VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BANGALORE INSTITUTE OF TECHNOLOGY

Department of
INFORMATION SCIENCE &ENGINEEERING

Final Year Project
on

TITLE : AUTOMATIC SCREENING OF TUBERCULOSIS USING CHEST RADIOGRAPHS

Presented By :

Aishwarya Mohan	**1BI15IS002**
Shabaz Khan	**1BI15IS038**
Smitha G M	**1BI15IS046**
Vignesh Raam B S	**1BI15IS058**

Under the Guidance of :

Mrs. H. Roopa
Assitant Professor
Dept of ISE,BIT

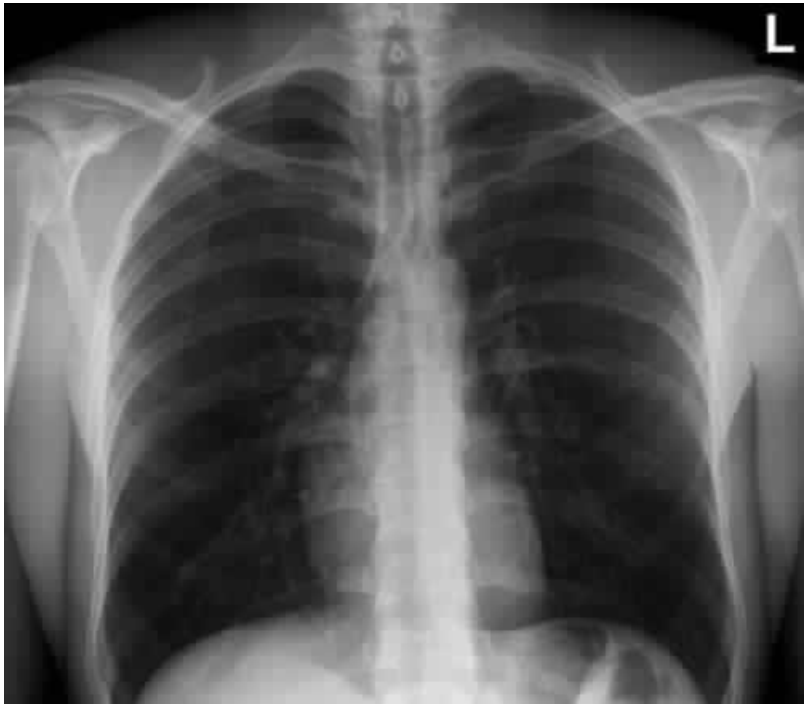
CONTENTS

- Introduction
- Literature Survey
- System Requirements
- System Design and Architecture
- Implementation
- Results and Discussions
- References

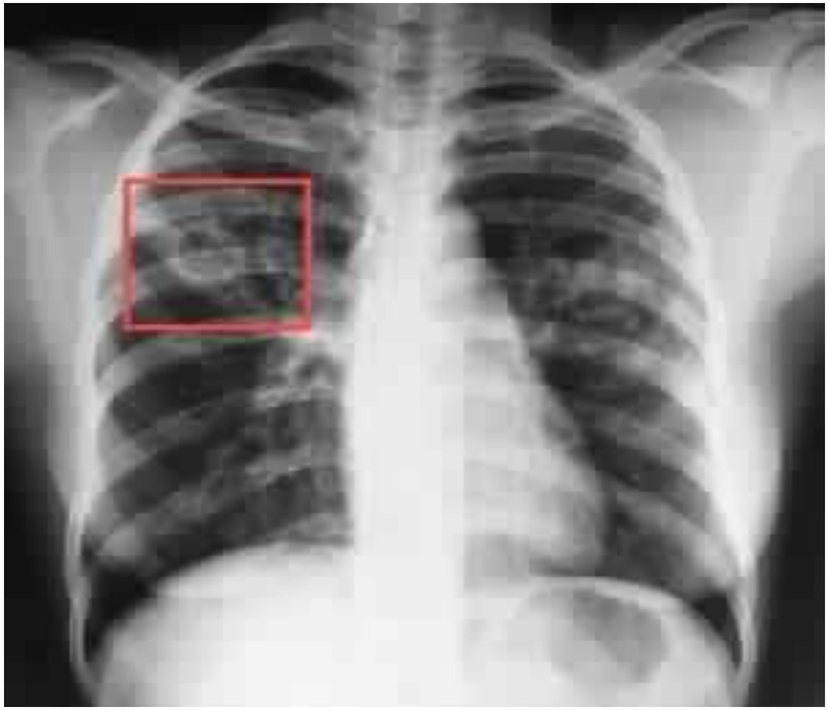
INTRODUCTION

- Tuberculosis is a chronic disease characterized by caseous necrosis and granuloma formation that affects lungs causing cough with hemoptysis (bloody sputum).
- The project presents an automated approach to detect tuberculosis using chest radiographs.
- Sample set of chest radiographs are taken, which contains radiographs of both healthy persons and Tuberculosis infected patients.
- From each of these radiographs the affected & not affected lung region is extracted using the segmentation method.
- These ROI are then used for shape and texture based feature extraction.
- Using these features, a classification model is trained using DM algorithms to classify the radiograph as TB infected or not.
- For testing, the chest radiographs of patients are run through above processes and then verified with the trained classification model to detect Tuberculosis.

