AN IMPROVED RANDOM FOREST APPROACH FOR PREDICTING TUBERCULOSIS

R. Beaulah Jeyavathana[1], Kalpana G[2] ,K.V.Kanimozhi[3]
Assistant Professors (SG)[1,3],Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,
Chennai, India.
Assistant Professors[2], Department of Computer Science and Engineering, Rajalakshmi
Institute of Technology, Chennai, India.

Abstract. Tuberculosis is one of the perilous infectious diseases that can be categorized by the evolution of tubercles in the tissues. This disease mainly affects the lungs and also the other parts of our body. Five stages are being used to detect tuberculosis disease. They are pre-processing an image, segmenting the lungs and Extracting the feature, Feature Selection and Classification. The optimal features are selected by Modified Random forest. Finally, Support Vector Machine classifier method is used for image classification. The proposed system accuracy results are better than the existing method inclassification.

Index Terms—Tuberculosis, Segmentation,K Means clustering, Feature extraction, GLCM approach, Modified Random Forest, SVM classifier.

1.Introduction

Tuberculosis is one of the communicable bacterial diseases and that may affect any tissues of the body but it primarily disturbs the lungs. TB is one of the airborne pathogens that can binge through air or by coughing or sneezing from one person to another. TB disturbs all age groups in all parts of the world. Tuberculosis bacteria are present in sputum trials and it is identified under a microscope. X-ray is not easily predicting the early stage of tuberculosis [1]. Hence, because of this wrong prediction of tuberculosis, an automated detection of tuberculosis is used. To overcome the problems in existing methods, CT lung images are used for diagnosis of tuberculosis[2].In image processing feature contains the information that is related to colour, shape, texture and context[3]. Modified Random Forest Algorithm technique is based on optimization searching technique and it is used to find the optimal solutions. It is used for selecting the best features after the feature extraction process. This will continue until a needed solution is obtained[4][5]. Classifying the images whether it is normal or abnormal is done by SVM classifier.

2. Proposed work

In our study, dataset containing lung CT images comprising abnormal lung and normal lung are taken from several patients was utilized. The lung diseases are categorized by the radiologist from the CT image. Images are collected from male and female patients whose ages are ranging from 15 to 78 years.

2.1 Preprocessing

Pre-processing is refining our image data by supressing all unwanted noise and distortions so that we can enhance the image. This step is very imperative in image processing. In our proposed system we have used wiener filter to remove noise. Since medical images in our dataset is of excessive noise and blurring, we have been used Wiener filter[6]. Since Weiner Filters are used in frequency domain it yields good results. It also preserves the edges and fine details of lungs. It is low pass-filter. The filter size of 5*5 is selected to avoid over smoothing of the image. 2D Wiener filter is used for lessening of additive gaussian white noise in images. We have compared the results with Gaussian Filter. It removes the noise efficiently when compared with Weiner filter. Since it is a linear filter it performs better more than weiner filter and it keeps the edges of our images relatively sharp. In figure 2 we have given an input image, Figure 3 represents pre-processed image using Weiner Filter, Figure 3 represents pre-processed image using Gaussian Filter.

2.1.1 Segmentation

Image Segmentation is being carried out to classify the images into different groups. For image segmentation we have used K Means Clustering algorithm. There are different methods for segmentation and we have chosen K-Means algorithm because if we keep k small it performs better than other algorithms. We have chosen cluster centre k value as 3 and accurate results have been obtained. Figure 4 represents the segmented lungs. The following steps describes K Means clustering algorithm.

Step 1: Choose the number of clusters K
Step 2: Select the centroids
Step 3: Assigning datapoints to form K Clusters
Step 4:Calculateand place the new centroid of each cluster.
Step 5:Recast each data point to the new centroid.
Step 6:If any reassignment repeat step 4, otherwise, stop.

2.2 ROIExtraction

ROI’s are taken out using the radiologists and thus it is authorized to attain the clinical relevance which progresses the performance of the system. Extract the defected tissues from the lung as ROI’s and then find the intensity level of the pixels and using the range of pixel intensity values discriminate the defected tissues and other lung tissues. If there is no defected tissues are present, then the slice is considered to be Normal. Then obtain the class labels for each ROI’s from the experts. Finally, ROI’s are extracted and also the class label information is obtained. Figure 5 represents the ROI extracted image.

