Tuberculosis Detection Using Machine Learning Algorithms

Ahana Dutta[1], Dr.C.Sweetlin Hemlatha[2]
School of Computer Science And Engineering, Vellore Institute of Technology, Chennai, India

***Abstract*—In today's world tuberculosis is a deadly disease. The whole world is affected by tuberculosis and also a huge number of people die everyday for it. This paper presents an implementation to detect if people is affected by tuberculosis or not with available datasets. For this detection purpose, chest x-ray images are required. Chest x-ray images with related details are collected from Montgomery Dataset and China Dataset. Exploratory data analysis, feature extraction and some machine learning algorithms like Support Vector Machine (SVM), K-Nearest Neighbour (KNN) are applied on those datasets to conclude the status of chest x-ray images as which are tb positive and which are tb negative.**

***Index Terms*—Tuberculosis, Chest X-Ray Images, Exploratory Data Analysis, Feature Extraction, SVM Classifier, KNN Classifier**

I. INTRODUCTION

Tuberculosis (TB) is an infectious, bacteria related deadly disease. This disease is caused by Mycobacterium tuberculosis bacteria (MTB). It generally affects the lungs, but can affect other parts of the body also. Most infections do not have any symptoms which is known as latent tuberculosis. About 10% of this latent tuberculosis, progress to active tuberculosis which, if left untreated, kills those affected persons.

The classic symptoms of active TB are a chronic cough with bloodcontaining mucus, fever, night sweats, weight loss etc.

Compared with other diseases caused by a single infectious agent, tuberculosis is the second biggest killer, globally. According to World Health Organisation (WHO) nearly about 9 million people get sick with TB in a year in which 3 million of people "missed" by health system.

The technique in medical science which is followed for detection of any body cell is affected by tuberculosis bacteria or not, is too much lengthy and time consuming. This is another cause which helps TB to become deadly disease in present world. So if there is any process to detect the cell is affected by this bacteria or not, the percentage of death of people affected by TB may be decreased.This process is built up with the help of various machine learning algorithms specially SVM and image processing technique.

Diagnosis of active TB is based on chest X-rays, microscopic examination and fluroscent image experiments.

Initially, image processing technique is applied to preprocess the X-ray images.
Then, machine learning algorithms are applied to detect whether the cell is affected by tuberculosis bacteria or not. It is observed that Support Vector Machine (SVM) classifier performs best for the chosen datasets.

The main moto of SVM is to classify the chest region into 4 stages-normal stage, beginning stage, moderate stage and severe stage and image processing is to examine chest x-ray report. In image processing, graph cut segmentation is used for extraction for lung region and wiener filter is used for removing noise.

II. LITERATURE SURVEY

A. List of Research Papers

Some research papers are attached which is the base of this survey paper. Those research papers are containing information how to detect which cell is affected by this bacteria, what is the procedure to detect the cell is affected or not and also which is the best process in machine learning algorithm and image processing to detect the affected cell. Following table explains about the whole matters.

NO. OF PAPER	TITLE OF PAPER	AUTHOR	JOURNAL/CONFERENCE	YEAR OF PUBLICATION	PROBLEM DISCUSSIONS
P1	ADVANCES IN AUTOMATIC TUBERCULOSIS DETECTION IN CHEST X-RAY IMAGES	WAI YAN NYEIN NAING ET AL.	An International Journal (SIPIJ) Vol.5, No.6	2014	CHEST X-RAY, SEGMENTATION AS PREPROCESSING