EXAMPLE



Normal chest radiograph



Tuberculosis affected chest radiograph

LITERATURE SURVEY

SL NO	TITLE	AUTHOR	METHODOLOGY
1	Image Segmentation of Ziehl-Neelsen Sputum Slide Images for Tubercle Bacilli Detection	R. A. A. Raof M. Y. Mashor R. B. Ahmad and S. S. M. Noor	This paper talks about the traditional manual methods of segmentation and techniques involved.
2	Novel Fuzzy Association Rule Image Mining Algorithm for Medical Decision Support System	P Rajendran and M Madheswaran	This paper talks about the mining algorithms in medicinal support systems.
3	Size of transaction-based association rule mining algorithm	Asha Pandian	This paper talks about the Associative Rule Mining Algorithms.

Existing System

- There is a dearth of trained staff who can quickly read CXR images, and there is a high degree of inter-reader variability, resulting in both over and under-diagnosis of tuberculosis.
- These limitations have generated interest in scoring systems for chest radiography as well as a growing interest in automatic systems to detect tuberculosis using chest radiographs.

Proposed System

- Automated Screening of Tuberculosis is modeled to quickly evaluate digital radiographs to improve the efficiency of detection of tuberculosis.
- In our approach, a sample set of chest radiographs are taken, which contains radiographs of both healthy persons and tuberculosis infected patients.
- These radiographs are pre-processed to reduce noise and segmented to obtain region of interest.
- The ROI's are then used for shape and texture-based feature extraction.
- Using the extracted feature, Classification model is trained using various DM algorithms to effectively classify the new chest radiographs.
- The results of various DM algorithms are compared to evaluate which classification model performs well with the collected data.

SYSTEM REQUIREMENTS

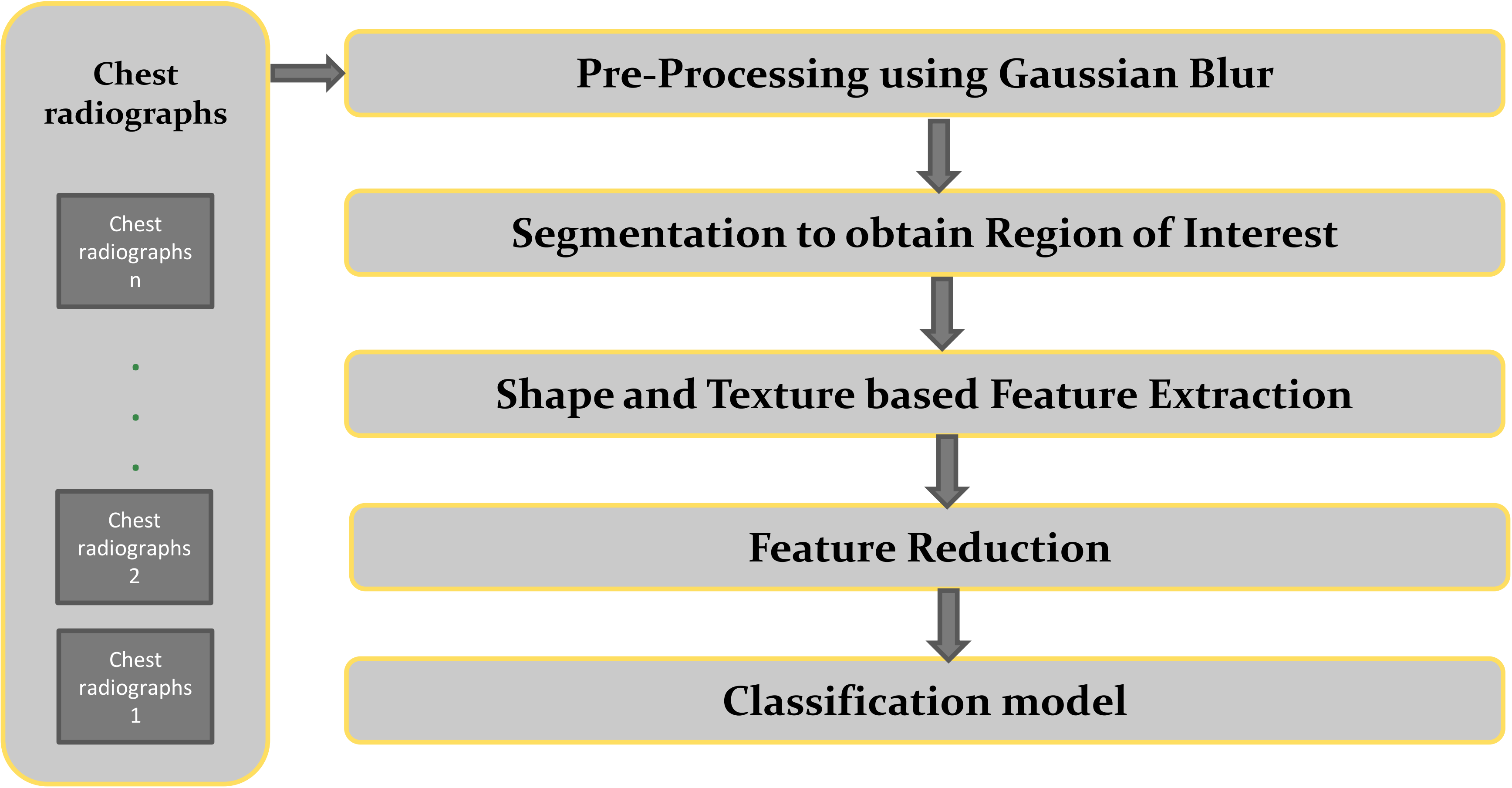
Software Requirements (minimum):