2.3 Feature Extraction-GLCM approach

The feature extraction based on Texture feature is carried out. GLCM approach is used for extracting the features in given image such as entropy, energy, contrast, correlation, variance,sum average, homogeneity cluster shade and etc[7][8].., are considered for feature selection. Extract these twenty-two features for each ROI in four orientations 0^o, 45^o, 90^o, 135^o using GLCM also called the Grey Tone Spatial Dependency Matrix. Then we have fixed the window size whose length is of 500 and height is of 300. Window size represents the area of samples we have been taken. Then we have specified parameters appropriately so that selected intensity can be assigned to gray levels in calculating GLCM. Gray levels describe the intensity values and we have chosen minimum, maximum and number of levels of parameters. Then we have defined the spatial relationship between our reference samples and moderate samples. Selecting the distance and direction for our samples is crucial one and we have taken the values for distance is 1 and direction is in horizontal direction.

2.4 ClassificationSubsystem - Naïve Bayes Classifier

It is not a single algorithm but a family of algorithms where all of them stake a communal principle, i.e. every pair of features being classified is self-determining of each other. It is a probabilistic machine learning model that is mainly used for classification. We have made an assumption that our selected features are independent. We have chosen Bernoulli Naïve Bayes classifier for classifying our dataset since it yields better results by predicting our class variables appropriately.

2.5 Feature Selection

The term Feature selection pacts with choosing a subset of features, amid the entire features, that displays the unsurpassed presentation in classification accuracy. Inorder to reduce our number of input variables we have decided to choose a better optimization searching process since it is a challenging one to select an appropriateprocess. Since it outperforms individual constituent models the correlation is very high. So we have decided to use random forest classifier and we have compared the results of random forest classifier with Bernoulli Naïve Bayes classifier and we have found that performance and accuracy of Random Forest classifier is high when compared with Bernoulli Naïve Bayes Classifier. We have given accuracy results in Table 2. We have got accuracy using Naïve Bayes classifier as 92.3% and by using Modified Random Classifier we have got accuracy as 98%. The predictions made by the trees which are individual have low correlation on one another. So with the help of Random Forest Model the chances of predicting Tuberculosis is accurate by giving correct predictions.

2.6 Modified Random Forest Algorithm

Random forests are an unification of tree predictors such that each tree rest on the values of a random vector sampled self-reliantly. There are more difficulties in predominant machine learning techniques.

Number of trees engendered in the Random Forest is the task for the Researchers because it devours extra space in memory and also increases run time of the algorithm. We have proposed the modification of Random Forest Algorithm by splitting the patches that have been trained instead of pixels. Fig 2 shows the data set splited for Random forest algorithm. We have taken three datasets and for all three datasets value is of 80%, validation comes around 20% and testing comes around 20%. When all the results are validated with another dataset the accuracy we have got is same when compared with the same algorithm.

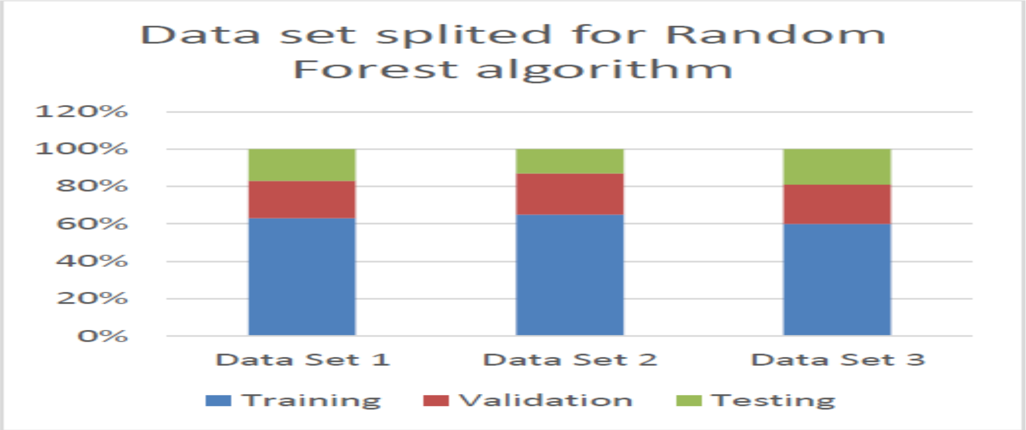


Figure 2

3.Results and Discussion

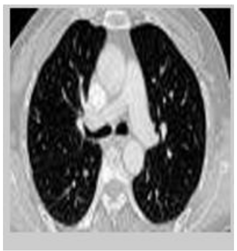
The CT images used for testing and training purpose for classification were collected from AARTHISCANS & LABS at TIRUNELVELI. We have several CT images, but we used 197 images for our work out of which 94 images have tuberculosis and the remaining images do not have tuberculosis. The segmentation of the image takes place. The set of Tuberculosis (TB) CT images and non-Tuberculosis CT images are tested to give precise result. Thus, the technique pacts with the accurate tuberculosis detection.

