Classification of Lung Disease using Local and Global Descriptors

R. Pradeebha
M.Phil Research Scholar,
Department of Computer Science,
PSGR Krishnammal College for Women,
Coimbatore-641 004, Tamil Nadu, India

S. Karpagavalli
Associate Professor,
Department of Computer Science,
PSGR Krishnammal College for Women,
Coimbatore-641 004, Tamil Nadu, India

ABSTRACT
Recent trends indicate that instances of chronic respiratory diseases are on the rise in India mainly due to vehicular pollution, air and dust pollution, habit of smoking and also increased population. A World Health Organization report indicates that India has a ranking number one in the world for lung disease deaths. Respiratory diseases like asthma, chronic obstructive pulmonary disease (COPD), Interstitial Lung Disease (ILD), pneumonia, tuberculosis (TB) are emerging as most important health problems in the country. The proposed work is aimed at establishing more advanced diagnostic strategy for lung diseases using CT scan images. Lung diseases such as Emphysema, Pneumonia, Bronchitis are classified using CT scan images which is collected from National Biomedical Imaging Archive (NBIA). A total of 366 images are used, out of which 300 images are used for training and 66 images are used for testing. The classification task carried out with classifier support vector machine (SVM) using Histogram of Oriented Gradient (HOG) –global descriptors and Local Binary Pattern (LBP) – local descriptors. The performance of the model built using Support Vector Machine indicates that it is effective in the prediction of lung disease with 98% predictive accuracy.

Keywords
Chronic Obstructive Pulmonary Disease, Histogram of Oriented Gradient, Local Binary Pattern, Support Vector Machine.

1. INTRODUCTION
The lungs are a pair of highly elastic and spongy organs in the chest. They are the main organs needed for breathing. They take in air from the atmosphere and also provide an area for oxygen to enter the blood and for carbon dioxide to leave the blood. There are various types of lung disease they are emphysema, pneumonia, bronchitis, lung cancer, sarcoidosis and pleural effusion [1]. Emphysema is a type of long-term lung disease. People with emphysema have difficulty in breathing due to a restriction in blowing air flow outside. Pneumonia is an inflammatory of the lungs. Generally the inflammation is caused by infection, however it can be caused due to deep breathing certain chemical substance fumes. Pneumonia can affect just one lobe of the right or left lung, a whole lung, or both lungs. Bronchitis is inflammation and also infection of the airways (bronchial tubes) of the lungs. Signs or symptoms consist of a determined cough that often produces mucus (sputum), fever, mild wheezing, and chest pain [2]. A sharp rise in respiratory disease in India due to infection, smoking and air pollution in the country. Respiratory diseases were no longer restricted to the elderly but were now being detected even in younger age groups. These factors motivated us to carry out the proposed lung disease classification task using CT scan images.

The rest of the paper is organized as follows: Section 2 provides the review of literature which covers similar work carried out to classify lung diseases by adopting various features and methods. Section 3 elaborate the proposed framework with clear sketch of various stages in this work. Section 4 provides the detailed description of feature extraction methods HOG and LBP. Section 5 discuss about the classification approach using support vector machine. In section 6 the experimental results and analysis of the models is presented. Section 7 gives summary of the work.

2. LITERATURE SURVEY
In recent years, many research work have been carried out and many effective methodologies, features and algorithms have been proposed by researchers.

C. Bhuvaneswari et al., [3] proposed a method to detect normal lung, bronchitis, emphysema, pleural effusion using 400 images of lung disease. The work involved preprocessing the images, novel fusion based feature extraction, feature selection by genetic algorithm and classification. The preprocessing carried out using median filter and morphological smoothening and extracting the features using the novel fusion Median Absolute Deviation technique by the combination of gabor feature and Walsh Hadamard transform. Feature selection had been done using genetic algorithm and classification using Decision trees, K-nearest neighbor and MLP. The results shows the achievement of detecting the lung disease is 91% in MLP. In another work, they presented a method for segmentation and recognition of lung disease CT images using the fuzzy c-means clustering and the texture features are extracted using the gabor filtering and selecting the features using the Information Gain, Principal Component Analysis, Correlation based feature selection are employed with genetic algorithm. They achieved high accuracy using correlation based feature selection method and Naive Bayes classifier [4].