Name of the Component	Specification
Operating System	Windows XP,Vista,7,8,10/Linux
Computing Environment	MATLAB R2014a
Database	MYSQL / MongoDB
Browser	Google Chrome/ Mozilla / Opera etc.
Web Server	Apache 2.0 OR ABOVE
Software Development Kit	XAMPP 7.1 OR ABOVE
Scripting Language	HTML, CSS, PHP
Database Connection	PHP Connection

SYSTEM DESIGN AND ARCHITECTURE

- System design is the process of defining the architecture , modules, interfaces and data for a system to satisfy specifies requirements.
- System architecture is the conceptual model that defines the structure, behavior, and more views of a system.
- A system architecture can consists of system components and the sub-systems developed that will work together to implement the overall system.

System Design



Chest radiographs

Chest radiographs n

.

.

.

Chest radiographs 2

Chest radiographs 1

Pre-Processing using Gaussian Blur

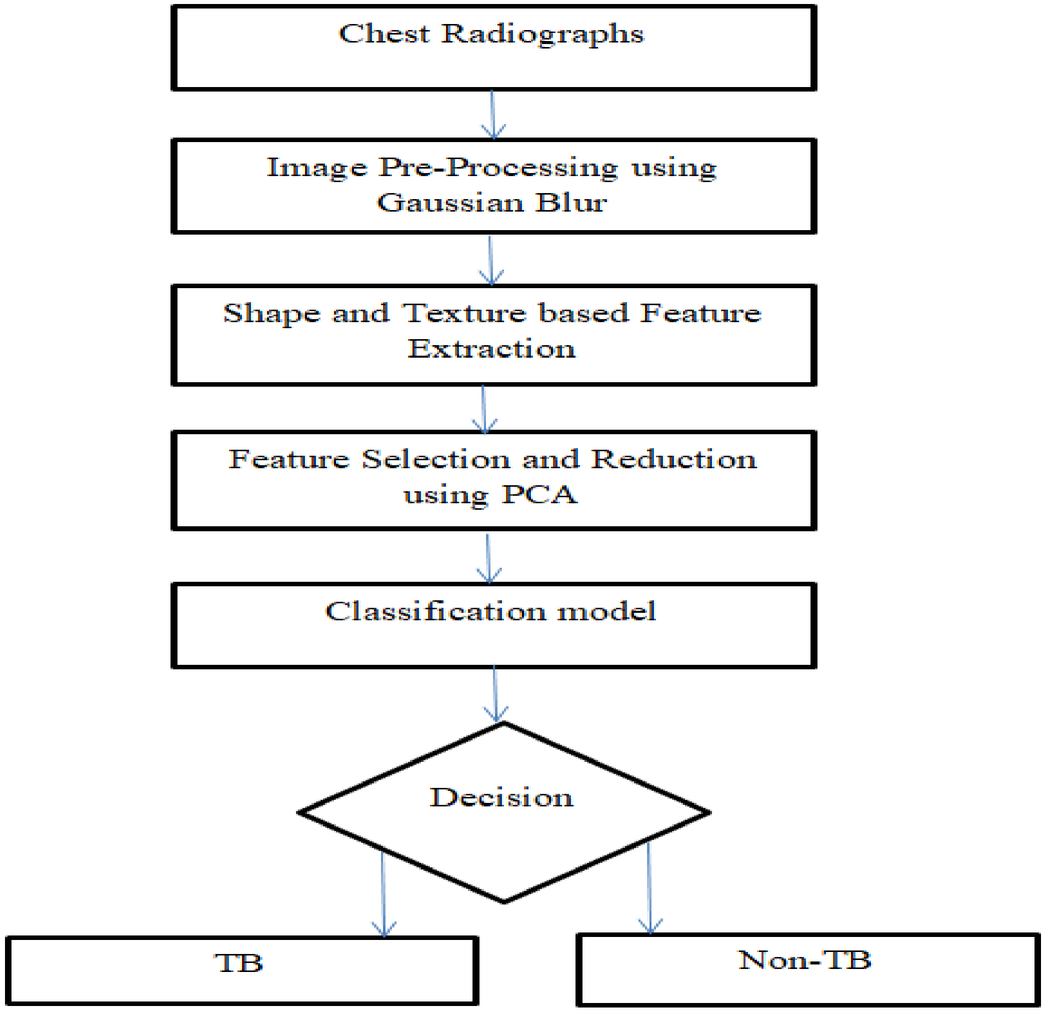
Segmentation to obtain Region of Interest

Shape and Texture based Feature Extraction

Feature Reduction

Classification model

Methodology


Chest Radiographs
Image Pre-Processing using Gaussian Blur
Shape and Texture based Feature Extraction
Feature Selection and Reduction using PCA
Classification model
Decision
TB
Non-TB

IMPLEMENTATION

MODULE-WISE BREAKDOWN

- Module 1 : Data Collection
- Module 2 : Pre-Processing
- Module 3 : Segmentation
- Module 4 : Feature Extraction
- Module 5 : Feature Reduction using PCA
- Module 6 : Classification Model

Module 1

Data Collection

- Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.
- The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.
- For the project, the data was collected from the hospitals, senior doctors and from The National Centre for Biotechnology Information (NCBI), apart of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH).

