International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

# A Data Mining Approach to Detect Tuberculosis Using Clustering and GA-NN Techniques

**Shakshi Garg[1], Navpreet Rupal[2]**

[1]Shaheed Udham Singh College of Engineering & Technology,(Tangori)

[2]Guide, Shaheed Udham Singh College of Engineering & Technology,(Tangori)

**Abstract:** ***During the past years, this has turn out to be clear that the range of Human Immunodeficiency Virus infection as well as persons immigrate from zones of high rate have ensued in bigger amount of Tuberculosis events. TB can affect all types of organs in a living being body. In previous years, TB classification has been done using various algorithms like color segmentation, thresholding, histogram equalization. The main objective of this research work is to create a data mining way out that makes identification of TB as exact as possible. In our proposed framework we have used various techniques such as centroid selection based clustering algorithm would be used to enhance the clustering scheme, PCA for feature extraction, genetic algorithm for feature optimization and neural network for training and testing purpose. In the end, results are being evaluated after classification and testing on the basis of performance parameter such as accuracy, recall, precision, false acceptance ratio, and false rejection ratio.***

**Keywords:** Tuberculosis, Data Mining, Principal Component Analysis, Neural network, Genetic Algorithm

## 1. Introduction

Tuberculosis (TB) was assumed to be practically in control; however it has once again turn out to be a severe issue world-wide. This is instigated through a bacterium that is entitled as mycobacterium TB [1]. This specific disease could easily spread between normal human beings in addition to the ill persons that suffer from TB might pass away except if any condition they get the proper medical treatment to recover. This specific microscopic organism extensively be present on birds, cattle, humans, and sheep. All type of organs present in the body could be disturbed by tuberculosis bacteria. Nevertheless, most of the TB cases occur in lungs area [2, 3].

With the intention of curing TB, four to five distinctive major anti-tuberculosis antibiotics are utilized for around six to twelve months. Some cases might heal deprived of any cure strategy if in any condition immune system is resilient enough. Afterwards full recovery, lung injuries that are instigated through TB disease still occur as calcify nerve. Unfortunately, in several cases that are not cured it may possibly result by the patient's demise [3]. The main objective of this research work is to make a solution of data mining that makes analysis of TB disease as precise as conceivable and it benefits determining if in any condition, it is reasonable to start TB cure on assumed patients deprived of any waiting for the precise results of test or not [4, 5].

Several researchers have now taken interest in several classification procedure for discovering the much reduced rate of error and enhanced rate of predication. Classification is a procedure to allocate an object into pre-defined groups by appraising their association into class according to some attribute values intended for that particular objects.

## 2. Data Mining in Medical Field

Data mining/knowledge-discovery in databases is a method of releasing conceivable data from raw information [6]. A software engine can filter a lot of information and naturally report fascinating examples without obliging human mediation. Other knowledge discoveries are Statistical Analysis, OLAP, Data Visualization, and Ad hoc inquiries. Dissimilar to these advances, data mining does not oblige a human to ask particular inquiries [7].

Data mining process can be extremely useful for Medical practitioners for extracting hidden medical knowledge. It would be impossible for traditional pattern matching and mapping strategies to be effective and precise in prognosis or diagnosis without application of data mining techniques[8]. Knowledge discovery and data mining have discovered in many business and scientific domain applications. Data mining has been connected with accomplishment in distinctive fields of human attempt, including advertising, managing an account, client relationship administration, designing and different regions of science. Be that as it may, its application to the investigation of restorative information has been generally constrained.

## 3. Literature Review

A.Sudha, et.al, paper provides a brief survey on forecasting the occurrence of life-threatening diseases that might causes to demise as well as they also list out the several classification procedures which has been utilized with several number of distinctive attributes for forecast.

Bakar et.al, they applied Rough NNs for categorizing the types of TB. Data set has two hundred thirty three records that has fourteen different attributes, initially diminished due to pre-processing of data_set. The decisive data_set is having eight distinctive attributes that are age, cough for more than three weeks, blood phlegm, fevers, gender, night sweats, sputum test and weight. 70 percent of the data set is utilized for training purpose and left thirty percent is utilized for testing purpose. Discretization is implemented on the numeric as well as continuous attributes utilizing Rough_Set

application. Afterwards, NN is implemented for training the dataset.

Collins K. Ahorlu et.al, here author define some factors which are affecting low TB incident recognition in the Ghana's Upper West Area. This was a vivid research work where partially-structured questionnaire form was directed to group of sixty-one respondents; with 6 focus group debates along with twenty detailed interviews were directed to produce both quantitative as well as qualitative information/ statistics for analysis purpose.