Xiabi Liu et al., [5] proposed a method to detect the 9 types of lung disease such as grand grass opacity, lobulation, cavity and vacuoles, spiculation, pleural identification, calcification, air bronchogram, bronchial mucus plugs and obstructive pneumonia. They used Fisher criterion and Genetic optimization (FIG) method to select the features from an image including the bag-of-visual-words based on HOG, LBP, wavelet transform and CT value Histogram in 511 ROIs captured from lung CT images. The lung disease are classified using the SVM, Bagging, Naive Bayes, K-Nearest Neighbor and Adaboost. They achieved high classification accuracy in SVM.

S.P. Kamalapriya et al., [6] proposed a method to detect the lung tissues in High Resolution Computed Tomography (HRCT) images. The ILD dataset consists of 130 images. The

lung disease are normal, emphysema, ground glass, fibrosis and micro nodules. They extracted LBP and HOG features on the ROI segmented image. Each image patch is then classified based on reference dictionaries with a new patch-adaptive sparse approximation (PASA) algorithm.

Swati P. Tidke et al., [7] proposed a method to detect the lung tumor whether benign or malignant in Computer Tomography images. Pre-processed the images by denoising and segmented the images using thresholding and morphological operation. They used GLCM features and classified by SVM and the accuracy is 92.5%.

3. PROPOSED WORK

In the proposed work, lung disease prediction from CT images has been implemented to solve using classification algorithms. An interactive user interface has been designed to detect the type of lung disease present in the given diseased CT scan image of lungs. The detection task involves Region Of Interest (ROI) segmentation, preprocessing, feature extraction, building models through training and testing the performance of the models. Each phase is described in following section and the framework for the detection of lung disease from CT images is shown in figure 1.
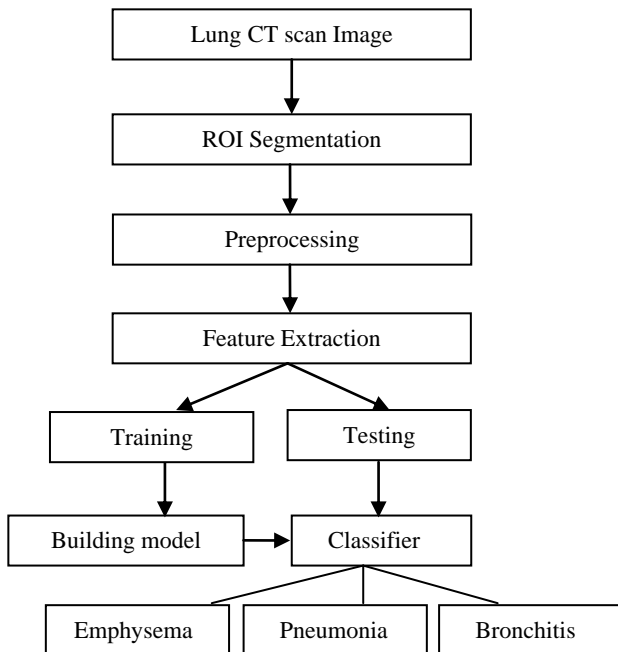

Figure 1. Framework for Detection of Lung Disease from CT scan Images

3.1 ROI Segmentation

A region of interest is a selected sub-region of a CT scan image which is to be processed further. A ROI of 256×256 size by choosing the centre of abnormality x and y value.

3.2 Preprocessing

The median filter is a nonlinear digital filtering technique, often used to remove noise. Such noise reduction is a typical pre-processing step to improve the results. In median filtering, the value of the pixel is replaced by the median of the pixel values in the 3x3 neighborhood.

4. FEATURE EXTRACTION

In the proposed work, the local and global features are extracted from the preprocessed CT scan image. The various

steps involved in HOG and LBP feature extraction techniques is elaborated below.

4.1 HOG Features

The Histogram of Oriented Gradients (HOG) is global descriptor of an image, mainly used in computer vision and image processing designed for object detection. The technique counts occurrences of gradient orientation in localized portions of an image [8].