Module 2

Pre-Processing

- Image pre-processing is a method to convert an image into digital form in order to get an enhanced image or to extract some useful information from it
- The purpose of image processing is divided into 5 groups. They are:

  1. Visualization - Observe the objects that are not visible.

  2. Image sharpening and restoration - To create a better image

  3. Image retrieval - Seek for the image of interest.

  4. Measurement of pattern – Measures various objects in an image.

  5. Image Recognition – Distinguish the objects in an image

Gaussian Blur

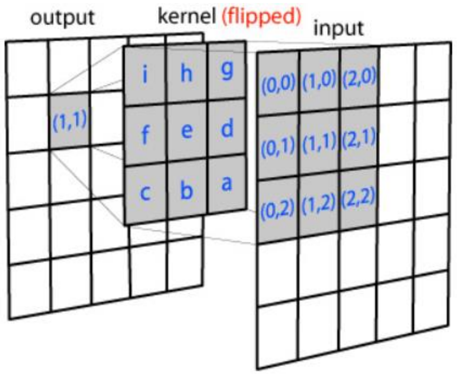
- The Gaussian blur is a type of image-blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image.
- Used to blur the image or to reduce noise
- The advantage of Gaussian filter is multiplying and adding is probably faster than sorting

Example :

- Gaussian blur uses kernel functions to blur the image.
- A Gaussian kernel (used for Gaussian blur) is a square array of pixels where the pixel values correspond to the values of a Gaussian curve (in 2D).

$\frac{1}{273}$

1	4	7	4	1
4	16	26	16	4
7	26	41	26	7
4	16	26	16	4
1	4	7	4	1



- Each pixel in the image gets multiplied by the Gaussian kernel which is done by placing the center pixel of the kernel on the image pixel and multiplying the values in the original image with the pixels in the kernel that overlap.
- The values resulting from these multiplications are added up and that result is used for the value at the destination pixel.

Module 3

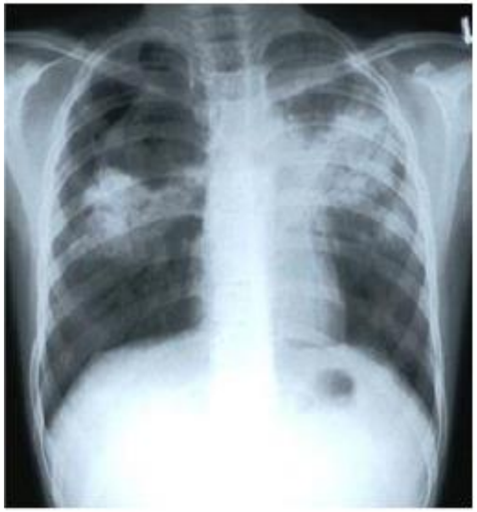
Segmentation

- Image segmentation is a procedure for extracting the region of interest (ROI) through an automatic or semi-automatic process.
- Image segmentation is used to locate objects and boundaries (lines, curves, etc.) in images.
- In medical research, segmentation can be used in separating different tissues from each other, through extracting and classifying features.
- We are using mark function to obtain the Region of Interest in the project.
- The various segmentation techniques are Thresholding, Watershed Transformation, Clustering, Semi automatic segmentation and many more.

Graythresh(I)

- Graythresh(I) computes global threshold T, from grayscale image I, using Otsu’s method.
- Otsu’s method chooses a threshold that minimizes the intra-class variance of the thresholded black and white pixels.
- Graythresh is employed in the project in order to detect cavities and consolidations.
- These regions will have to marked in order to extract region of interest.

Example Of Segmentation



TB infected image

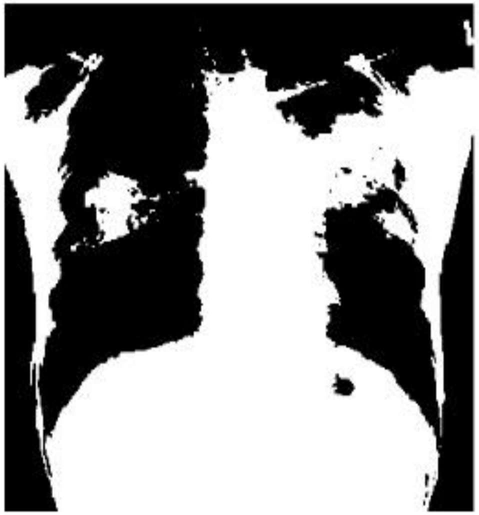
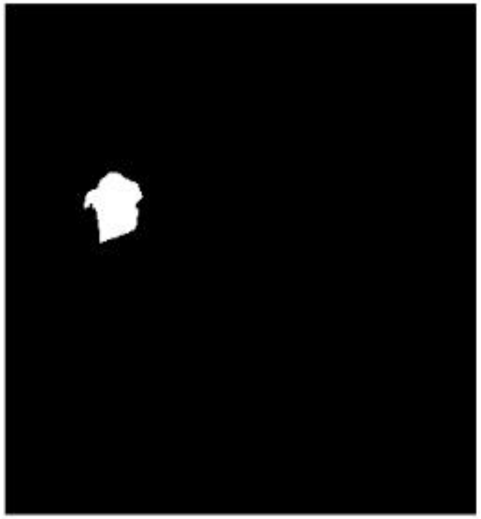


Image after applying Graythresh



Masked image

Module 4

Feature Extraction

- Feature extraction starts from an initial set of measured data and builds derived values intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations.
- These features are used to identify the image characteristics, for the purpose of classification.
- The various features extracted entropy, energy, correlation, homegeneity, contrast, area ,orientation etc.