K.R Lakshmi et.al, in this paper author summarizes several technical articles along with numerous review on TB diagnosis as well as prognosis also they mainly focus on present research which is being done by utilizing several data mining methods so that to augment the TB prognosis in addition to diagnosis much easier. At this point, they took benefit of those existing high-tech developments to advance the best forecast system for TB survivability.

Orhan Er et.al, they present a study on tuberculosis diagnosis, carried out with the help of multilayer NNs. In their proposed work they have used a multi-layer neural network with 2 different hidden layers along with a genetic algorithm which is used for training algorithms.

P. Seppo et.al, they researched about two categories of conflicts, out of two one of them is generated by data discrepancy inside the specific area of the intersection of the data_bases and the another conflict is generated after the meta technique chooses dissimilar data mining approaches with unreliable competence maps for the specific objects of the particular intersected portion and their amalgamations and it also propose various approaches to control them.

Rusdah et.al, Scientists keep evolving several exact data mining approaches for quick TB diagnosis to decrease the rate of development of the global population of TB patients. This paper aims to offer state-of-the-art of data mining approaches in identifying TB utilizing clinical signs as I/P parameters. Initially, it familiarizes TB and existing methods utilized for TB diagnosis. Then it talk over methods for pre-processing data and data mining approaches for TB diagnosis which are used currently.

## 4. Proposed Work

The proposed methodology will follow following steps:

**Step 1 :** Upload data set.
**Step 2 :** Apply k-mean clustering for creating two cluster of data set.
**Step 3 :** After this apply Principal Component Analysis for feature extraction from cluster1 and cluster2 separately.
**Step 4 :** Then, apply genetic algorithm for optimizing features extracted from cluster1 and cluster2 separately.
**Step 5 :** Apply neural network on cluster1 and cluster2 for classification along with testing them for the results separately.
**Step 6 :** Then evaluate the results using parameters such as recall, precision, False Acceptance Rate, False Rejection Rate and Accuracy.
**Step 7 :** STOP.

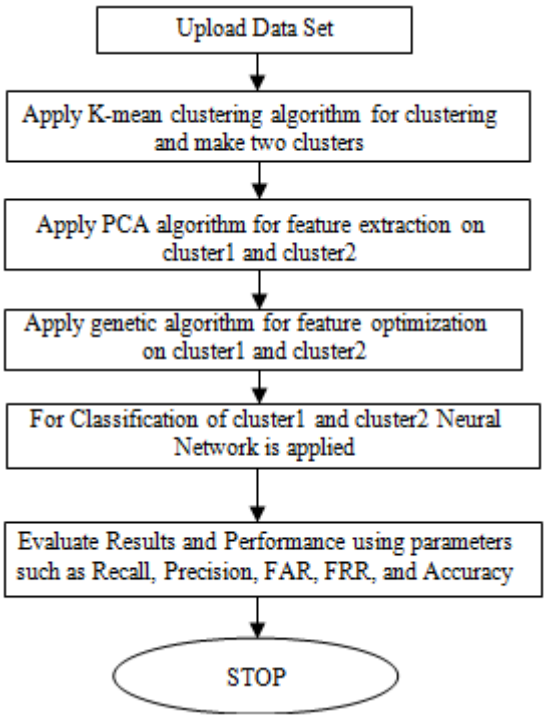



**Figure 2:** Flowchart of Proposed Work

## 5. Results & Snapshots

The whole simulation has been done in MATLAB 2010 a and below figure shows the graphs achieved for TB classification.

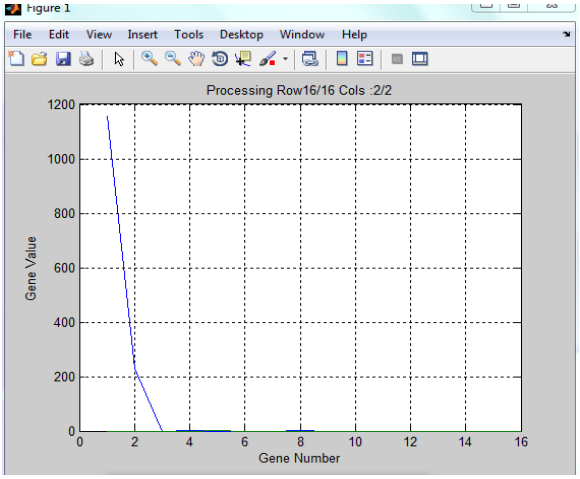



**Figure 3:** Apply genetic algorithm for cluster1

After PCA implementation on cluster 1. We will implement genetic algorithm on cluster-1. In this process, we will reduce as well as optimize the feature sets extracted using PCA. A graph will generate after the complete implementation of genetic algorithm as shown above which is plotted between gene number and gene value by processing 16/16 rows and 2/2 columns. For instance, at gene number 2, the gene value is 220.

