

# A Review on Tuberculosis Using Data mining Approaches

Shakshi Garg<sup>1</sup>, Navpreet Rupal<sup>2</sup>  
Shaheed Udham Singh College of Engineering & Technology, Tangori

**Abstract** - Tuberculosis is one of the prominent diseases for all persons in established countries as well as India. It is the utmost communal reason of demise in human being. The high occurrence of Tuberculosis has augmented expressively in previous years. In previous years TB classification has been done using various algorithms like color segmentation, thresholding, histogram equalization. The previous work is appreciable, but the work is consistent for similar types of dataset. In this paper we have discussed about tuberculosis, various data mining approaches for tuberculosis data set. And also give a brief about previous work done in this field as well as two techniques that are genetic algorithm and neural network can be utilized for future work to enhance the results of previous approaches to diagnose and prevent TB.

**Keywords** - Tuberculosis, Data mining, Neural Network, Genetic Algorithm.

## 1. INTRODUCTION

The rise of health care cost is one of the world's most important problems. Due to increase in world's population, the health care industries are facing many challenges in addition to issues dependent upon patient's relentlessness is needed to be lessened then identify it former in a more effective way. The medical infirmaries are trying to progress in addition to upsurge the proficiency of medical resource use in order to reduce the health care cost. Tuberculosis (TB) is a very common disease in India which is caused by mycobacterium and established as severe disease with fatal effects. According to WHO (World Health Organization) statistics available for 2013 provides an assessed incidence amount of 2.1 million instances of tuberculosis occurred in India beyond a worldwide occurrence of 9 million instances. The expected tuberculosis occurrence amount for 2013 is specified as 2.6 million instances. It can spread from one person to other person by using air as medium, when they come in contact. The microorganism can go inside the body by simple inhalation process through lungs and can affect lungs first and after that other parts of the body by mixing with blood. Sometimes it is very difficult to diagnose it at initial level by physicians because symptoms of TB initial level are similar to any other normal disease.

Prediction and diagnosis of tuberculosis at initial phase is the immense problem in India and other Asian countries also. It has highest mortality rate in comparison to other diseases which are caused by single microorganism. Many researchers have already taken interest in classification algorithm for finding the better predication rate and reduced error rate. Classification is a process to assign an object into predefined classes by evaluating their membership into class according to attribute values for that objects. For making significant improvement in classification process many researchers have employed different approaches and adopt different learning methods which are better than previous traditional approaches but very few researchers have tried for classification process preceded by clustering approach.

S. No.	Status of Tuberculosis In India
1.	Out of five one is the patient of TB in the world is an INDIAN.
2.	Around 2.6 million TB patients present in INDIA.
3.	Around 66,000 annual new cases occur of multi-drug resistant TB in INDIA.
4.	Sometimes there is shortage of DR-TB medicines due to fewer quality suppliers, afford ability and stock-outs.

## 2. DATA MINING APPROACHES FOR HEALTHCARE

Data Mining is one of the most dynamic as well as motivating part of exploration using objective of finding meaningful information from huge data sets. In current years, Data Mining is becoming widespread in healthcare area for the reason that there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud happened in availability of medical solution to the patients at lower cost, health insurance, discovery of causes of diseases and identification of medical treatment methods. The aforementioned also aids healthcare investigators intended for constructing competent healthcare strategies, making drug approval systems, developing health profiles of individuals etc. [1]. The data generated by the health organizations is very vast and complex due to that this one is problematic to evaluate the information with the intention of taking significant decision concerning about just patient health. This specific information compasses particulars about patients, treatment cost, hospitals, medical claims, and so on. Consequently, there is a requirement to produce a prevailing tool designed for examining in addition to removing significant information from this complex data. The investigation of health information progresses the healthcare through augmenting the working of patient management tasks. The outcome of Data Mining technologies are to provide benefits to healthcare organization for assembling patients having alike category of diseases or health problems to facilitate healthcare association to provides them effective treatments. This one could probably be valuable for forecasting the duration of stay of patients in hospital,

for medical diagnosis and making plan for effective information system management. Current technologies are utilized in medical arena to augment some medical facilities in price effective manner. Data Mining techniques are also used to analyze the various factors that are responsible for diseases for instance different working environment, sort of food, living conditions, education level, obtain ability of pure water, health care services, cultural ,environmental and agricultural factors.

