Tuberculosis Disease Forecasting Among Indian Patients

Rupali Zakhmi
Department of Computer Science and Engineering Desh Bhagat University, Fatehgarh Sahib, INDIA
Email-ID:rupali2218@gmail.com

***Abstract*----**Tuberculosis is a conspicuous syndrome for all individuals in developing countries including India. It is an uttermost causation of bereavement in personage. It is an ailment triggered by bacteria which strikes hominid body parts, primarily lungs. The desideratum of this paper is to foretell tuberculosis disease using data mining techniques, which tends to make a medical diagnosis of tuberculosis rigorous. Data Mining Techniques will help to glean that whether it is plausible to start tuberculosis treatment on suspected victims or not, without waiting for pernickety medical test outcomes. This scrutiny emphasis on patients health and provides treatment at low outlay through forecasting systems. There are assorted parameters such as Cough, Chest Pain, Night Sweats, Age, Weight Loss, Gender and Fever, Coughing up Blood, No Appetite which are used for predicting tuberculosis. Both Genetic algorithm and Neural network backwash better than other techniques. Tuberculosis disease forecasting is accomplished by soft computing technique. Genetic algorithm offers best fitness value, disembroil optimization problems whereas Neural Network takes parameters as input and also utilize genetic operators to train the neural network and spawn an output for presaging tuberculosis disease. This research outlines the main review and technical papers on tuberculosis detection that are implemented using multifarious data mining techniques. Review of papers surmises that soft computing technique acquires the highest accuracy.

Keywords*----Data mining, Genetic Algorithm, Neural Network, Parameters, Tuberculosis disease, Soft Computing.*

*****

I. INTRODUCTION

Data Mining is a technique of fascinating non-trivial, implicative, formerly unidentified and potentially valuable information or patterns from the giant quantity of data. It has been using in medical, known as medical data mining. The type of data mining contains a vast amount of medicinal data which includes patient's records. Doctor uses medical records to give right treatment to the patients. Two preeminent goals of data mining are prediction and other is description. Prediction incorporates attributes which prognosticate future value of some another attribute whereas description emphasis on patterns that delineates data which can be conveniently interpreted by an ethnologic [12]. The prediction and analysis of TB are one of the challenging tasks for researchers. Data mining techniques play a momentous role in the prediction of tuberculosis. Tuberculosis is plague which usually attacks feeble immune system. It is the most familiar cause of demise in a human being.

Prophecy of tuberculosis using legion techniques at the condign time is one of the challenging tasks. Tuberculosis disease in patients is very high because when an individual is not vigorous then they cannot combat the disease. There are manifold advantages of using data mining techniques for the treatment in advance, without waiting for an appropriate upshot. There have also been various obstacles associated with their use. When treatments of tuberculosis are employed, they do not always find definite results at a low price. They are transportable in the atmosphere and often proliferate through talk, sneeze, and cough by the infected person. The problem with tuberculosis disease is that when the bacteria of tuberculosis come in contact with other persons, thus cause harm and get infected. A way of predicting tuberculosis, treatment is required via collecting pre-information regarding various parameters such as fever, cough, chest pain, night sweats, age, gender, weight loss etc. To get better treatment and timely administration of TB systems one should use soft computing techniques to abate the risk of human loss.

The aim of this survey is to utilize various parameters that are used for prognosis of tuberculosis disease. In other terms, the desideratum is to study the data for finding its exact results, so that patient can be treated as required without any snag. There are so many data mining techniques proposed by researchers to detect tuberculosis disease such as Support vector machine, Artificial Neural Network, K-Mean Clustering etc. In this probe, Genetic Algorithm along with Neural Network is introduced.

Tuberculosis forecasting systems reduce the cost of treatment by enhancing the timing and frequency of application to control measures and ensure patient to get reasonable treatment at felicitous time span at a lower cost. Previous results use incommensurable data mining methods to give an exact result but attain the lowest accuracy than both genetic algorithm and neural network acquire. To foresee tuberculosis disease, the Genetic algorithm is used to select best chromosomes, afterward apply genetic operators such as crossover and mutation. Neural network performs training and also verify chances of tuberculosis to a particular person.