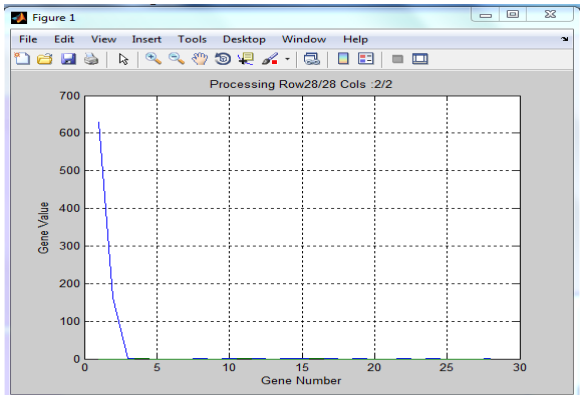
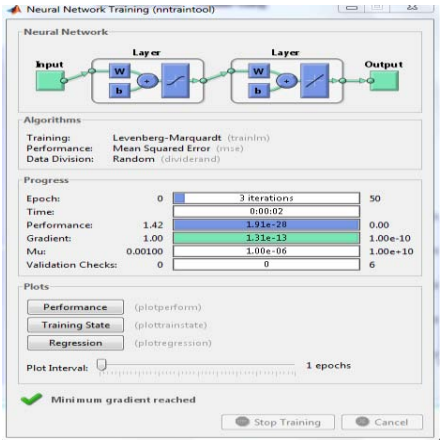



**Figure 4:** Apply genetic algorithm for cluster1

After GA implementation on cluster 1. We will implement genetic algorithm on cluster-2 and repeat the same process as cluster-1. In this process, we will reduce as well as optimize the feature sets extracted using Principal Component Analysis. A graph will generate after the complete implementation of genetic algorithm on cluster-2 as shown above which is plotted between gene number and gene value by processing 28/28 rows and 2/2 columns. For instance, at gene number 2, the gene value is 150.




**Figur 5:** Apply neural network training on cluster-1

In this figure, after applying genetic algorithm for feature reduction along with feature optimization we will apply neural network is applied on cluster-1 for training and testing. In this first we will apply neural network training on cluster-1 feature optimized set and then test it by classify on the basis of feature that they have tuberculosis or not. Here, epoch value is 3 iterations, time taken is 02 second, performance value is 1.91e-28 and gradient value is 1.31e-13. It shows that minimum gradient reached. After testing results will be generated from cluster-1 and cluster-2. Here, we will evaluate results by using parameters such as recall, precision, False Acceptance Rate, False Rejection Rate and accuracy. As, shown in below table values. Repeat same process of training and testing on cluster-2 and obtain results.

**Table 1:** Results of cluster 1 and cluster 2

| Parameters | Cluster 1 | Cluster 2 |
|---|---|---|
| Accuracy | 99.9455 | 99.7302 |
| FAR | 0.003182 | 0.002442 |
| FRR | 0.002273 | 0.002558 |
| Precision | 0.005833 | 0.009052 |
| Recall | 0.004167 | 0.000948 |


**Figure 6:** Graph of result parameters used for cluster 1 and cluster 2.

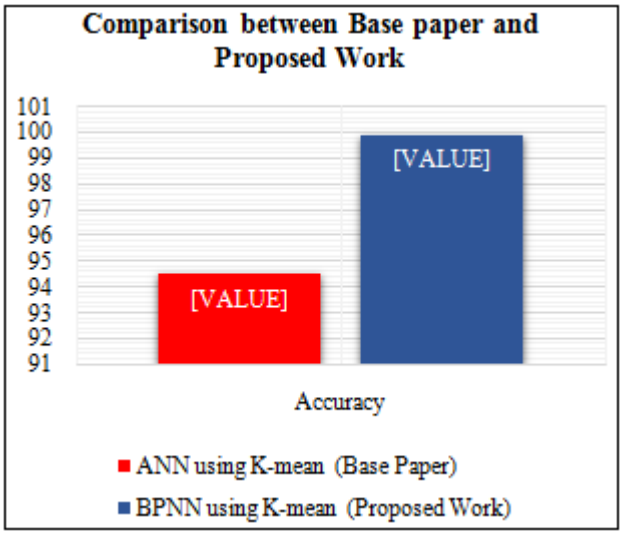



**Figure 7:** Comparison graph between Base Paper and Proposed Work

Above comparison graph is given between base paper and proposed work using accuracy parameter. It shows that our proposed work result is much better as compared to base paper accuracy which is 94.56 and much lesser than 99.94 proposed work result value.