The important concept behind the histogram of oriented gradients descriptor is local object appearance and also shape within an image are defined by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and also for the pixels within each cell, a histogram of gradient orientations is compiled. The descriptor is then the concatenation of these histograms. For improved accuracy, the local histograms can be contrast-normalized by calculating an approach of the intensity across a larger region of the image, called a block, and by using this value to normalize each cells within the block. This normalization results in better invariance to changes in illumination and shadowing. The steps involved in HOG features extracting process is shown in figure 2.
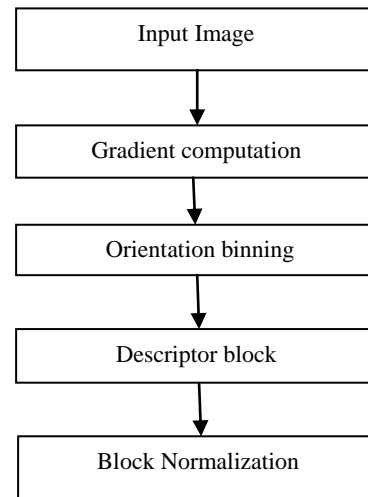

Figure 2. HOG feature extraction

4.1.1 Gradient Computation

In the HOG features, dividing the image window into small spatial regions called cells, for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The calculation is the computation of the gradient values. The most common method is to apply the 1-D centered, point discrete derivative mask in one or both of the horizontal and vertical directions. Specifically, this method requires filtering the colour or intensity data of the image with the following filter kernels:

$$[-1, 0, 1] \text{ and } [-1, 0, 1]^T$$

Testing the gradients computed using Gaussian smoothing followed by one of several discrete derivative masks.

4.1.2 Orientation Binning

In the HOG features, the next step of calculation is creating the cell histograms. Each pixel calculates a weighted vote for an edge orientation histogram channel based on the orientation of the gradient element centred on it, and the votes are accumulated into orientation bins over local spatial regions is called cells. Cells can be either rectangular or

radial. The orientation bins are evenly spaced 0 to 180 degrees or 0 to 360 degrees, depending on whether the gradient is "unsigned" or "signed". In this experiment, the image pixel size cells are computed as 128x128 with 9 orientation bins for [0°,180°] interval. For each pixel's orientation, the corresponding orientation bin is found.

4.1.3 Descriptor Block

In the HOG features, the gradient strengths must be locally normalized, which requires grouping the cells together into larger, spatially connected blocks. The HOG descriptor is the concatenated vector of the components and parts of the normalized cell histograms from every one of the block regions. These blocks typically overlap, meaning that each cell contributes more often than once to the final descriptor. Two main block geometries occur: rectangular R-HOG blocks and also circular C-HOG blocks.

In this experiment, the rectangular R-HOG is used for calculating the optimal parameters 2x2 cell blocks of 4x4 pixel cells with 9 histogram channels. For each CT scan image 36 HOG -global features are extracted.

4.1.4 Block Normalization

There are four different methods for block normalization. Let V be the non-normalized vector containing all histograms in a given block, $\|v\|_K$ be its k-norm for k=1,2 and e be some small constant (the exact value, hopefully, is unimportant). Then the normalization factor can be one of the following:

L2-norm: $$f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}}$$

L2-hys: L2-norm followed by clipping (limiting the maximum values of v to 0.2) and renormalizing, as in

L1-norm: $$f = \frac{v}{\|v\|_{1+e}}$$

L1-sqrt: $$f = \sqrt{\frac{v}{\|v\|_1 + e}}$$

In addition, the scheme L2-hys can be computed by first taking the L2-norm, clipping the result, and then renormalizing.

4.2 LBP Features

A Local Binary Pattern (LBP) is a local descriptor that encapsulates the appearance of an image in a small neighborhood around a pixel. The basic local binary pattern operator, was depending on the assumption that texture provides locally two different elements, a pattern and its strength. The local binary pattern operator works in a 3×3 pixel block of an image. The pixels in this particular block are thresholded by the center pixel value, multiplied by powers of two and then summed to get a label for the center pixel. As the neighborhood consists of 8 pixels, a total of $2^8 = 256$ various labels can be obtained depending on the related gray values of the center along with the pixels in the neighborhood [9]. The mean and standard deviation of LBP-Local features of an image are calculated and used in classification along with global features