Feature Extraction

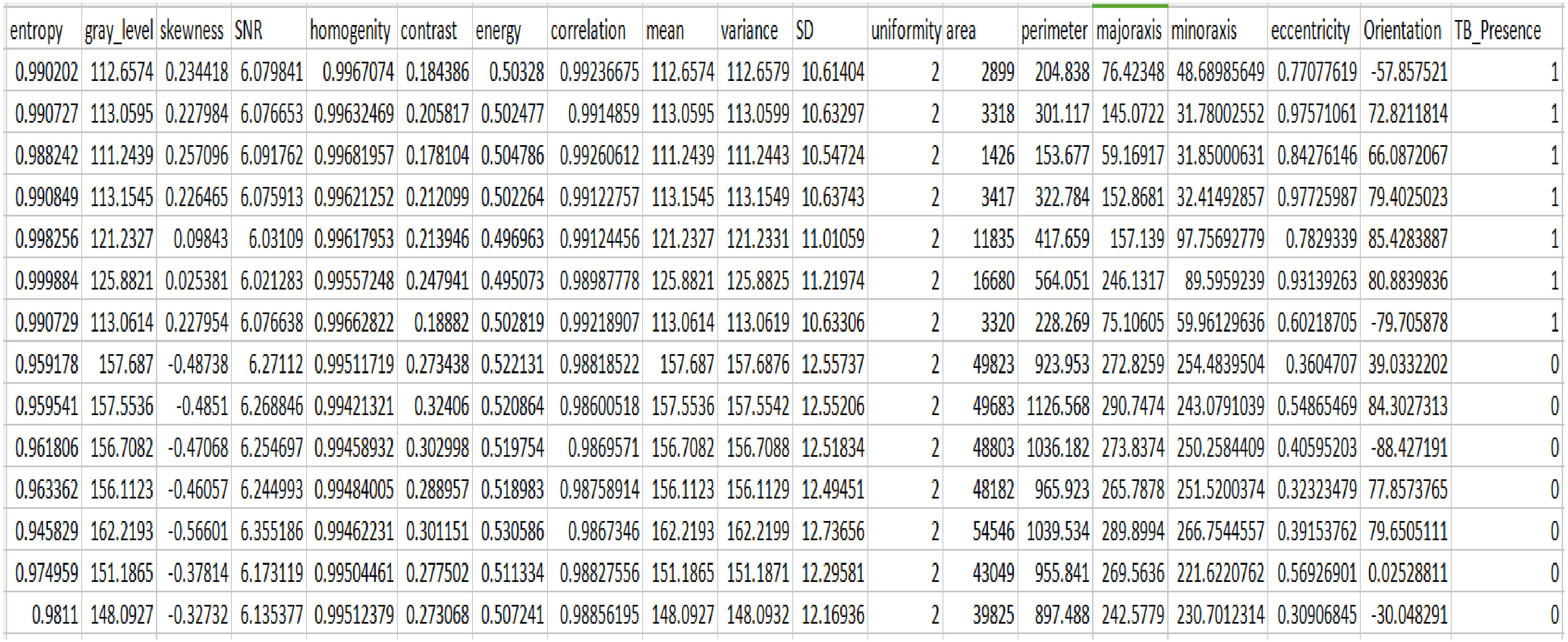
- The masked TB and normal images are fed as inputs to the feature extraction module.
- The features gives distinct values for TB and normal images.
- This module extracts 18 features from each image.

Entropy	Gray Level	Energy	Homogeneity	Contrast	Correlation
Skewness	Mean	Variance	Standard Deviation	Major Axis	SNR
Minor Axis	Eccentricity	Orientation	Uniformity	Area	Perimeter

- The features are exported to an excel sheet.
- The excel sheet serves as a database for the purpose of classification.

Feature Extraction in Project

The following figure show the dataset formulated after the extraction of features.