### 3. DATA MINING USING TUBERCULOSIS DATASET

Tuberculosis is a typical and frequently savage irresistible disease brought about by mycobacterium; in people it is primarily Mycobacterium tuberculosis. It is an incredible issue for most creating nations in light of the low determination and treatment opportunities. Tuberculosis has the most noteworthy mortality level among the maladies brought about by a solitary sort of microorganism. Along these lines, tuberculosis is an awesome well being concern everywhere throughout the world and in India also.

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. Therefore, there is a developing weight for clever information examination, for example, data mining to encourage the extraction of learning to backing clinical pros in deciding. Restorative datasets have come to gigantic limits. This information may contain profitable data that anticipates extraction. The learning may be embodied in different examples and regularities that may be covered up in the information. Such information may turn out to be inestimable in future therapeutic choice making. Information investigation underlies numerous figuring applications, either in an outline stage or as a component of their on-line operations. Information investigation systems can be dichotomized as either exploratory or affirming, in view of the accessibility of proper models for the information source, yet a key component in both sorts of methods (whether for speculation development or choice making) is the gathering, or arrangement of estimations in light of either integrity of-fit to a hypothesized model, or common groupings (grouping) uncovered through examination.

Clustering is the unsupervised grouping of examples (perceptions, information things, or highlight vectors) into gatherings (bunches). The clustering issue has been tended to in numerous connections and via analysts in numerous orders; this mirrors its expansive offer and helpfulness as one of the progressions in exploratory information investigation. Then again, clustering is a troublesome issue combinatory, and contrasts in presumptions and settings in diverse groups have made the exchange of helpful generic ideas and approaches moderate to happen [1].

Data classification procedure utilizing learning acquired from known authentic information has been a standout amongst the most seriously considered subjects in measurements, choice science and software engineering. Information mining strategies have been connected to medicinal administrations in a few zones, including expectation of viability of surgical methods, restorative tests, drug, and the revelation of connections among clinical and conclusion information.

### 4. GENETIC ALGORITHM

Genetic algorithms are stimulated by Darwin's theory related to evolution. Explanation to issues resolved as a result of using genetic algorithms is progressed. This procedure is taking place using a group of explanations (epitomized thru chromosomes) entitled as population. Answers obtained after one population are occupied as well as utilized to arrange a novel population. This is encouraged through an anticipation, in which the new-fangled population would be enhanced as compared to previous one. Solutions that are designated to construct fresh offspring are chosen as stated through their fitness operator- more appropriate they are the more probabilities they have to regenerate. This is continual in anticipation of few specific circumstance is mollified.

### 5. NEURAL NETWORK

Neural networks are normally systematized in layers. Layers are prepared up of a amount of interrelated 'nodes' that usually encompass an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates in the direction of one or more 'hidden layers' wherever the actual handling is completed through a system of weighted 'connections'. The hidden layers at that time linkage to an 'output layer' wherever the solution is output as shown in the figure below.

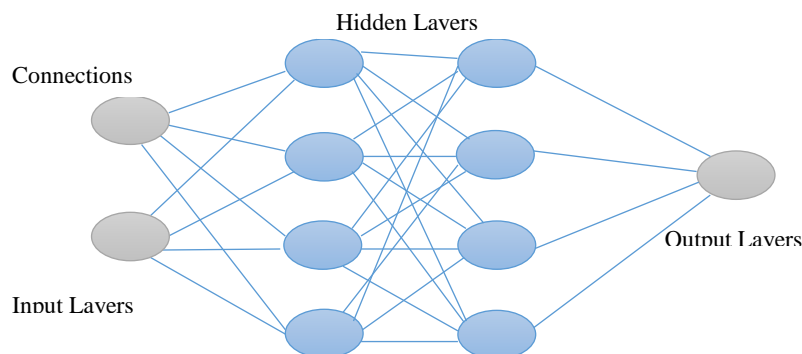


Figure 1 Schematic representation of neural network

#### *Advantages of Neural Network*

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.

2. Self-Organization: An ANN can create its own organisation or representation of the information it receives during learning time.
3. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.
4. Fault Tolerance via Redundant Information Coding: Neural Network is able to deal with noisy data.