5. SUPPORT VECTOR MACHINE

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that examine data and identify patterns. A support vector machine develops a hyper plane or number of hyper planes in a high or unlimited

dimensional space, which is used for classification. If hyper plane is achieved the good separation and consists of the largest distance to the nearby training data points of any class. When classifier having the larger margin then generalization error will be lower. Assigned a set of training examples, each one marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one group or the other, making it very a non-probabilistic binary linear classifier [10]. The Maximal Margin Classifier has been shown in figure 3.
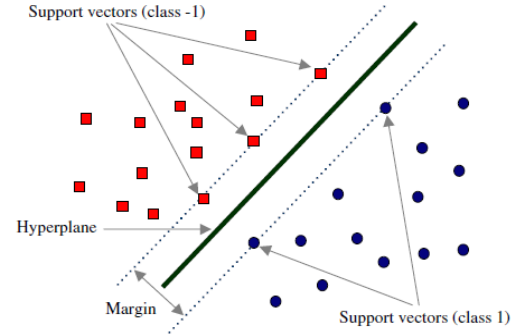

Figure 3. Maximal Margin Classifier

5.1 Multi-class SVM

Instead of creating several binary classifiers, a more natural way is to distinguish all classes in one single optimization processing [11]. For a k-class problem, these methods design a single objective function for training all k-binary SVMs simultaneously and maximize the margins from each class to the remaining ones. Given a labelled training set represented by $\{(x_1,y_1),...,(x_1,y_1)\}$ of cardinality l, where $x_i \in R^d$ and $y_i \in \{1,...,k\}$, the formulation proposed in [12] is given as follows:

$$\min \quad \frac{1}{2}\sum_{m=1}^{k} w_m^T w_m + C \sum_{i=1}^{l} \sum_{t \neq y_i} \xi_{i,t}$$
$$w_m \in H, b \in R^k, \xi \in R^{l \times k}$$

subject to $\quad w_{y_i}^T \varphi(x_i) + b_{y_i} \geq$
$$w_t^T \varphi(x_i) + b_t + 2 - \xi_{i,t}$$

$$\xi_{i,t} \geq 0,$$
$$i =$$

$$1, \dots, l, t \in \{1, \dots, k\}\backslash y_i.$$

The resulting decision function is
$$argmax_m \ f_m(x)$$
$$= argmax_m(w_M^T \varphi(x) + b_m)$$

6. EXPERIMENT AND RESULTS

The experiment has been carried out using the CT scan images collected from the National Biomedical Imaging Archive (NBIA) database [13]. In this experiment, a total of 360 images with emphysema, pneumonia, and bronchitis were taken for processing.

In the dataset, 80% data are used for training and 20% used for testing. A classifier for lung disease classification is built using Support Vector Machine with linear, polynomial and RBF kernel in MATLAB. The results are presented in Table 1 and shows that Support Vector Machine with RBF

substantially better in prediction of type of lung disease with prediction accuracy of 98.5%. An effective user interface has been designed to predict the type of lung disease in the given CT scan image in MATLAB. Few screenshots of the interface are given in figure 4, 5, 6, 7.

Table 1. Performance of SVM classifier using Local and Global features

Classifier	LBP - HOG features	
	Correctly classified instances	Incorrectly classified instances
SVM - Linear Kernel	74.15%	25.85%
SVM - Polynomial Kernel	86.42%	13.58%
SVM - Radial Basis Function	98.48%	1.54%

Figure 4. Read image from database

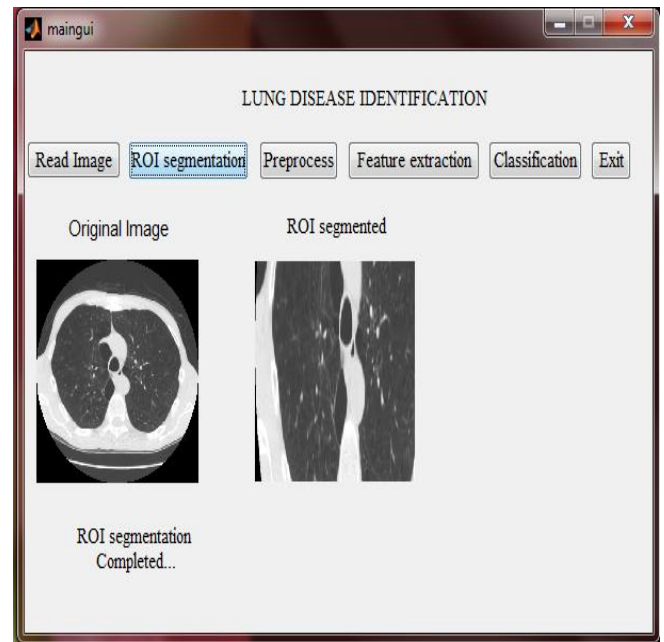Figure 4 shows how the user can load / select the CT scan image of a patient from the folder.

