Identification of Lung Cancer from Chest CT Images using SVM

Aravind S 21MM61R12
MTech Medical Imaging and Informatics,
SMST, IIT Kharagpur
Kharagpur, India
aravindsri30@gmail.com

Surendar Raj M 21MM61R05
MTech Medical Imaging and Informatics
SMST, IIT Kharagpur
Kharagpur, India
rajsurendar07@gmail.com

I. Team Member's Contribution

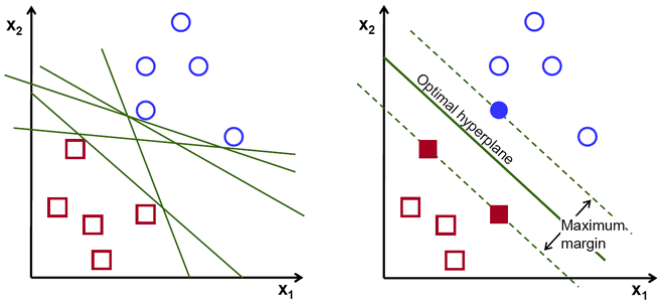
Aravind S- Image Analysis (Pre-processing, Segmentation)
Surendar Raj M- Feature Extraction and SVM classifier.

II. Introduction

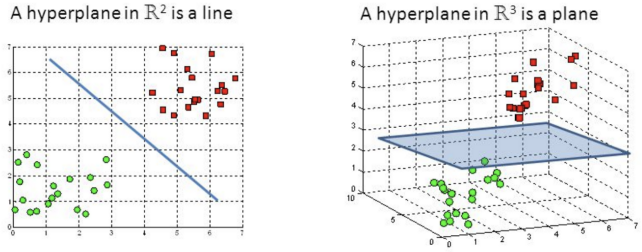
Lung cancer is one of the most predominant and widespread form of cancer with a high mortality rate of 50 percent. It is assessed that around 12,203 people had lung tumor diagnosed in 2016, out of which 7130 were male and 5073 female.Mortality from lung malignant growth in 2016 was 8839. Lung cancer is silent in initial stages but becomes aggressive at later stages making it difficult to treat. Currently, the clinical diagnosis of lung cancer is dependent on methods like biopsy, Sputum cytology and Bone scan. CT scan or Computerized Axial Tomography (CAT) scan is the most sensitive and specific detection modality which produces cross-sectional images of specific areas of scanned object by the use of computer processed combination of many X-ray images taken from different angle. Images obtained from CT scans can be used for non-invasive diagnosis of Lung Tumor using Machine Learning (ML) methods for faster and more accurate diagnosis.

A. Support Vector Machines (SVM)

Support vector machine (SVM) is a supervised learning algorithm which intends to find a hyper-plane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features.If the number of input features is 2, the hyperplane is a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane.Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.



Possible Hyperplanes | Optimal Hyperplanes



Hyperplanes in 2D and 3D feature space

III. Objective of the Project

To analyse and automatically segment Lung CT images and classify them as normal or cancerous using Image analysis and SVM based classification.

IV. Methodology and Detailed Algorithm

A. Image Filtering

The input image *input img* given for classification to the model is first filtered using a median filter. Median filter runs through the image pixel by pixel, replacing each pixel with the median of neighbouring entries over a [3x3] window, removing noise and preserving edges *medfiltered*. The median is calculated by first sorting all the pixel values from the window into numerical order, and then replacing the pixel being considered with the middle (median) pixel value *medfilt2* command is used in MATLAB to apply Median filtering to the input image.