P2	ANALYSIS OF TUBERCULOSIS IN CHEST USING SVM CLASSIFIER	S.SIVARANJANI ET AL.	National Conference on Research Advances in Communication, Computation, Electrical Science and Structures(NCRACCESS-2015)	2015	CHEST X-RAY, SEGMENTATION AS PREPROCESSING, MODIFIED LHTGF IN FEATURE EXTRACTION, SVM AS ALGORITHM
P3	An Efficient Feature Extraction Method for Tuberculosis detection using Chest Radiographs[8]	R. Beaulah Jeyavathana et al.	International Journal of Applied Environmental Sciences ISSN 0973-6077 Volume 12[8]	2017	CHEST X-RAY, SEGMENTATION AS PREPROCESSING, MODIFIED LHTGF IN FEATURE EXTRACTION, SVM AS ALGORITHM
P4	Detecting drug-resistant tuberculosis in chest radiographs	Stefan Jaeger et al.	International Journal of Computer Assisted Radiology and Surgery	2018	CHEST X-RAY, SVM AS ALGORITHM
P5	Current Applications and Future Impact of Machine Learning in Radiology	Garry Choy et al.	NA	2018	CHEST X-RAY, SEGMENTATION, DECISION TREE,RANDOM FOREST,KNN, NEIVE BAYES, SVM AS ALGORITHMS
P6	Hybrid RID Network for Efficient Diagnosis of Tuberculosis from Chest X-rays	Rabia Rashid et al.	IEEE PAPER	2018	CHEST X-RAY, SVM AS ALGORITHM
P7	Combining Deep Convolutional Neural Network with Support Vector Machine to Classify Microscopic Bacteria Images[7]	TASNIM AHMED ET AL.	International Conference on Electrical, Computer and Communication Engineering (ECCE)[7]	2019	MICROSCOPIC BACTERIA IMAGES, SVM AS ALGORITHM
P8	Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis[9]	Rakhmetulayeva S.B. et.al.	The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018)[9]	2018	FLUROSCENT IMAGES, SVM AS ALGORITHM
P9	Tuberculosis Bacteria Detection based on Random Forest using Fluorescent Images[10]	Chi Zheng et al.	9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics[10]	2016	FLUROSCENT IMAGES, RANDOM FOREST, SVM AS ALGORITHMS

P10	A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification	Asha T et al.	NA		DECISION TREE, RANDOM FOREST, KNN, NEIVE BAYES AS ALGORITHMS
P11	Tuberculosis Recognition &- it's Analysis using Adaptive Neuro Fuzzy Inference System-ANFIS[6]	Reema Sharma et al.	International Conference on Energy, Communication, Data Analytics and Soft Computing(ICECDS-2017)[6]	2017	MACHINE LEARNING ALGORITHMS

B. Types of Data

Paper ID	Data Source	Type of Data	Description
P1	Manually collected	Chest X-Ray	Image of Data
P2	Institute of Respiratory Medicine(IPR), Malaysia, Montogomery DB(US)	Chest X-Ray	CT(Computed Tomography) images
P3	Montgomery DB and China DB(publicly available)	Chest X-Ray	Image of Data
P4	Manually coleected	Chest X-Ray	X-Ray Image of Data
P5	Bone age based Data	Chest X-Ray	Image of Data
P6	Shenzhen Dataset	Chest X-Ray	662 chest x-rays in which 326 are normal and 336 are abnormal
P7	Manually collected	Microscopic Bacteria Images	800 image samplesof seven separate bacteria spices
P8	Patients come hospital for collection	Numerical	Calculating math problem
P9	Msnually collected data	Fluorescent Microscopic Images	Devices set up in sunny optics. which consists a personal computer, a fluorescent microscope and a digital CMOS camera
P10	Manually collected	Numerical Data, Categorical Data	Based on 12 attributes like name,age, loss of weight etc.
P11	Manually collected	nerve	Involving nervous system

C. Machine Learning Algorithm

Paper ID	Algorithm	Performance Measure
P1	SVM, Decision Tree	SVM: establish hyper plane in the space where input is given Decision tree: Single decision is decided by each node.
P2	SVM	Multi-class SVM
P3	KNN,SVM	SVM: better
P4	SVM	Used
P5	SVM	Used
P6	SVM	Used