Tuberculosis is a pestilential disease that affects hominid. Tuberculosis patient's infection gets treated with medicine and panacea from this disease is very arduous. It is feasible by adopting the full course of medicine. The physician takes long time period to recognize tuberculosis. Tuberculosis infection needs treatment on right instance otherwise, it can transform in tuberculosis disease. If tuberculosis is present on lungs, then it can infuse other people.

Figure 1.1 Parameters used for Tuberculosis Disease Forecasting

There are some parameters such as Cough, Chest Pain, Night Sweats, Age, Weight Loss, Gender and Fever etc. are used to predict which person has tuberculosis or not. In this research, different patient records are utilized to detect and prevent TB using soft computing technique.

Tuberculosis is a pandemic illness that affects the immune system. It needs right treatment at a right time. Tuberculosis occurs in all age groups. It is a bug which attacks organs of the human body such as lungs, kidney etc. Cattle, Sheep and other animals are also affected by this disease. Tuberculosis disease test results require patient data. The result of this disease takes too much time; if it is not corrected at the right time then it becomes incurables and attacks the whole immune system badly. So the diagnosis of TB can be done as early as possible.

The remaining paper is organized as follows. Section II shows soft computing technique, Section III depicts related work using various data mining techniques, ceases that genetic algorithm and neural network approach is best for research work and lastly Section IV about the conclusion.

II. SOFT COMPUTING TECHNIQUE

Soft Computing is an aggregation of methodologies that were concocted to model and provide solutions to those problems that are onerous to handle in real world situations. To bestow robustness and low-cost solutions, soft computing technique is introduced. The basic principle is to devise computational methods that result in a satisfactory solution at a low price, by exploring for the comparative solution to an indefinite or definite problem. The use of a hybrid system is increasing day by day with successful applications in the field of engineering design, medical diagnosis, prediction, stock market analysis etc. Fuzzy Logic, Artificial Neural Network and Genetic Algorithm are some soft computing techniques used for disease prediction [3].

The aspiration of hybridization is to surmount the frailty of one technique while applying and emphasize the vigorous of other technique to get the solution by integration. In early days, data mining techniques have been used in the field of medicine and genetics termed as medical data mining [3]. It is progressive research field that is camouflaged by data mining. The medical database incorporates gargantuan data about the patient and their symptoms. Diverse data mining techniques can be applied to different types of disease datasets. These data sets will also facilitate doctors to ensure parameters that are liable for tuberculosis detection.

This study has been implemented by using parameters that are related to a particular disease such as a cough, chest pain weight loss, night sweats, fever etc. are the symptoms taken as parameters of tuberculosis disease etc. Pre-Prediction is useful for both doctor as well as patients in the following way:-

i. If tuberculosis is detected on time by medical practitioner then patient remains alive and get treated as soon as possible.

ii. The physician can easily recognize whether an individual has tuberculosis or not.

iii. The patient can alleviate from such an epidemic disease at the proper time.

iv. Provides treatment to TB patient at low outlay.

v. Suffers commences treatment without waiting for specific test results such as sputum smear microscopy test takes a lot of time to give exact results.

vi. If tuberculosis is identified at first stage then the chances of spreading TB become less.

2.1 GENETIC ALGORITHM

A genetic algorithm is a soft computing technique, first proposed by Holland. A genetic algorithm has been used in the field of pattern recognition, bioinformatics. To upsurge the chance of thriving treatment, early detection of disease is very imperative. There are soft computing techniques that have been employed for identification of a medical problem. Genetic Algorithm begins with a population of arbitrarily generated chromosomes. Each chromosome represents a solution to the existing problem being solved. By applying genetic operators, achieves better chromosomes which are based on genetic process arising in nature. Due to its robust nature, Genetic Algorithm had a great measure of success in search and optimization problems. It is specially developed for big complex and poorly understood search spaces where standard tools are not available, inefficient or time-consuming [2].