B. Image Analysis and Segmentation

- **Calculating Gradient and Gradient Magnitude of the Image**
  - Gradient of the image is calculated to find the directional change in the intensity of the image in the x y directions to find the edges. *imgradientxy* function is used for calculating Gx and Gy which are the gradient along X and Y direction in the filtered image. The gradient magnitude *Gmag* is also calculated using *imgradient* function.
- **Image Analysis**
  - Opening operation which is erosion followed by a dilation, using the same structuring element (disk of size=10) is done on the filtered image to emphasize and connect darker areas of the image *openedimg*.
  - This is followed by Erosion operation using *imerode* function which performs grayscale erosion (emphasizes the darkest region in the given window) on the filtered image.(*erodedimg*)
  - Morphological reconstruction is done next on the eroded image to extract or enhance marked objects from an image. The eroded image (*erodedimg*) is given as the marker and the Median filtered image (*medfiltered*) as the mask to the *imreconstruct* function. (*reconstrimg*)
  - This is then followed by Closing operation using *imclose* on the *openedimg* which emphasizes and connects brighter spots of the image and removes darker areas that are smaller than the structuring element. (*closedimg*)
  - Dilation operation is performed on *reconstrimg* to highlight any regions of higher pixel value than the structuring element.(*dilatedimg*)
  - Reconstruction operation is again performed on *dilatedimg* and it's complement/negative is found so as to highlight any regions of abnormality when superimposed on the original image. (*reconstructandcomp*)
  - Regional maxima using *imregionalmax* function is found on *reconstructandcomp*. (*regionmax*)
  - Closing operation is done on *regionmax* which gives *regionmaxclose* followed by erosion operation on it resulting in *regionmaxcloseerode*.
  - To remove all connected components that have fewer than 20 pixels, *bwareaopen* function is used which gives *regionmaxareaopen*.
  - *reconstructandcomp* is binarized and Distance transform of the image is found which computes the Euclidean distance transform and assigns a number that is the distance between that pixel and the nearest nonzero pixel of the binary image. (*disttrans*)
  - Watershed transform is then performed on *disttrans* which segment contiguous regions of interest into distinct objects by finding catchment basins or watershed ridge lines in an image by treating it as a surface where light pixels represent high elevations and dark pixels represent low elevations. (*watershedimg*)
  - *imimposemin* function used to modify *Gmag* and *bgm* using morphological reconstructing so it only has regional minima wherever binary marker image *regionmaxareaopen* is nonzero.
  - The obtained image is then superimposed on the original image to highlight the regions of abnormalities as markers and boundary regions.
- **Discrete Wavelet Transform** (Noise Removal) DWT is a wavelet transform for which the wavelets are sampled at discrete intervals. DWT provides a simultaneous spatial and frequency domain information of the image. In DWT operation, an image can be analyzed by the combination of analysis filter bank and decimation operation. The analysis filter bank consists of a pair of low and high pass filters corresponding to each decomposition level. The low pass filter extracts the approximate information of the image whereas the high pass filter extracts the details such as edges.The application of 2D DWT decomposes the input image into four separate sub bands: low frequency components in horizontal and vertical directions (cA), low frequency component in the horizontal and high frequency component in the vertical direction (cV), high frequency component in the horizontal and low frequency component in the vertical direction (cH) and high frequency components in horizontal and vertical directions (cD). cA, cV, cH and cD can also be represented as LL, LH, HL and HH respectively.The representation of an image I after 1-level DWT with its sub-bands is given by

$$I = I_a^1 + (I_h^1 + I_v^1 + I_d^1)$$

where I_a^1 a represents the approximation of input image (smaller scaled form) and I_h^1 , I_v^1 , I_d^1 represent horizontal, vertical and diagonal details respectively, where the powers of the terms represent the level of decomposition. Further decomposition can be achieved by decomposing the LL sub band successively and the resultant image is split into multiple bands. An image after 3-level DWT decomposition is represented by

$$I = I_a^3 + \sum_{i=1}^{3}(I_h^1 + I_v^1 + I_d^1)$$

The use of DWT as a feature extractor allows the transformed data to be sorted at a resolution which matches its scale. The multi-level representation of the transformed image allows both small and large features to be discernable as they can be studied separately. The discontinuities in data are handled better by DWT than Discrete Cosine Transform (DCT) as the wavelet transform is not a Fourier-based transform.
- **Principal Components Analysis (PCA)**
  Principal Components Analysis (PCA) is a mathematical formulation used in the reduction of data dimensions. Thus, the PCA technique allows the identification of

standards in data and their expression in such a way that their similarities and differences are emphasized. Once patterns are found, they can be compressed, i.e., their dimensions can be reduced without much loss of information. In summary, the PCA formulation may be used as a digital image compression algorithm with a low level of loss.

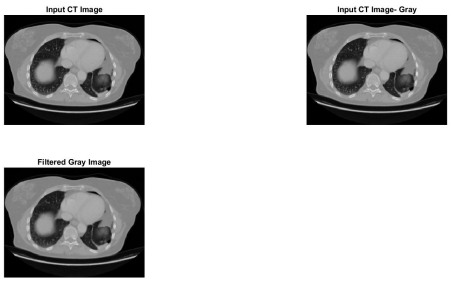
In the PCA approach, the information contained in a set of data is stored in a computational structure with reduced dimensions based on the integral projection of the data set onto a subspace generated by a system of orthogonal axes. The optimal system of axes may be obtained using the Singular Values Decomposition (SVD) method. The reduced dimension computational structure is selected so that relevant data characteristics are identified with little loss of information. Such a reduction is advantageous in several instances: for image compression, data representation, calculation reduction necessary in subsequent processing, etc.

