An Efficient Prediction Model to Analyze Tuberculosis Chest X-ray Images

Roopa H, Dr. Asha T

Abstract - Infectious disease like tuberculosis can be diagnosed by its symptoms like weight loss, cough, fever, x-ray image finding's, etc. X-ray is one of the ways to identify whether the patient is infected by tuberculosis or not. Features representing the cause of tuberculosis are examined from x-ray and these features are chosen out of it by implementing concepts of image processing. These features are then modeled using multiple linear regressions to obtain an accuracy of 96.07%. Area of all the considered features related to the disease are calculated using Simpson's rule and then modeled using multiple linear regression to get an improved accuracy of 97.4%. Thus improved classification accuracy is obtained when numerical and statistical techniques are applied together for a x-ray image to classify tuberculosis disease.

Index Terms - Tuberculosis, Linear Regression Model, Simpson's Rule.

1. INTRODUCTION

Among many infectious diseases, tuberculosis (TB) [1] is one of them caused by mycobacterium. TB can be identified by symptoms like fever, cough, weight loss, x-ray image finding's, etc. TB usually affects lungs but can also infect other parts of the body like kidney, bone, spine, brain, etc. To create an awareness of TB, March 24 is observed as world TB day . The level of TB is analyzed by World Health Organization (WHO) [2] for all countries in the world.

To find relationship between two or more attributes in any dataset, Linear Regression Model (LRM) is used. The output that represents the relationship between changes of dependent attribute with respect to change in independent attribute is called a regression value predicted. When dependent attribute either increases or decreases throughout with the changing attribute in the independent manner, then the predicted value is said to be linear.

More than one independent variables in any data set is analyzed and predicted attribute value is computed by same process as that of LRM which is known as multiple LRM. Classification performance of any medical data can be improved if numerical and statistical methods are combined.

To analyze the concept of numerical and statistical model on TB data is the main aim of our work. In this paper, section 2 describes about related research work, section 3 explains about proposed method, section 4 gives details of experimental implementation and section 5 concludes the work.

2. RELATED WORK

Image mining framework explained by Perner et al [3] is very helpful in analyzing any images in medical field. To improve TB disease prediction Asha et al [4] executed Association Rule Mining (ARM) concepts on TB data. Roopa H et al [5] segmented chest x-ray image then applied city block distance measure to analyze the presence of TB. Li –Yeh Chuang et al [6] considered DNA

microarray data applied Taguchi genetic feature selection algorithm and used KNN with Leave One Out Cross Validation (LOOCV) to understand its final outcome. Features based on shape and texture were extracted from breast tumor data by M.Suganthi et al [7] using Multi objective Genetic Algorithm (MOGA).Statistical association rules were applied by Pedro et al [8] on the extracted texture and shaped based features of MRI data.

For brain MRS data Mahmoodabadi et al [9] used PCA to extract features then features were discriminated by Simple Genetic Algorithm (SGA). Zyout et al [10] considered mammogram images, from these texture features were extracted then used and Particle Swarm Optimization (PSO) to discriminate features for further analysis. From microarray data wavelet features were extracted by Liu Yihui et al [11] and then applied Support Vector Machine (SVM) for classification. Ranking of features by concave approximation of zero-norm function was given by Luca Bravi et al [12] and Support Vector Machines (SVMs) with Gaussian kernel was used for performance analysis.

Max- Relevance-Max-Distance (MRMD) was used as a feature extraction method by Quan Zou et al [13] which provides a means to rank the features resulting in improved classification efficiency of the data. In an incremental manner Alaleh Razmjoo et al [14] ranked features which can be used in classification methods.

3. PROPOSED METHODOLOGY

Image contains information which cannot be analyzed by normal person so it has to be explained by an expert in that field for knowledge extraction. Chest x-ray image contains information like infiltrates, opacity, pleural effusion, etc. which a physician can analyze and conclude that a person is infected by TB or not. To analyze TB using chest x-ray, feature extraction and selection has to done. Then statistical and numerical model are applied on these features for analysis. The method proposed for TB data analysis is depicted in Fig.1.