entropy	gray_level	skewness	SNR	homogenity	contrast	energy	correlation	mean	variance	SD	uniformity	area	perimeter	majoraxis	minoraxis	eccentricity	Orientation	TB_Presence
0.990202	112.6574	0.234418	6.079841	0.9967074	0.184386	0.50328	0.99236675	112.6574	112.6579	10.61404	2	2899	204.838	76.42348	48.68985649	0.77077619	-57.857521	1
0.990727	113.0595	0.227984	6.076653	0.99632469	0.205817	0.502477	0.9914859	113.0595	113.0599	10.63297	2	3318	301.117	145.0722	31.78002552	0.97571061	72.8211814	1
0.988242	111.2439	0.257096	6.091762	0.99681957	0.178104	0.504786	0.99260612	111.2439	111.2443	10.54724	2	1426	153.677	59.16917	31.85000631	0.84276146	66.0872067	1
0.990849	113.1545	0.226465	6.075913	0.99621252	0.212099	0.502264	0.99122757	113.1545	113.1549	10.63743	2	3417	322.784	152.8681	32.41492857	0.97725987	79.4025023	1
0.998256	121.2327	0.09843	6.03109	0.99617953	0.213946	0.496963	0.99124456	121.2327	121.2331	11.01059	2	11835	417.659	157.139	97.75692779	0.7829339	85.4283887	1
0.999884	125.8821	0.025381	6.021283	0.99557248	0.247941	0.495073	0.98987778	125.8821	125.8825	11.21974	2	16680	564.051	246.1317	89.5959239	0.93139263	80.8839836	1
0.990729	113.0614	0.227954	6.076638	0.99662822	0.18882	0.502819	0.99218907	113.0614	113.0619	10.63306	2	3320	228.269	75.10605	59.96129636	0.60218705	-79.705878	1
0.959178	157.687	-0.48738	6.27112	0.99511719	0.273438	0.522131	0.98818522	157.687	157.6876	12.55737	2	49823	923.953	272.8259	254.4839504	0.3604707	39.0332202	0
0.959541	157.5536	-0.4851	6.268846	0.99421321	0.32406	0.520864	0.98600518	157.5536	157.5542	12.55206	2	49683	1126.568	290.7474	243.0791039	0.54865469	84.3027313	0
0.961806	156.7082	-0.47068	6.254697	0.99458932	0.302998	0.519754	0.9869571	156.7082	156.7088	12.51834	2	48803	1036.182	273.8374	250.2584409	0.40595203	-88.427191	0
0.963362	156.1123	-0.46057	6.244993	0.99484005	0.288957	0.518983	0.98758914	156.1123	156.1129	12.49451	2	48182	965.923	265.7878	251.5200374	0.32323479	77.8573765	0
0.945829	162.2193	-0.56601	6.355186	0.99462231	0.301151	0.530586	0.9867346	162.2193	162.2199	12.73656	2	54546	1039.534	289.8994	266.7544557	0.39153762	79.6505111	0
0.974959	151.1865	-0.37814	6.173119	0.99504461	0.277502	0.511334	0.98827556	151.1865	151.1871	12.29581	2	43049	955.841	269.5636	221.6220762	0.56926901	0.02528811	0
0.9811	148.0927	-0.32732	6.135377	0.99512379	0.273068	0.507241	0.98856195	148.0927	148.0932	12.16936	2	39825	897.488	242.5779	230.7012314	0.30906845	-30.048291	0

Module 5

Feature Reduction Using PCA

- Dimensionality reduction is the process of reducing the number of random variables.
- Reduces the problem of curse of dimensionality
- Feature projection transforms the data in the high-dimensional space to a space of fewer dimensions.
- Principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.

Principal Component Analysis

- PCA is a method of extracting important variables from a large set of variables available in a data set.
- Extracts low dimensional set of features, called principal components from a high dimensional data set with a motive to capture as much information as possible.
- The accuracy of the system is enhanced after applying PCA.

PCA Algorithm

- Standardization
- Computing Covariance matrix
- Computing Eigen values and Eigen Vectors to identify Principal Components
- Recast the data along principal component axes

Module 6

Data Mining

- Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and systems.
- An overall goal is to extract information (with intelligent methods) from a data set and use it to derive required results.
- The different data mining techniques are:

1. Supervised Learning : Uses the dataset to train the DM model

   Example : Classification and Regression Algorithms

2. Unsupervised Learning : No previous data examples are used in training.

   Example : Clustering and Association Rules

Regression

- Regression analysis is a form of predictive modeling technique that investigates the relationship between a dependent (target) and independent variable(s) (predictor).
- We have two kinds of Regression models. They are

  1. Linear Regression
  2. Logistic Regression

- **Linear Regression** establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).It used Mean Squared Error to predict the outcome
- **Logistic regression** is a statistical method for analyzing a dataset that has one or more independent variables which determine an outcome. It used Log-Loss function to Predict the outcome.
- Logistic Regression works better compared to Linear Regression when the target variable is a binary value (yes/no)

Logistic Regression

- The logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- The equation for logistic regression is :

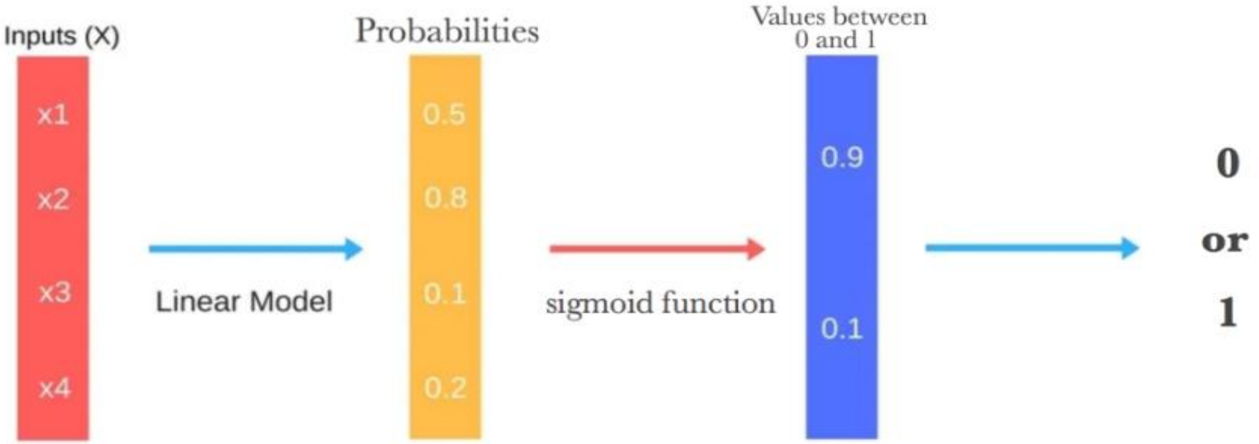
  $\text{logit}(P) = b_0 + b_1 x_1 + b_2 X_2 + \ldots\ldots + b_n X_n$, where