The basic principle is to uphold a population of chromosomes. This population progress with respect to time by performing successive iteration for completion and controlled variation. The fitness value is an association of each chromosome that defines the quality of the solution, which is depicted by the chromosome values.

It is search algorithm employed to decipher escalation problems which are based on natural evolution. GA is used for an individual and producing the child for next generation. Genetic Algorithm is also useful for practical purposes due to ease of availability, high-speed computers. This algorithm uses a crossover, mutation and fitness function evaluation. Selection operator is used to electing chromosomes from the present generation to be parents for next generation. To gauge the performance, representation of chromosomes has interpreted. The main element of a genetic algorithm for tuberculosis disease prediction is as follow:

(i) **Fitness Function:** The fitness function is the crux part of every optimization problem. This function is used to yardstick the performance of each chromosome [3].

(ii) **Crossover Chromosomes:** Crossover genetic operator may ensue in two ways. The first member of newly recreated strings is consolidated randomly. Second, carry out crossover on two offspring. In crossover function, chromosomes are taken from a population and spawn young ones.

(iii) **Mutation:** Mutation genetic operators are global searches. It is a mechanism to modify recent solutions to uncover better upshot.

2.2 NEURAL NETWORK

Neural Network proceeds information in the similar way as an individual do. The network is tranquil with a mammoth quantity of interconnected processing neurons, which are running in parallel to interpret the meticulous problem. To allow amalgamated relationships between data of input and output, guilelessly explore the neural network. In order to perform extrusive accuracy in prediction of tuberculosis, supervised learning model has been proposed. The supervised learning model has three distinct layers which are input layer then hidden and the output layer. Back propagation algorithm relies upon weight and it forecast the denouement of the neural network.

Information regarding the neural network is programmed in the genome of the genetic algorithm. At the commencement, a number of random individuals are formed. Neural Network implicates three layers in tuberculosis disease forecasting that are input layer, hidden and the finally the last one is output layer. Its performance can be confirmed after training, with back-propagation.

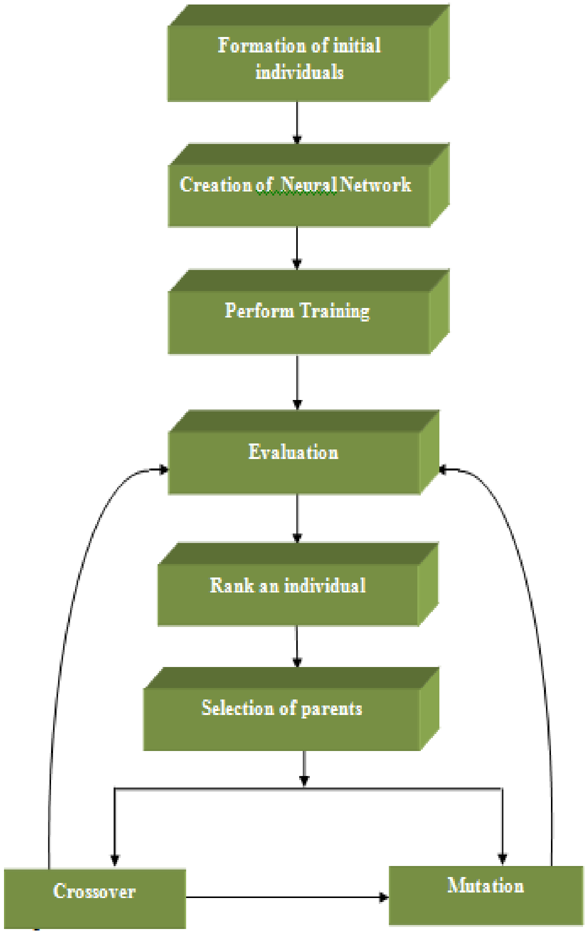

Figure 1.2 Block diagram of Genetic Algorithm and Neural Network

Individuals are evaluated and then provide ranks to it. Evaluation of parameter string must be performed that means a neural network has to be intended according to the genome information. Some Genetic Algorithm Neural Network (GANN) strategies rely upon the Genetic algorithm to uncover an optimal network. The fitness evaluation contemplates not only on the performance of the individual but on consideration of individuals. Selection of parent chromosome is very imperative. To find out best individuals, simple uproot bad patterns. Some approaches acquire the network size in order to spawn small networks. Finally, crossover and mutation generate new individuals. Supplant the worst one with the new population. Genetic algorithm and neural network together endeavor in different ways:-