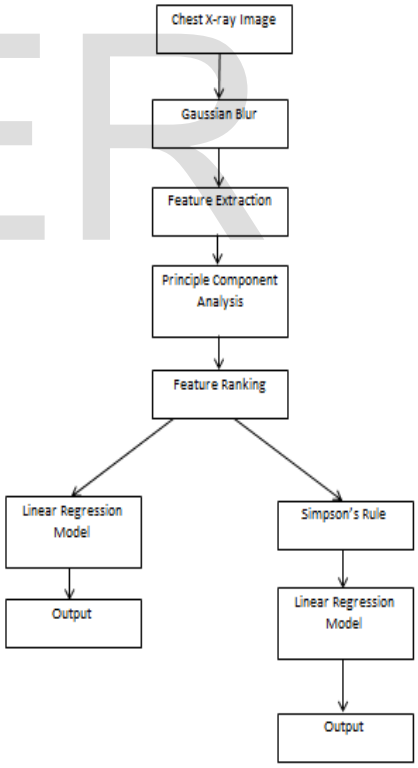

Fig.1. x-ray image analysis model.

Steps involved in x-ray image analysis model are

1. Input data:

   Images of x-ray are taken as an input for the model. We have collected around 77 images from a city hospital, out of this 47 are TB infected images and 30 are normal images.

2. Preprocessing:

   Images of x-ray may contain redundant, noise and irrelevant data which is removed by using method called Gaussian blur.

3. Feature Extraction:

   The quality of a x-ray image can be understood by geometric and texture based features [15]. Geometric features [16] like perimeter, major axis, area, orientation, minor axis, eccentricity, etc., are got from x-ray image using function called regionprops().

   Texture features [16] like entropy, homogeneity, correlation, contrast, energy, etc., are found from a x-ray image using entropy(), graycrops(), numel(U) where U represents uniformity, etc., respectively. Seventeen attributes were extracted from x-ray image .The extracted feature values are projected to a new space using Principal Component Analysis (PCA) where the variance of first principle component is highest when compared to others which are described in Table.1.

Table.1. Extracted features of a x-ray image.

No.	Attribute	NewAttribute	Value
1.	Area	Comp1	Integer
2.	Perimeter	Comp2	Numeric
3.	Major Axis	Comp3	Numeric
4.	Minor Axis	Comp4	Numeric
5.	Orientation	Comp5	Numeric
6.	Uniformity	Comp6	Integer
7.	Gray level	Comp7	Numeric
8.	Skewness	Comp8	Numeric
9.	Energy	Comp9	Numeric
10.	Signal to	Comp10	Numeric
11.	Homogeneity	Comp11	Numeric
12.	Eccentricity	Comp12	Numeric
13.	Variance	Comp13	Numeric
14.	Standard	Comp14	Numeric
15.	Entropy	Comp15	Numeric
16.	Contrast	Comp16	Numeric
17.	Correlation	Comp17	Numeric
18.	Class	Class	Integer (0

4. Ranking of features:

   The importance of attributes can be analyzed and understood by certain defined measure [17]. The weights of attribute of TB data are obtained by applying weight by PCA method which gives the ranking of attributes with respect to class variable.

5. Multiple variable Linear Regression Model:

   The relationship between dependent attribute 'y' and set of multiple independent attribute values x_1, x_2, x_3,....., x_k is given by the equation 1 .

$$y = q_0 + q_1 x_1 + q_2 x_2 + \ldots + q_k x_k + \varepsilon \quad (1)$$

where 'y' is the class variable of TB data,

'q_0' is an intercept,

ε is an error term and $q_1, q_2, q_3, \ldots\ldots\ldots, q_k$ are coefficients of $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_k$ respectively. Here dependent class variable is predicted using equation (1), based on values of $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_{17}$ which are independent set of attribute values of TB data given by comp1, comp2, .. , comp17 respectively.

6. Simpson's Rule:

The steps involved in finding the area of all attributes of TB data are given by

1) All the attributes are arranged according to their ranking and the model information given by multiple LRM.

2) Read this file containing the values of TB data.

3) Let n=number of attributes of TB data which is equal to 17.

4) Let $I = \frac{\text{value of preceding value attribute} - \text{value of previous column attribute}}{\text{total number of attributes}}$

5) Sum=values of first column + values of 17th column.

```
for (i in 2 : (n-1))
{
  if (i%%2 ==0)
        {
          Sum= Sum+2*(valuesof ith column)
        }
    else
        {
          Sum=Sum+4*(values of ith column)
        }
```

6) I/3 * Sum gives the area of all attributes.

7) Once area is found apply multiple LRM to analyze the result.

7. Output :

The model will classify the features of TB data as TB affected or normal.