Figure 5. ROI segmented image

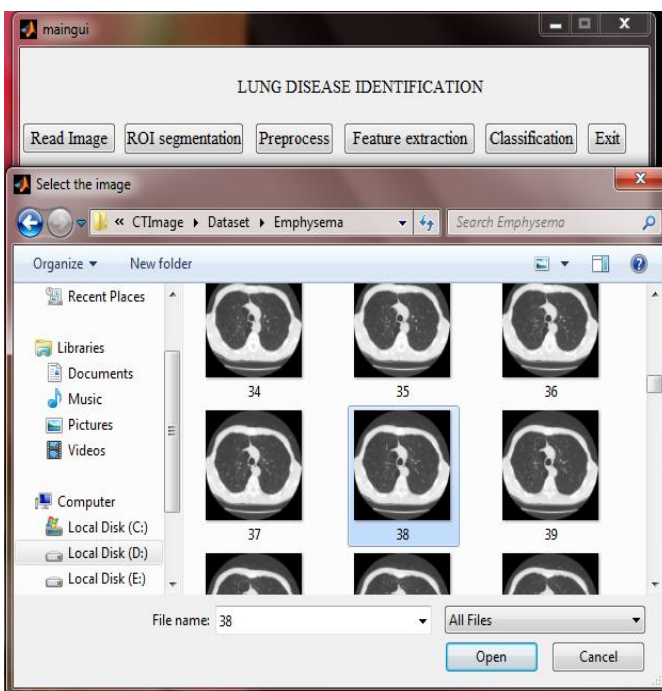Once the image is loaded, from the original image Region Of Interest is segmented which is shown in figure 5.
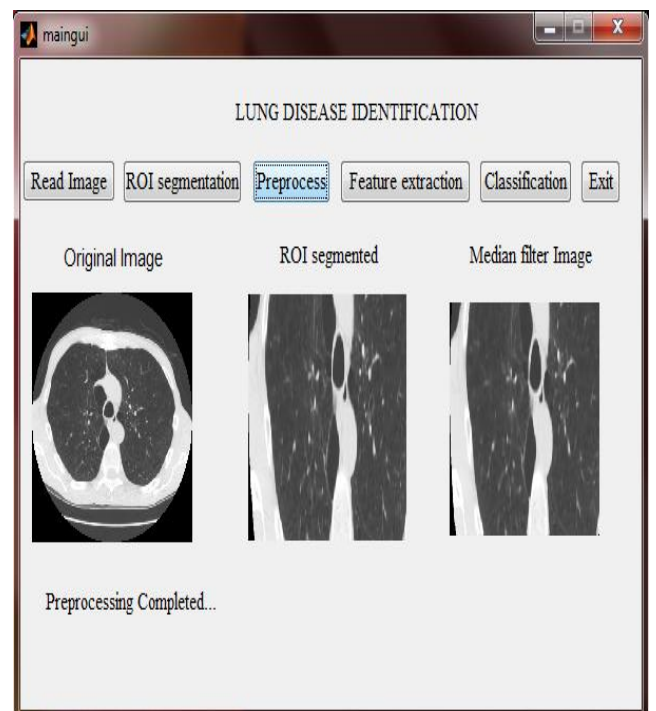