1. Genetic algorithm depicts the information about how many hidden nodes are presented in the neural network.
2. To set weights of hidden nodes, simply propose genetic algorithm.
3. For decision problems of the neural network, researchers train feed forward neural network using a genetic algorithm.
4. Genetic algorithm is used to elect training data and depicts the output performance of the neural network.
5. It delineates the way that how nodes are amalgamated with each other.
6. Gradient methods have been developed to train weights.
7. To dwindle the number of inputs nodes to get effective results.

III. LITERATURE REVIEW

Riries Rulaningtyas et. al [1] described techniques for image segmentation such as adaptive color thresholding, K-means clustering and K-nearest neighbors. This clustering approach took a long period to find outcomes; hence it did not execute well for an entire image. To surmount local optima problem, K-Means Clustering was used to determine pulmonary TB with segmentation. K-Means Clustering Method was introduced with accuracy 97.90%.

Rusdah et. al [4] determined that support vector machine have the highest accuracy as compared to Bagging and Random Forest. In Early day's tuber test, sputum sneer, microscopy and chest radiography methods were used to identify tuberculosis but these methods wasted time and money and gave poor results. It is a very problematic process if sputum is collected from babies. It needs to train personnel to obtain the desired result without any error; hence it was costly procedure so this research showed that ANFIS was the best approach for diagnosis purpose. **Asha T. et. al [5]** introduced Association Rule Mining which is the process of finding, interesting and unforeseen rules from vast data sets. Association rule mining is used to spawn rules using genetic algorithm. In this probe, the author took attributes to create higher level policy. HIV was an imperative attribute of this analysis. The intention of using a genetic algorithm was to create forecasting principle so that they perform association of attributes in a good manner. This studied to serve as a milestone for the physicians as to how the affected person of abdominal T. B presents.

Shakshi Garg et. al [6] presented a study on the variety of an individual having HIV infection as well as TB. Tuberculosis may repercussion all types of body parts of human being. Over time, TB category has done using a variety of methods like histogram equalization, thresholding method etc. The aspiration of this survey was to produce data mining solution that recognizes TB as perfect as possible. Centroid choice based clustering algorithm PCA, Genetic and Neural Network has used. Result analysis was done by false acceptance ratio, Mean error rate, false rejection rate, recall etc. To achieve better outcomes supplant PCA i.e. Principal component analysis with independent component analysis so that it acquire functions more meticulously. At the end, author concludes 99.7302% accuracy.

TABLE I DATA MINING TECHNIQUE WITH THEIR ACCURACY

Data Mining Technique	Accuracy
K-means Clustering	97.90%
Genetic algorithm and Neural Network	99.73%
Classification	91.33%,
Artificial Neural Network	93.10%
Support Vector Machine	98.70%

Muhammad Khusairi Osman et. al [7] proposed Online Sequential Extreme Learning Machine and Classification of tuberculosis bacilli was categorized into three types-TB, Overlapped TB, nonTB. Sigmoid activation function and 40-by-40 learning mode gave 91.33% as testing accuracy. For improved generalization, OS-ELM was used.

K.W. Becker et. al [8] proposed Neural network, attain lungs sounds of suffers who gets affected by Pulmonary TB from chest walls and of a healthy person as well. Statistical Overlap Factor was used to pinpoint signal features concerned with pulmonary TB. To weigh up the accuracy performance in the better way, other respiratory disorders have to be evaluated in future.

Yang Benfu et. al [9] proposed Artificial Neural Network (ANN). In this survey, whole data was divided into modeling sample and validating sample. By using Training sample, the diagnosis model of smear-negative pulmonary tuberculosis was designed. Validating sample was utilized for generalization of the model. At the end validating sample was used with ANN model and obtained accuracy of 93.10%.