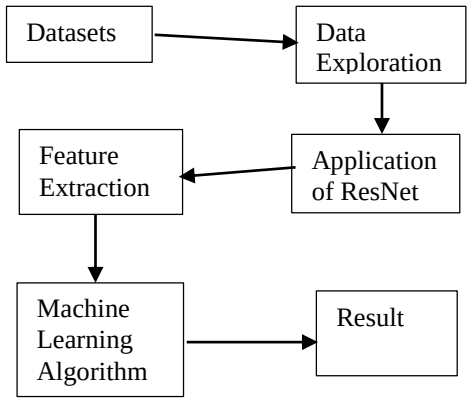
P7	SVM	Used
P8	SVM	Used
P9	SVM, Random Forest	High: Random forest(Based on MGB and HoG) Cross Validation SVM(Training data)
P10	SVM	Used
P11	SVM	Used

III. EXPERIMENTAL RESULTS

A. Datasets

There are two datasets which contain chest x-ray reports. One is Montgomery datasets which contains Montgomery Country – Chest X-ray database and another is China datasets which contains Shenzhen set – Chest X-ray database. Montgomery set contains total 138 images whereas 58 cases with manifestation of tuberculosis and 80 with normal cases. China set contains total 662 number of images where 336 images are manifestation of tuberculosis and rest 326 are images with normal cases. Image file names are coded as CHNCXR_#####_0/1.png for China datasets and MCUCXR_#####_0/1.png for Montgomery datasets, where '0' represents the normal and '1' represents the abnormal lung. All images are in .png format.

The clinical readings of the x-rays are saved as text file following the same file format: CHNCXR_#####_0/1.txt for China dataset and MCUCXR_#####_0/1.txt for Montgomery datasets. Each text file contains the patient's age, gender, and abnormality of the lung.



Block Diagram

B. Exploratory Data Analysis

The data in both datasets are collected from hospitals. So, there may be some null values in the datasets. First, null values are removed. Images for tuberculosis positive and tuberculosis negative are identified with the class '0/1'. Histogram is plotted on the basis of age, gender to find out at which age or in which gender tuberculosis is mainly affected.

Positive cases (denoted as 1) and Negative cases (denoted as 0):

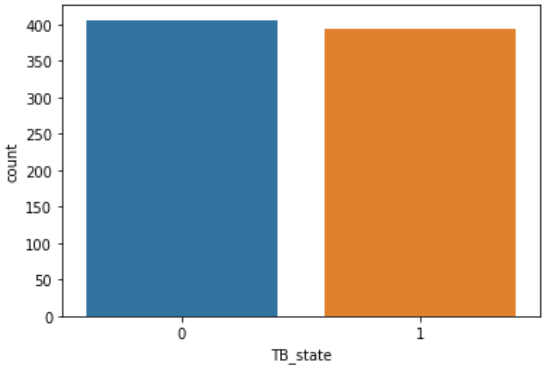


Fig 1: Count of tuberculosis positive and tuberculosis negative cases

In figure 1, bar plot is drawn for counting tuberculosis positive and tuberculosis negative cases. Blue colored bar denoted as 0 represents the count for tuberculosis negative patients and orange colored bar graph denoted as 1 represents tuberculosis positive patients.

China dataset(denoted as CHNCXR) and Montgomery dataset(denoted as MCUCXR):

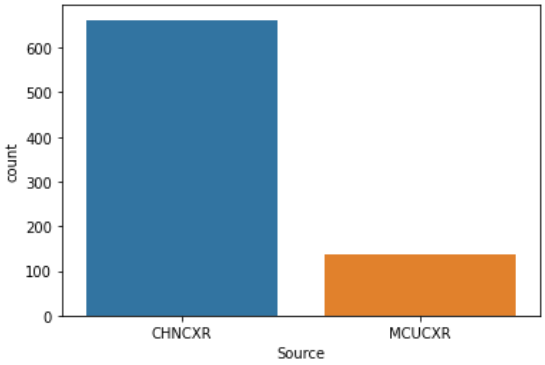


Fig 2: number of data in datasets

In figure 2, bar graph represents the number of cases presents in China and Montgomery datasets. Blue color indicates China dataset and orange color indicates Montgomery dataset.