Figure 6. Preprocessing the Image

After the ROI segmentation, the image is preprocessed using median filter which is shown in figure 6. Once preprocessing is done, mean and standard deviation of LBP-Local descriptor along with 36 features of HOG-Global descriptor are extracted.
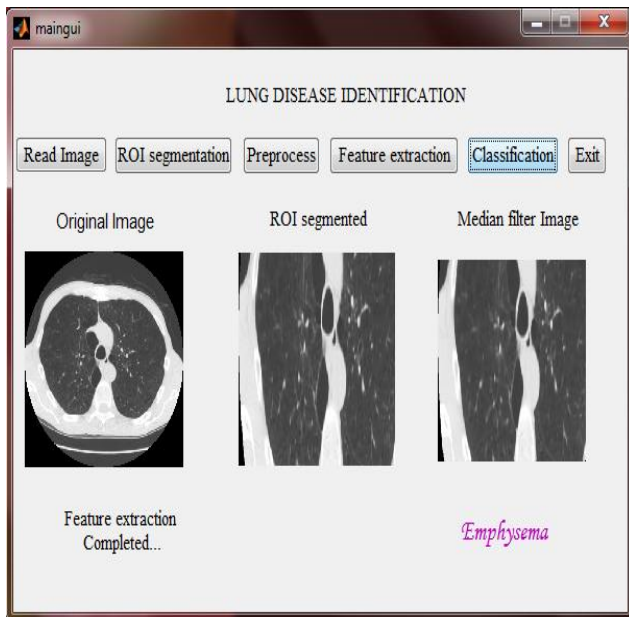
Figure 7. classified image

After extracting features, the SVM model predicts lung disease present in the given CT scan image shown in figure 7.

7. CONCLUSION

Lung disease classification through CT imagery analysis is a important clinical information to diagnose the lung disease. The proposed work automated the lung disease classification task using CT scan image of lungs by employing supervised machine learning technique, Support Vector Machine. It is likely that the interface for lung disease classification would be helpful in quick diagnosis of lung disease and improve the quality of life of patients with suitable therapies. The work can be extended to detect more number of lung diseases like pleural effusion, grand glass opacity, lobulation, cavity and vacuolus, spiculation, pleural identification, calcification, air bronchogram, mucus plugs. The work can be further extended using variants of LBP like Local Ternary Patterns (LTP), Volume Local Binary Pattern (VLBP), Direction coded Local Binary Patterns (DLBP), Local Derivative Pattern (LDP), Local Multiple Patterns (LMP).

8. REFERENCES

[1] Lung Disease & Respiratory Health Center. 2014. WebMD, LLC.

[2] Kraft M. 2014. Approach to the patient with respiratory disease. Goldman's Cecil Medicine, 24th edition, Philadelphia, PA: Saunders Elsevier.

[3] Bhuvaneswari C, Aruna P, Loganathan D. 2014. A new fusion model for classification of the lung disease using genetic algorithm. Egyptian Informatics Journal. 15(2): 69-77.

[4] Bhuvaneswari C, Aruna P, Loganathan D. 2013. Advanced segmentation technique using genetic algorithm for recognition of lung disease from CT scans of thorax. International Journal of Engineering Research and Applications. 3(4): 2517-2524.

[5] Xiabi Liu, Ling Ma, Li song, Yanfeng Zhao, Xinming Zhao, Chunwu zhou. 2014. Recognizing common CT imaging signs of lung disease through a new feature selection method based on fisher criterion and genetic optimization. IEEE Journal of Biomedical and Health Informatics. 19(2): 635 - 647.

[6] Kamalapriya SP, Pathur Nisha S, Thangarasu VS. 2014. Enhanced image patch approximation for lung tissue classification using feature based extraction. Indian Journal of Research. 3(2): 226-228.

[7] Swati P. Tidke, vrishali A, Chakkarwar. 2012. Classification of lung tumor using SVM. International Journal Of Computational Engineering Research. 2(5): 1254-1257.

[8] Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1: 886 - 893.

[9] Ojala T, Pietikainen M, Maenpaa T. 2002. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Trans. Pattern Analysis and Machine Intelligence. 24(7): 971 - 987.

[10] Soman K.P, Loganathan R, Ajay V. 2009. Machine Learning with SVM and other kernel method. PHI, India.

[11] Vapnik V. 1998. Statistical Learning Theory. Wiley-Interscience, New York.

[12] Weston J, Watkins C. 1999. Multi-class support vector machines. In Proceedings of European Symposium on Artificial Neural Networks.

[13] Lung disease CT scan image database. https://niams.imaging.nci.nih.gov/ncia/login.jsf.