4. EXPERIMENTAL RESULTS

Images must be preprocessed to remove irrelevant information and then features are extracted from it. Consider normal or TB image, and then extract the features from these and save it to a file. TB image feature extractions are shown in Fig.2, Fig. 3 and Fig .4 respectively.

Fig.2. TB Image

Fig.3. Marked Image

Fig.4. Masked Image

The infected parts of TB patient chest x-ray are identified by considering the suggestion of physician .Then this region is marked, masked and values are extracted which are shown in Fig. 3 and Fig. 4 respectively. The extracted values of Fig. 2 image are given in Table.2.

Table.2. Infected TB image feature values.

Area	9121
Perimeter	418.8620
Major Axis	138.5736
Minor Axis	94.9123
Orientation	87.0484
Uniformity	128
Gray level	115.3345
Skewness	0.1846
Energy	0.4684
Signal to Noise	6.2017
Homogeneity	0.9947
Eccentricity	0.7286
Variance	117.7717
Standard	10.8523
Entropy	1.3867
Contrast	0.1810
Correlation	0.9923

Attributes of TB data are projected to a new space using PCA and modeled using multiple LRM as shown in Fig. 5.

The residuals error between the prediction of LRM and actual results is smaller ranging between -0.29183 and 0.40370.And the median value is closer to zero. F-statistics shows that model has at least one variable that is significantly different than zero. The variables comp1, comp2, comp3, comp4, com5, comp7 and comp15 are more significant indicated by *** which is greater than ** and * when compared to comp8, comp12 and comp16 respectively. Here comp6, comp9, comp10, comp11, comp13, comp14 and comp17 of TB data are least important. The components of TB data are partitioned where 80% are considered for training and 20% of data for testing respectively. These data are classified using multiple LRM, giving an accuracy of 96.07% [16].