Gender wise positive and negative cases:

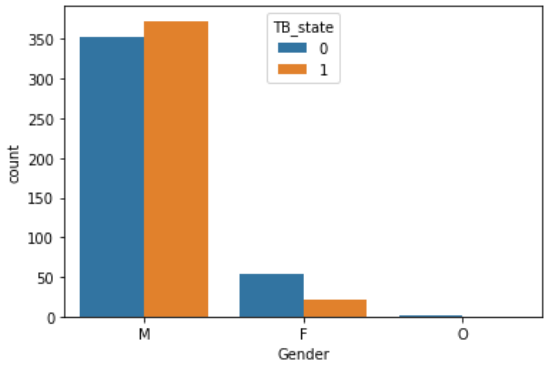


Fig 3: number of positive and negative cases according to gender

In figure 3, side by side bar plot describes that how many male, female and other gender persons are affected and unaffected by tuberculosis.

Tuberculosis Positive cases according to ages:

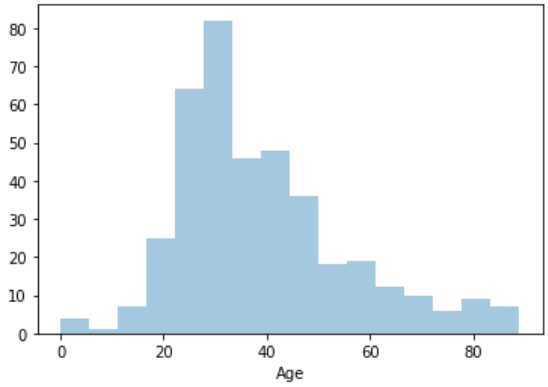


Fig 4: Variation with ages

In figure 4, histogram plot describes the variation of ages of patients who are affected by tuberculosis.

C. Feature Extraction

Deep learning is applied for feature extraction of the images. Resnet, VGG-16 methods are used as a method of feature extractions. Both methods are applied and results are stored in different variables.

D. Machine Learning Algorithm

A few machine learning algorithms are applied on the extracted image datasets. First Have it as uniformly SVM. Then do the change wherever applicable and find out the accuracy value using svm classifier. After that same process is done for knn classifier and random forest.

SVM(Support Vector Machine): SVM is a supervised machine learning algorithm to find a hyperplane in n-dimensional space to distinctly classify the data points.

KNN(K-Nearest Neighbors) algorithm is a machine learning algorithm which is used for both classification and regression model. This algorithm mainly uses feature similarity method to predict the values of the new data points. These new data points are mainly considered as test datasets. The value of test datasets are measured with the help of the training datasets. Different types mainly Euclidean distance is used to measure the distance. As a result there raises some error.

Random Forest: Random Forest is such a machine learning algorithm which fits a number of decision trees on various sub samples of datasets.

E. Result

SVM:

Classification Report:

	Precision	Recall	F1-score	Support
Class 0	0.72	0.75	0.73	88
Class 1	0.68	0.64	0.66	72
Accuracy			0.70	160
Macro avg	0.70	0.69	0.70	160
Weighted avg	0.70	0.70	0.70	160

AUROC(Area Under The Receiver Operating Characteristics):- To check or to visualize the performance of the multi-class classification problem, AUC(Area Under The Curve) ROC(Receiver Operating Characteristics) curve is used. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much a model is capable of distinguishing classes. AUC value tends to 1 gives more accurate value. In this paper, higher the AUC, better the model is at distinguishing between patients with tuberculosis and no tuberculosis. The ROC curve is plotted with TPR against FPR where TPR(True Positive Rate) is y-axis and FPR(False Positive Rate) is on the x-axis.