Asha.T et. al [10] described a strategy for computerized diagnosis and classification of TB. Tuberculosis is a condition induced by mycobacterium which develops via the air and impacts entire mechanism poorly. To foresee tuberculosis two different techniques were conglomerated i.e. K-mean clustering and different classifiers. These two techniques are used for those entire individual having HIV sickness. This strategy not only allows physicians to analyze tuberculosis disease but also to carry out various other features engaged within each category in preparing the treatments. When compared with current NN classifiers and NN with GA, author's design produced a perfection of 98.7% with SVM.

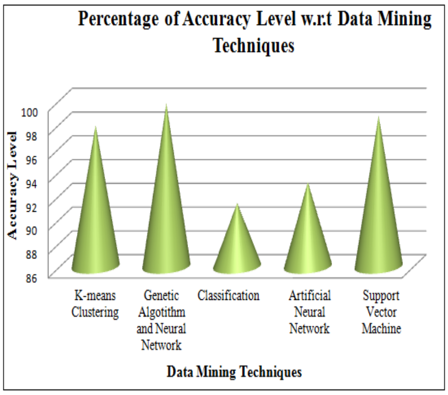

Figure 1.3 Accuracy Comparison used for Tuberculosis Disease Analysis

In above figure, accuracy comparison of diverse data mining technique is shown. Genetic algorithm and Neural network achieves the highest accuracy among other data mining techniques.

IV. CONCLUSION

From above of the papers survey by different authors, conclude that Genetic algorithm and Neural network accomplished the highest accuracy 99.7302% as compared to other techniques. So this probe consummates that soft computing is the most felicitous approach for disease analysis. Tuberculosis is prevalent disease and the cost of treatment is high. If pre-information of disease is predicted at very low cost then it is very effective for human society.

REFERENCE

[1]. Riries Rulaningtyas, Andriyan Bayu Suksmono, Tati Mengko and Putri Saptawati," Multi Patch Approach in K-Means Clustering Method for Color Image Segmentation in Pulmonary Tuberculosis Identification," In 2015 ICICI BME,pp. 75-78.

[2]. Ranno Agarwal," Genetic Algorithm in Data Mining," In 2015 IJARCSSE, Vol 5, Issue 9, pp. 631-634.

[3]. Amarbir Singh and Ramanpreet kaur," A Study of Hybrid Soft Computing Techniques,"In 2015 *IJAFRSE*.

[4]. Rusdah, Edi Winarko," Review on Data Mining Methods for Tuberculosis Diagnosis," In 2013 ISICO.

[5]. Asha T. , S. Natarajan and K.N.B. Murthy," Optimization Of Association Rules For Tuberculosis Using Genetic Algorithm," In 2013 International Journal of Computing, Vol. 12, Issue 2, pp. 151-159.

[6]. Shakshi Garg and Navpreet Rupal," A Data Mining Approach to Detect Tuberculosis Using Clustering and GA-NN Techniques," In 2013 IJSR, Vol 4 Issue 10, pp. 1841 1844.

[7]. Muhammad Khusairi Osman, Mohd Yusoff Mashor, Hasnan Jaafar," Online Sequential Extreme Learning Machine for Classification of Mycobacterium tuberculosis in Ziehl-Neelsen Stained Tissue",In 2012 International

[8]. Conference on Biomedical Engineering, pp. 139-143.

[9]. K.W. Becker, C. Scheffer, M.M. Blanckenberg and A.H. Diacon," Analysis of Adventitious Lung Sounds Originating from Pulmonary Tuberculosis, "In 2013 IEEE EMBS, pp. 4334-4337..

[10]. Yang Benfu,PHD, Song Hongmei,MS, Song Ye,MS, Liu Xiuhui,MS, Zhuang Bin," Study on the artificial neural network in the diagnosis of smear negative pulmonary tuberculosis ," In 2009 World Congress on Computer Science and Information Engineering,pp. 584-588.

[11]. Asha.T, S. Natarajan, and K.N.B. Murthy," A Data Mining Approach to the Diagnosis of Tuberculosis by CascadingClustering and Classification."