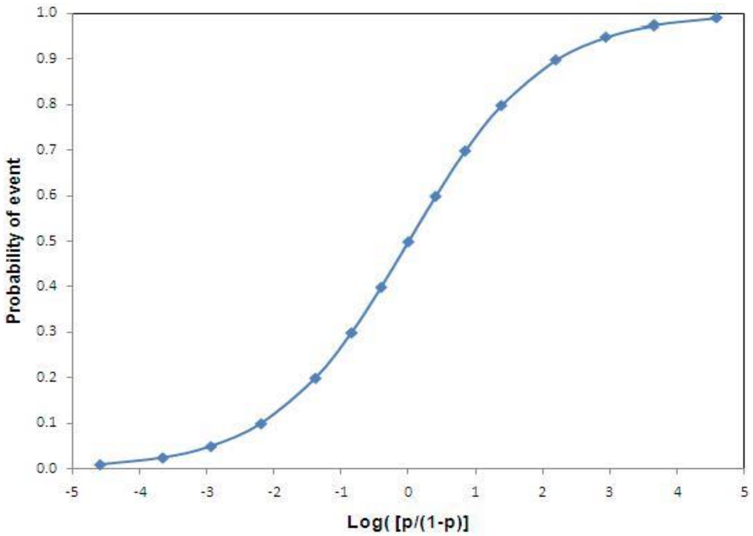
  $\text{logit}(P) = \ln(P/1-P)$
- Sigmoid function is used to map Predicted values to probabilities. Given By

  $S(z)=1/1+e^{-z}$

  - s(z) = output between 0 and 1 (probability estimate)
  - z = input to the function (your algorithm’s prediction e.g. mx + b)
  - e = base of natural log