## 6. LITERATURE SURVEY

S.No.	Authors	Work Done
1.	Manish Shukla and Sonali Agarwal	They presented a new approach for centroid selection in k-mean algorithm for health datasets which gives better clustering results in comparison to traditional k-mean algorithm.
2.	K.R Lakshmi, M.Veera Krishna and S.Prem Kumar	This study paper summarizes various review and technical articles on Tuberculosis diagnosis and prognosis. They emphasis on present research which are performed utilizing the data mining methods in the direction of enhancing the Tuberculosis diagnosis and prognosis. Here, they took advantage of those available technological advancements to develop the best prediction model for Tuberculosis survivability.
3.	Seppo Puuronen, Vagan Terziyan and Alexander Logvinovsky	They analyze the two types of conflicts, one created by data inconsistency within the area of the intersection of the data bases and the second is created when the meta technique chooses dissimilar information mining approaches using varying competence maps intended for the substances of the intersected part and their combinations and suggest ways to handle them.
4.	Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu	They present a study on tuberculosis diagnosis, carried out with the help of multilayer neural networks (MLNNs). Designed for this determination, an MLNN having two separate hidden layers as well as a genetic procedured esigned for training algorithm has been used.
5.	Wai Yan Nyein Naing, Zaw Z. Htike	In this, medical contextual past of TB release in upper body X-rays and a survey of the various approaches in TB detection and classification are presented.
6.	Collins K. Ahorlu, Frank Bonsu	To define elements affecting low tuberculosis case recognition. This was a descriptive study where semi-structured questionnaire was administered to 61 respondents; six focus assembly conferences as well as 20 in-depth interviews were directed to produce together qualitative as well as quantitative data for analysis. There is therefore the need for vigorous health education to inform the people about biomedical reasons of TB along with obtain ability of suitable cure meant for disease at health facilities.
7.	Tamer and Adem Karahoca	They proposed the use of Sugeno-type “adaptive-network-based fuzzy inference system” (ANFIS) to predict the existence of mycobacterium tuberculosis. 667 diverse patient archives that are acquired through a clinic are utilized in entire process of this research. Every single of the patient archives contain of 30 distinct input constraints. ANFIS model is produced through utilizing 500 of these records archives. They also implemented a multilayer perceptions and PART model using the same data set.

## 7. CONCLUSION

Tuberculosis is a typical and frequently savage irresistible disease brought about by mycobacterium; in people it is primarily Mycobacterium tuberculosis. It is an incredible issue for most creating nations in light of the low determination and treatment opportunities. There is urgent need to prevent TB in humans. So, in this paper we have discussed about various existing techniques to prevent tuberculosis like neural network, SVM in addition to data mining method.

## 8. REFERENCES

- [1] Manish Shukla and Sonali Agarwal, “Hybrid approach for tuberculosis data classification using optimal centroid selection based clustering” DOI: 10.1109/SCES.2014.6880115 Conference: 2014 Students Conference on Engineering and Systems (SCES).
- [2] 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
- [3] Seppo Puuronen, Vagan Terziyan and Alexander Logvinovsky, “Mining Several Data Bases with an Ensemble of Classifiers”.
- [4] 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.
- [5] Wai Yan Nyein Naing, Zaw Z. Htike, IIUM, Malaysia, “Advances in Automatic Tuberculosis Detection in Chest X-Ray Images ”December 2014 volume 5, number 6, SIPIJ.
- [6] 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](https://doi.org/10.4236/jtr.2013.13006).
- [7] Tamer, “Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches”, World Conference on Information Technology, Vol.3, pp.1404-1411, 2011.
- [8] Asha.T, S. Natarajan and K.N.B. Murthy, “Diagnosis of Tuberculosis using Ensemble methods”, IEEE, 2010, 978-1-4244-5539-3/10.
- [9] K.S. Al-Sultan, “A Tabu Search Approach to the Clustering Problem,” Pattern Recognition, vol.28, no.9, pp. 1443-1451, 1995.

- [10] R.V. Singh, and M. P S “Bhatia, "Data clustering with modified Kmeans algorithm," International Conference on Recent Trends in Information Technology (ICRTIT), 2011, pp.717-721.
- [11] J. Han, and M. Kamber, Data mining: concepts and techniques. 2nd ed.The Morgan sKaufmann Series, 2006.