## 6. Conclusion & Future Scope

Several new methods which go away from the standard passive case-discovering methods currently used are needed to make progress towards confirming all those people with Tuberculosis are acknowledged and given appropriate treatment. But this work has utilized PCA, K-Means clustering method and neural network for classification of TB classes. In this we have three panels named upload panel, gene information panel and data panel. In upload panel, we have several buttons of upload data set, apply k-mean, apply

PCA for cluster-1, apply PCA for cluster-2, apply GA for cluster1, apply GA for cluster2, apply neural network for cluster-1 and testing, and apply neural network for cluster-2 and testing. In gene information panel, it displays total number of genes, total number of features, processing gene number and evaluating feature number. In data panel, we will represent the uploaded data. We have evaluated the results on the basis of performance parameter such as Recall, Precision, False Acceptance Ratio, False Rejection Ratio and Accuracy. The whole simulation has been taken place in MATLAB environment. And obtained values are Recall value is 0.00094803., Precision value is 0.009052, Accuracy value is 99.7302., False Acceptance Rate value is 0.0024422 and False Rejection Rate value is 0.00025577. Its future scope lies in the use of ICA i.e. independent component analysis in place of Principal component analysis technique which might give better results by extracting features more precisely.

## References

[1] Tuncay, N., Uzunboylu, H. (2010). Trend of Distance Education in the last three Decades. World Journal On Educational Technology, 2(1). Retrieved November 15, 2010,
[2] Davidson S. Davidson's Principles and Practice of Medicine.Churchill Livingstone; 1999.
[3] Harrison TR. Harrison's Principles of Internal Medicine. McGraw-Hill Education; 1999.
[4] Stefan Jaeger, Alexandros Karargyris, Sameer Antani, and George Thoma, "Detecting Tuberculosis in Radiographs Using Combined Lung Masks" 34th Annual International Conference of the IEEE EMBS San Diego, California USA, 28 August - 1 September, 2012.
[5] Sánchez MA, Uremovich S, Acrogliano P. Mining Tuberculosis Data. In: Berka P, Rauch J, Zighed DA, editors. Data Mining and Medical Knowledge Management: Cases and Applications. New York: Medical Information Science Reference; 2009.
[6] J. Han and M. Kamber. Data mining: concepts and techniques: Morgan Kaufmann Pub, 2006.
[7] H. Dağ, K. E. Sayın, I. Yenidoğan, S. Albayrak, and C. Acar, "Comparison of Feature Selection Algorithms for Medical Data," in 2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2012, pp. 1–5.
[8] M. H. Dunham, "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc., (2003).
[9] A.Sudha, P.Gayathri and N.Jaisankar, "Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability" International Journal of Computer Applications (0975 – 8887) Volume 41– No.17, March 2012.
[10] Bakar AA, Febriyani F. Rough Neural Network Model for Tuberculosis Patient Categorization. In: Proceedings of the International Conference on Electrical Engineering and Informatics; 2007; Indonesia. p. 765-768.
[11] Collins K. Ahorlu, Frank Bonsu, "Factors affecting TB case detection and treatment in the Sissala East District, Ghana" Journal of Tuberculosis Research, 1, 29-36. doi: 10.4236/jtr.2013.13006.
[12] K.R.Lakshmi, M.Veera Krishna, S.Prem Kumar, "Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability" DOI: 10.5815/ijmecs.2013.08.02.
[13] Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu, "Tuberculosis disease diagnosis using Artificial Neural networks ", Journal of Medical Systems, Springer, 2008, DOI 10.1007/s10916-008-9241-x online.
[14] Puuronen, S.; Terziyan, V.; Logvinovsky, A., "Mining several databases with an ensemble of classifiers," in Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on, vol., no., pp.218-222, 1999 doi: 10.1109/DEXA.1999.795169.
[15] Rusdah and Edi Winarko, "Review on Data Mining Methods for Tuberculosis Diagnosis" Information Systems International Conference (ISICO), 2 – 4 December 2013.