How does Logistic Regression Works



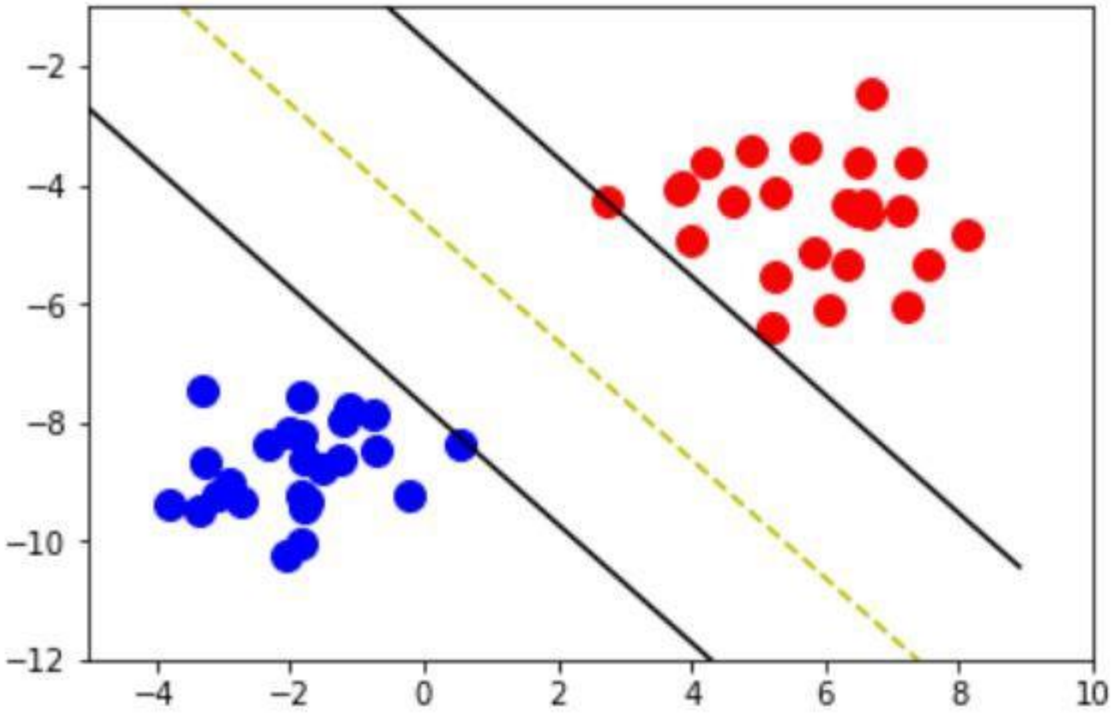


Support Vector Machines

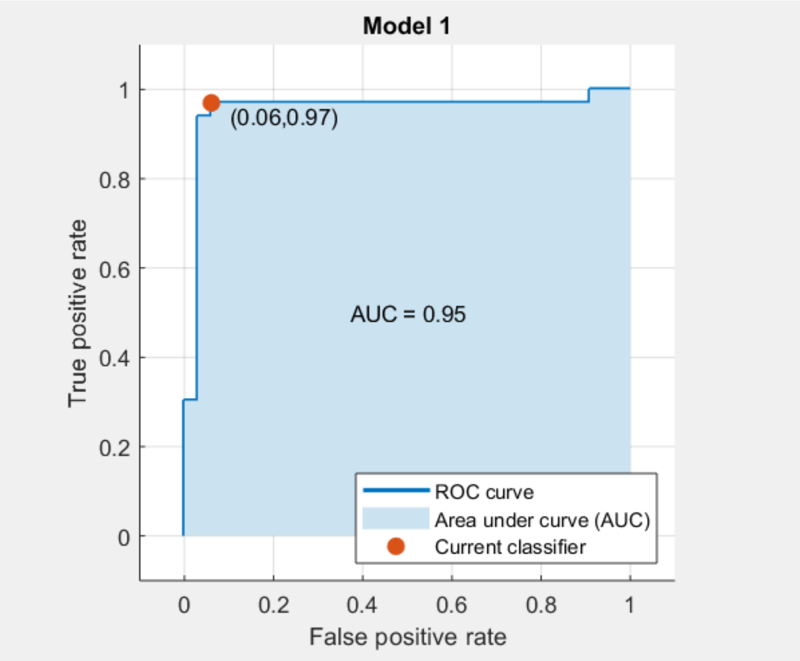
- An SVM model is a representation of the dataset as points in space, mapped so that the data of the separate categories are divided by a clear gap that is as wide as possible.
- New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
- The objective of the support vector machine algorithm is to find a hyper plane in an N-dimensional space(N—the number of features) that distinctly classifies the data points.
- In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values([-1,1]) which acts as margin.

- **Linear Kernel SVM**
  - The dot-product called the kernel can be written as K(x, xi) = sum(x * xi)
  - The kernel defines the similarity or a distance measure between new data and the support vectors.
  - Other kernels can be used that transform the input space into higher dimensions such as a Polynomial Kernel and a Radial Kernel. This is called the Kernel Trick.
- **Polynomial Kernel SVM**
  - Instead of the dot-product, we use a polynomial kernel, K(x,xi) = 1 + sum(x * xi)^d
  - Where the degree of the polynomial must be specified by hand to the learning algorithm. When d=1 this is the same as the linear kernel. The polynomial kernel allows for curved lines in the input space.

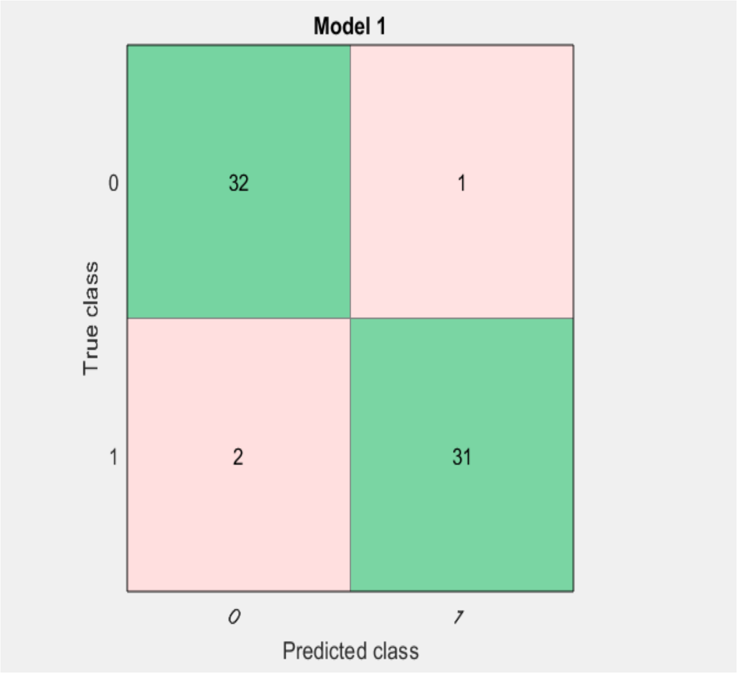
Linear SVM Hyperplane



RESULTS



Region of Curve



	0	1
0	32	1
1	2	31

Confusion Matrix

Future Enhancement

- The proposed system can be enhanced in order to detect other pulmonary disorders.
- The proposed system make use of manual intervention to prune the region of interest which can be automated using efficient segmentation techniques.
- The logistic regression gives considerably good accuracy which can be further enhanced for the better prediction.

CONCLUSION

- In the proposed system, the Chest radio graphs are taken and the unwanted information that is inconsistent to the proposed work is eliminated using the pre-processing technique.
- From the pruned images the Regions of Interest are extracted and the attribute values are derived for detected Regions of Interest (ROI).
- The logistic regression algorithm is applied on the attributes that are extracted from the ROI and on applying the logistic regression algorithm, we identify if CXR are infected or not.
- The results derived from the technique provide better efficiency that helps the physicians for making better decision by providing the keywords and identifying if chest radiographs are infected by TB or not.

REFERENCES

[1]. Raof, R. A. A., M. Y. Mashor, R. B. Ahmad, and S. S. M. Noor. "Image segmentation of Ziehl-Neelsen sputum slide images for tubercle bacilli detection." In *Image Segmentation*. IntechOpen, 2011.

[2]. Asha, T., S. Natarajan, and K. N. B. Murthy. "Data mining techniques in the diagnosis of tuberculosis." In *Understanding tuberculosis-global experiences and innovative approaches to the diagnosis*. IntechOpen, 2012.

[3]. Zyout, Imad, Joanna Czajkowska, and Marcin Grzegorzek. "Multi-scale textural feature extraction and particle swarm optimization based model selection for false positive reduction in mammography." *Computerized Medical Imaging and Graphics* 46 (2015): 95-107.

[4]. Rajendran, P., and M. Madheswaran. "Novel fuzzy association rule image mining algorithm for medical Decision support system." *International Journal of Computer Applications* 1, no. 20 (2010): 87-94.

THANK YOU