3.1 Processing Time Analysis

In our study, to implement our proposed algorithm, we used MATLAB software (R2016a) on a laptop, Intel Core i3 (2.0 GHZ) and 4GB memory. The resolution of images in our database was 512x 512. To assess our proposed algorithm competently, we scrutinized each step of our algorithm based on processing time. Table1 presents average processing time of each module of proposed algorithm.

Table1. Average Processing Time

Proposed algorithm.	Processing time of each Module
Module	Processing time (s)
Pre-processing	1.1456 seconds
K-Means segmentation	20.0196 seconds
Feature extraction	0.24534 seconds



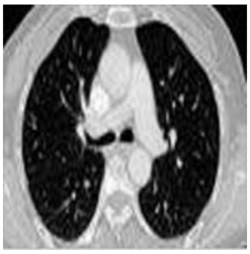


Fig 2 Input Image**Fig 3**Preprocessed image using GaussianFilter

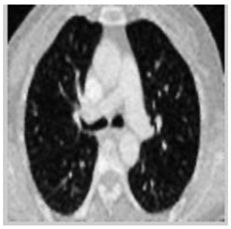


Fig 4 Preprocessed image using Weiner Filter

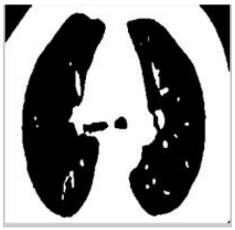


Fig 4 Segmented Lungs

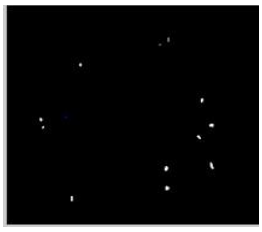


Fig 5ROI Extraction

Table 2. Accuracy

Parameter	Using PSO	Using MPSO	**Using Modified Random Forest**
TP	**43**	**46**	**49**
TN	**41**	**45**	**48**
FP	**09**	**05**	**02**
FN	**07**	**04**	**01**
Accuracy	**84**	**91**	**98**
Sensitivity	**86**	**92**	**96**
Specificity	**82**	**90**	**94**

In Table 2, we have taken the parameters like True positive, True negative False positive and False Negative. We have compared our proposed Modified Random Forest Classifier with Particle Swarm Optimization and Modified Particle Swarm Optimization. We have got accuracy more using our proposed approach when compared to PSO and MPSO. We have got accuracy as 98%.

Table3. Classification Accuracy

Classifier	**Performance Parameters**	**percentage**
Bayes classifier	**Accuracy**	**92.30%**
	Sensitivity	**96%**
	Specificity	**49%**

4.Conclusion

In our work an approach for automatic TB detection from CTs image has been proposed. So far, there are no automatic detection algorithms developed that can detect TB from CTs accurately. We have used Random Forest Classifier in detecting Tuberculosis disease effectively. The performance is being compared with Naïve Bayes Classifier and we got good accuracy using Modified Random Forest Approach. Experimental results demonstrate that our method achieves good accuracy. In future we have planned to use Deep learning Techniques in detecting Tuberculosis disease.

5. References

[1] Laurens H*, Clara I. Sánchez, Pragnya M, Rick P, Alistair S,"Automatic Detection of Tuberculosis in Chest Radiographs Using a Combination of Textural, Focal and Shape Abnormality Analysis" 2015 IEEE Transactions on Medical Imaging.

[2] Shenshen S, Wenbo L, Yan K "*Lung Nodule Detection Based on GA and SVM*" 2015 8th International Conference on Bio Medical Engineering and Informatics.

[3] G. Vijaya, A. Suhasini, R. Priya "*Automatic Detection of Lung Cancerin CT images" 2019 IJRET: International Journal of Research in Engineering and Technology*.

[4] Elmar R and Volodymyr P "*Automatic Lung Nodule Segmentation and Classification in CT Images Based on SVM*" 2016 International conferences IEEE.

[5] Mumini O Omisore proposed the genetic neuro-fuzzy inferential model for the diagnosis of tuberculosis, 2014, IEEE transactions.

[6] Dhinakaran. K, R. Anand, et.al (2020)," Video Surveillance Wildfire Detection using Dark Convolutional Neural Network", Test Engineering and Management, Vol.83, pp No. 11601 - 11604.

[7] Sema C, Stefan J, Palaniappan K, Jonathan P, "Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Non-rigid Registration" 2014 IEEE Transactions on Medical Imaging.

[8] Marius G*, William J, Jianfei L, Jeremy M, "Tumor Burden Analysis on Computed Tomography by Automated Liver and Tumor Segmentation"2016, IEEE.

Acknowledgments

We thank our colleagues for valuable comments and suggestions. We also thank Dr. Dhinakaran for his valuable support.