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.29183 -0.03385 -0.00112  0.03241  0.40370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.234e-01  1.201e-02  51.915  < 2e-16 ***
Comp.1      -1.298e-05  3.957e-07 -32.806  < 2e-16 ***
Comp.2       8.692e-04  8.754e-05   9.929 3.36e-14 ***
Comp.3       7.818e-04  1.804e-04   4.333 5.80e-05 ***
Comp.4      -4.816e-03  3.734e-04 -12.898  < 2e-16 ***
Comp.5      -4.257e-03  4.002e-04 -10.637 2.42e-15 ***
Comp.6      -8.613e-04  1.454e-03  -0.592  0.55593
Comp.7       1.128e-02  1.948e-03   5.789 2.89e-07 ***
Comp.8       1.664e-02  8.030e-03   2.072  0.04263 *
Comp.9      -1.672e-02  1.349e-02  -1.240  0.21983
Comp.10      9.735e-02  1.412e-01   0.689  0.49324
Comp.11      3.274e-01  3.051e-01   1.073  0.28759
Comp.12      4.583e+00  1.443e+00   3.176  0.00237 **
Comp.13      2.316e+00  2.049e+00   1.130  0.26306
Comp.14      2.999e-01  2.734e+00   0.110  0.91301
Comp.15     -2.617e+01  4.238e+00  -6.174 6.63e-08 ***
Comp.16      6.565e+01  3.173e+01   2.069  0.04292 *
Comp.17      1.126e+02  8.688e+01   1.296  0.20007
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1054 on 59 degrees of freedom
Multiple R-squared:  0.9638,    Adjusted R-squared:  0.9533
F-statistic: 92.31 on 17 and 59 DF,  p-value: < 2.2e-16
```

Fig.5. Result of model on TB data.

Analyzing the results of Fig.5 and attribute weights of TB data, rearrange the attributes of TB data by considering importance of attributes and its ranking which is used as an input for Simpson's rule. The areas of all the attributes are found and modeled using multiple LRM to obtain the results as shown in Fig .6.

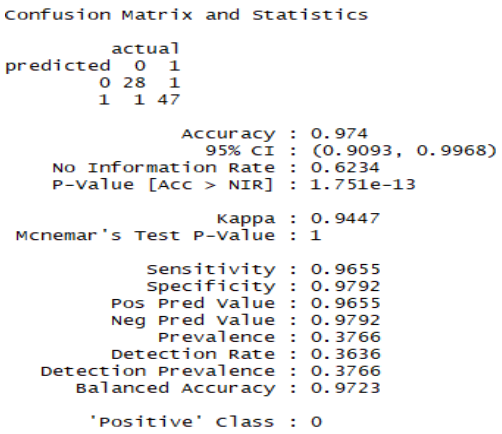

Fig.6. Performance of TB data using Simpson's rule.

The TB data performance using Simpson's rule is represented by confusion matrix as depicted in Figure 6. Here 28 observations are correctly predicted negative for TB disease and 47 observations are correctly predicted positive for TB. So, 75 observations have been correctly classified for TB data using the model. 1 observation represent incorrect prediction of model that resulted positive for normal person and 1 observation represent incorrect prediction of model that resulted negative for TB affected person. So, 2 observations have been wrongly classified by model for chest x-ray data.

Table.3. Result and comparison of TB data.

Source	Method	Accuracy
Roopa H et al [15]	LRM	96.07%
Our Research Study	LRM + Simpson's Rule	97.4%

In Table.3 diagnosis of TB data is improved using our proposed numerical and statistical method than existing method [15]. The TB data generated and analyzed for the current research work is not publicly available. It is obtained by our request from a hospital in our district.

5. CONCLUSION

The proposed research work on TB data gives an insight of extraction of feature and weights of features of x-ray image. The features are then modeled using multiple linear regression model to classify whether a patient is TB infected or not. Then for the same features, areas of all features are found using Simpson's rule and then modeled to classify TB data. It is observed that the classification performance improved when numerical and statistical methods are applied together on TB data.

REFERENCES

[1] Asha, T., K. N. B. Murthy, and S. Natarajan. "Data mining techniques in the diagnosis of tuberculosis." *INTECH Open Access Publisher*,(2012).

[2] GlobalTuberculosisReport2015.

http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf

[3] Perner, Petra. "Image mining: issues, framework, a generic tool and its application to medical-image diagnosis." *Engineering Applications of Artificial Intelligence* 15.2 (2002): 205-216.

[4] Asha. T , Dr. S. Natarajan , Dr. K.N.B.Murthy "A Study of Associative Classifiers with Different Rule

Evaluation Measures for Tuberculosis Prediction." *IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" AIT*, (2011).

[5] Roopa H and Asha T "Segmentation of X-Ray Image using City Block Distance Measure." *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kumaracoil, (2016), pp. 186-189.

[6] Chuang, Li-Yeh, et al. "A hybrid feature selection method for DNA microarray data." *Computers in biology and medicine* 41.4 (2011): 228-237.

[7] Suganthi, M., and M. Madheswaran. "Mammogram tumor classification using multimodal features and Genetic Algorithm." Control, Automation, *Communication and Energy Conservation, 2009. INCACEC 2009. International Conference on. IEEE*, 2009.

[8] Bugatti, Pedro Henrique, et al. "Content-based retrieval of medical images by continuous feature selection." *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on. IEEE*, 2008

[9] Mahmoodabadi, S. Zarei, et al. "PCA-SGA implementation in classification and disease specific feature extraction of the brain MRS signals." *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE* 2008.

[10] 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.

[11] Liu, Yihui. "Wavelet feature extraction for high-dimensional microarray data." *Neurocomputing* 72.4 (2009): 985-990.

[12] Bravi L, Piccialli V, Sciandrone M. "An optimization-based method for feature ranking in nonlinear regression problems." *IEEE transactions on neural networks and learning systems*. (2017)Apr;28(4):1005-10.

[13] Zou Q, Zeng J, Cao L, Ji R."A novel features ranking metric with application to scalable visual and bioinformatics data classification." *Neurocomputing. (2016)* Jan 15;173:346-54.

[14] Razmjoo A, Xanthopoulos P, Zheng QP. "Online feature importance ranking based on sensitivity analysis." *Expert Systems with Applications.* (2017) Nov 1;85:397-406.

[15] Roopa H, Asha T." Feature Extraction of Chest X-ray Images and Analysis Using PCA and kPCA". *International Journal of Electrical and Computer Engineering (IJECE).* 2018 Oct 1;8(5).

[16] Haralick, Robert M. "Statistical and structural approaches to texture." *Proceedings of the IEEE* 67.5 (1979): 786-804.

[17] Roopa, H., and T. Asha. "Analysis of Feature Ranking Methods on X-Ray Images." *International Conference on ISMAC in Computational Vision and Bio-Engineering*. Springer, Cham, 2018.