TPR=TP/TP+FN

FPR=FP/TN+FP

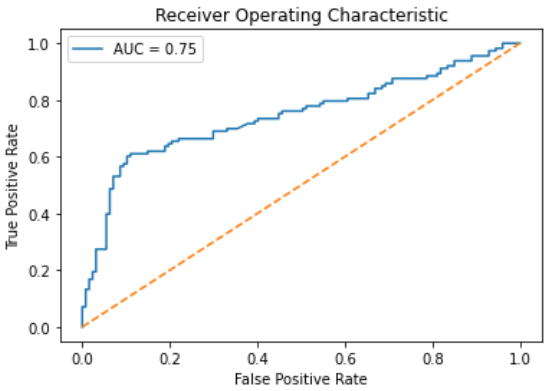


Fig 5: ROC curve (True Positive Rate vs False Positive Rtae)

In figure 5, accuracy level is 0.75 which is near to 1. So, by applying this algorithm tuberculosis affected or unaffected patients can be distinguishable.

KNN:

Classification Report:

	Precision	Recall	F1-score	Support
Class 0	0.79	0.88	0.83	88
Class 1	0.83	0.72	0.77	72
Accuracy			0.81	160
Macro avg	0.81	0.80	0.80	160
Weighted avg	0.81	0.81	0.80	160

According to KNN algorithm, there raises some error at the time to compare test data with training data.These error is varying with assign the k-value. Here this graph is titled as ‘error rate k value’ where k-value is placed in x-axis and error is placed in y axis

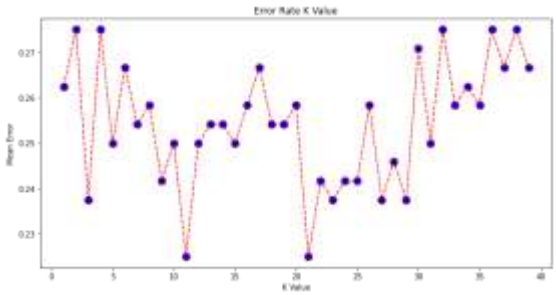


Fig 6: Mean value vs K-value

In figure 6, k value is variable with the mean error value.

Random Forest:

Acccuracy score is 0.78 and roc_value is 0.8

IV. Conclusion And Future Work

The results prove that the proposed methodology is a promising tool for diagnosis of Tuberculosis from chest x-ray images. Combining machine learning algorithms like svm, knn, random forest with deep learning algorithms like convolution neural network, resnet, vgg-16 improve the accuracy of given dataset. The process is done with using any perticuler machine learning algorithm like only svm or only knn with any particular deep learning algorithm like only resnet or only vgg-16. But using multiple algorithms in both deep learning and machine learning field, it becomes easy to figure out the most accurate process to detect tuberculosis positive x-ray image. But there is one drawback that deep learning models are very time consuming. As feature is extracted by deep learning method, application of any machine learning algorithm on that extracted feature gives almost same result.

References

[1] https://www.mayoclinic.org/diseases-conditions/tuberculosis/symptoms-causes/syc-20351250
[2] https://www.learnopencv.com/svm-using-scikit-learn-in-python/
[3] https://machinelearningmastery.com/support-vector-machines-for-machine-learning/
[4] https://www.kaggle.com/qyleong13/tb-dataset
[5] https://doi.org/10.1148/radiol.2018171820
[6] International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)
[7] International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019
[8] International Journal of Applied Environmental Sciences ISSN 0973-6077 Volume 12
[9] The 9th International Conference on Ambient Systems, Networks and Technologies(ANT 2018)
[10] 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics
[11] https://medium.com/@dataturks/understanding-svms-for-image-classification-cf4f01232700
[12] http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/142-knn-k-nearest-neighbors-essentials/
[13] https://discuss.analyticsvidhya.com/t/how-to-choose-the-value-of-k-in-knn-algorithm/2606/5
[14] https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76
[15] https://androidkt.com/resnet-implementation-in-tensorflow-keras/
[16] https://www.modelzoo.co/model/resnet
[17] https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/
[18] https://in.mathworks.com/matlabcentral/answers/123999-images-classification-using-svm-classifier
[19] https://www.researchgate.net/publication/265151934_Image_Classification_using_Support_Vector_Machine_and_Artificial_Neural_Network
[20] https://medium.com/@YearsOfNoLight/intro-to-image-classification-with-knn-987bc112f0c2