- **GLCM Features**
  graycomatrix() function is used to create a gray-level co-occurrence matrix (GLCM) from the input image. GLCM is calculated by how often a pixel with gray-level value i occurs horizontally adjacent to a pixel with the value j. Each element (i,j) in GLCM specifies the number of times that the pixel with value i occurred horizontally adjacent to a pixel with value j.The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. GLCM features Contrast, Correlation Energy and Homogeneity are calculated from the input image.
- **Other Features** Other features of the input image like Mean, Standard Deviation, Entropy, RMS, Variance, Smoothness, Kurtosis, Skewness and IDM are also calculated and used as predictors to the SVM model to improve model accuracy.
- **Model Generation**
  fitcsvm function is used to generate a fine Gaussian SVM model which trains or cross-validates model for one-class and two-class (binary) classification on a low-dimensional or moderate-dimensional predictor data set. fitcsvm supports mapping the predictor data using kernel functions, and supports sequential minimal optimization (SMO), iterative single data algorithm (ISDA), or L1 soft-margin minimization via quadratic programming for objective-function minimization.

V. Data Description

- The dataset used is from Kaggle titled Chest CT-Scan images Dataset.
- The dataset was modified for convenience and split into 3 groups- Train, Test and Validate which consisted of 70%, 20% and 10% of the dataset respectively.
- There are 2 classes in the dataset- Normal and Cancerous. Training dataset contained 148 normal images and 465 cancerous images while the Test dataset contained 54 normal and 261 cancerous images. Validation set contained 13 normal and 59 cancerous images.
- A custom batch feature extraction program was also written to generate labelled data from the training set.
- The aforementioned modified dataset and batch feature extraction program can be found can be found here .

VI. Results



Reading Input Image and Filtered Image

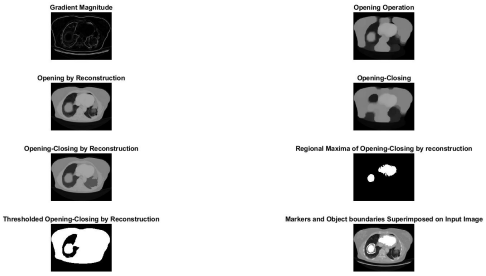
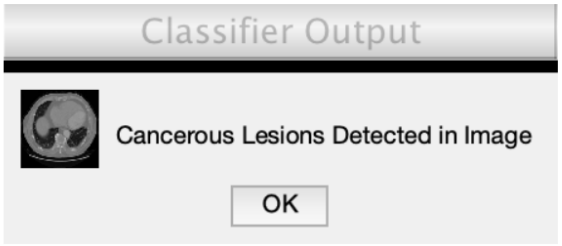


Image Analysis on Filtered Image

Extracting Features of Input Image for Classification

```
Contrast = 0.1297
Correlation = 0.0855
Energy = 0.8498
Homogeneity = 0.9603
Mean = 0.0013
Standard_Deviation = 0.0693
Entropy = 3.0814
RMS = 0.0693
Variance = 0.0048
Kurtosis = 11.4825
Skewness = 0.6925
IDM = -0.2592
```

Input Image GLCM features




```
Model Accuracy is 97.72
```

Model Classification Output and Accuracy

VII. Observations

This proposed model presents an image processing technique using an SVM algorithm with improved performance compared to other linear classifiers. The model presents an improved image processing technique for the detection of lung cancer. The greatest challenge faced came from image processing techniques. Also, the training data images for the SVM classifier were found to be insufficient in number.

VIII. Conclusion and Future direction of research

In this paper, an image processing techniques has been used to detect early stage lung cancer in CT scan images. The CT scan image is pre-processed followed by segmentation of the ROI of the lung. Discrete waveform Transform is applied for image compression and features are extracted using a GLCM. The results are fed into an SVM classifier to determine if the lung image is cancerous or not. The classifier achieves an accuracy of 95%. Future works should be focused on improving Image Segmentation algorithms and accuracy of the SVM classifier.

References

[1] D. P. Kaucha, P. W. C. Prasad, A. Alsadoon, A. Elchouemi and S. Sreedharan, "Early detection of lung cancer using SVM classifier in biomedical image processing," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 3143-3148, doi: 10.1109/ICPCSI.2017.8392305.

[2] R. Sathishkumar, K. Kalaiarasan, A. Prabhakaran and M. Aravind, "Detection of Lung Cancer using SVM Classifier and KNN Algorithm," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1-7, doi: 10.1109/ICSCAN.2019.8878774.

[3] Asuntha, A. Brindha, A. Indirani, S. Andy, Srinivasan. (2016). Lung cancer detection using SVM algorithm and optimization techniques. 9. 3198-3203.

[4] Firmino, M., Angelo, G., Morais, H. et al. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. BioMed Eng OnLine 15, 2 (2016).

[5] V. Govindaraj, B. Arunadevi. (2021) Machine Learning Based Power Estimation for CMOS VLSI Circuits. Applied Artificial Intelligence 35:13, pages 1043-1055.

[6] B R Manju et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1012 012034

Appendix A- Code

This document was created using LaTeX software and hence does not support MATLAB code syntax.

The code accompanying this model can be found here .