents one of the most promising approaches for reducing the growing cancer burden. The purpose of early detection is that it will identify cancer while still localized and curable, preventing not only mortality, but also reducing costs.

IV.FUTURE WORK

In the future scope, the accuracy of the system can be improved if training is performed by using a very large image database. Similarly, we have planned to use MRI scan images to detect different types of lung cancer or other cancers using the same approach.

V.CONCLUSION

The current best system does not classify into different stages of cancer such as stage I, II, III or IV which results into non-satisfactory results. Therefore new system is proposed. The proposed system is successfully able to classify both benign and malignant tumors more accurately. It is also possible for early stage detection of lung cancer by using this system, thereby overcoming the limitations of the existing system. SVM can handle better complex classifications as compared to KNN classifier.

REFERENCES

[1] Xiuhua,G., Tao, S., & Zhigang, L.(2011) "Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image." In Theory and Applications of CT Imaging and Analysis. DOI: 10.5772/14766.
https://www.intechopen.com/download/pdf/14768

[2] Aggarwal, T., Furqan, A., & Kalra, K. (2015) "Feature extraction and LDA based classification of lung nodules in chest CT scan images." 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI),DOI: 10.1109/ICACCI.2015.7275773.
https://ieeexplore.ieee.org/abstract/document/7275773

[3] Jin, X., Zhang, Y., & Jin, Q. (2016) "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID). DOI:10.1109/ISCID.2016.1053.
https://ieeexplore.ieee.org/abstract/document/7830327.

[4] Sangamithraa, P., & Govindaraju, S. (2016) "Lung tumour detection and classification using EK-Mean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet). DOI:10.1109/WiSPNET.2016.7566533.
https://ieeexplore.ieee.org/abstract/document/7566533/authors#authors

[5] Roy, Sirohi., & Patle (2015)" Classification of lung image and nodule detection using fuzzy inference system". International Conference On Computing, Communication Automation. DOI: 10.1109/CCAA.2015.7148560.
https://www.sciencedirect.com/science/article/pii/S1877050917327801

[6] Ignatious, S., & Joseph, R. (2015) "Computer aided lung cancer detection system." 2015 Global Conference On Communication Technologies (GCCT), DOI: 10.1109/GCCT.2015.7342723.
https://ieeexplore.ieee.org/abstract/document/7342723

[7] Rendon-Gonzalez, E., & Ponomaryov, V. (2016) "Automatic Lung nodule segmentation and classification in CT images based on SVM." 2016 9Th International Kharkiv Symposium On Physics And Engineering Of Microwaves, Millimeter And Submillimeter Waves (MSMW). DOI: 10.1109/MSMW.2016.7537995.
https://ieeexplore.ieee.org/document/7537995/authors#authors

[8].The Cancer Imaging Archieve(TCIA) Dataset
https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics

[9]American Cancer Society, Costs of Cancer, 2002 [online], (cited 25 Jan 2003), http://www.cancer.org/docroot/MIT/content/MIT_3_2X_Costs_of_Cancer.asp

[10] Sruthi Ignatious , Robin Joseph, Jisha John Dept. of Computer Science and Engineering, Mar Baselios college of Engineering and Technology Trivandrum, Dr. Anil Prahladan Dept. of Imageology, Regional Cancer Centre Trivandrum, " Computer Aided Lung Cancer Detection and Tumor Staging in CT image using Image Processing" India International Journal of Computer Applications (0975 – 8887) Volume 128 – No.7, October 2015
https://pdfs.semanticscholar.org/e8bf/3d6b4d897fd3c9e13feed03636d3ee0f1845..pdf