Applying Support vector machine for effective pulmonary tuberculosis diagnosis

Israel Mitiku Ersado
Computer Science and engineering departement
Parul University
Gujarat state,vadodara, India

***Abstract*— Tuberculosis (TB) is one of those dangerous diseases caused by bacteria whose scientific name is mycobacterium. The spread of Tuberculosis is increasing mainly in developing countries for recent two decades. Oppositely the availability and the size of datas are getting increased but the problem is that handling those problems in good manner and extracting useful pattern of the data to explore hidden knowledges and pattern of the disease. In order to solve the problem of the complexity the Machine learning and data mining techniques should be applied. This study is proposed to apply those techniques in order to diagnosis the disease efficiently and effectively. In this study classification is used to identify patterns and predict the occurrence of TB and to increase the accuracy of the result before previous studies.**

(*Abstract***) *Keywords—Tuberculosis, Data mining, Support vector Machine, (****key words****)***

I. INTRODUCTION

The healthcare industry currently is supported by different and many technologies and science products in various aspects so that is improving its service the customers day to day and time to time. There are too many tasks in the area of medical services that cannot be done manually, if done manually the outcome may not be accurate, timely and problem solving, so that it is believable using different computer related techniques.Data mining on area of solution to the problems in medical services such as data management, data retrieving pattern identification, prediction and diagnosis of different types of diseases.

Developing countries mostly and some developed countries are experiencing a huge problem of disease mainly related with transmitted diseases because the life level of the people is low so that the problem gets so difficult. Shortage of medical service providers and high turnover of peoples from place to place; inadequacy of essential drugs and supplies have also contributed to the burden, [1] Tuberculosis (TB) is one of those dangerous diseases caused by bacteria whose scientific name is mycobacterium. This disease without effective diagnosis and treatment cannot be diagnosed and treated easily so that it needs strong and effective diagnosing and predicting techniques. There are two broad classification of tuberculosis diseases those are Pulmonary tuberculosis and infant or inactive tuberculosis, not contagious and it has no symptoms so that it is difficult to detect the second type of tuberculosis, so that in this dissertation we are dealing with the first one pulmonary (active) tuberculosis .TB can remain in an inactive (dormant) state for long time without causing symptoms or transmitted to other peoples. After the immune system of a patient with dormant TB weakens, the TB can became active and cause infection in the lungs or other parts of the body

Now a days in different organizations and areas datas are getting larger and larger day today so that to analysis those datas and to find out the genuine and hidden pattern between different variables and instances .In recent decades data mining is becoming popular in medical areas, hospitals and clinics to manage and analyse important information by identifying hidden and unknown patterns important information. Data mining is one of the Artificial Intelligence areas that enhances the employees of medical fields and such as doctors, admistrative persons and other low level employees and on the other side supports the customers to remove unnecessary time cost and incorrect prediction and other many benefits. Data's especially in medical areas are complex so that cannot be analysed by using human power .so that using machine learning and data mining approaches it is possible to analyse and increase the performance the diagnosis of Tuberculosis by collecting the important symptoms the patients has by analysing those potential attributes its about to predict the wether the patient is free or not from Tuberculosis dieseases.

Statement of problem

Medical areas are one of the areas that a hug data are available , handling those data is big problem ,as the data getting larger more errors can occur and it leads to wrong prediction . Tuberculosis (TB) is a big challenging problem in most developing countries. This disease sometimes become difficult to diagnose by hiding itself in the blood of the patient it may stay for a long years so that it need very strong diagnosis tool rather than currently used in most of medical services it become more complex when it became combined with other diseases.[3]

The problem occurs as the time data's are getting large and large in size that they need more strong and flexible data modeling techniques and diagnosis mechanism in order to remove problems related with data handling and to increase the accuracy and performance of the result. The other problem is also that the tools that mostly used in diagnosis and prediction of disease and the problem happens during data mining times due to the complexity of the diseases health professionals may not be able to find out the disease simply they need more computerized systems rather than manual and traditional approaches .

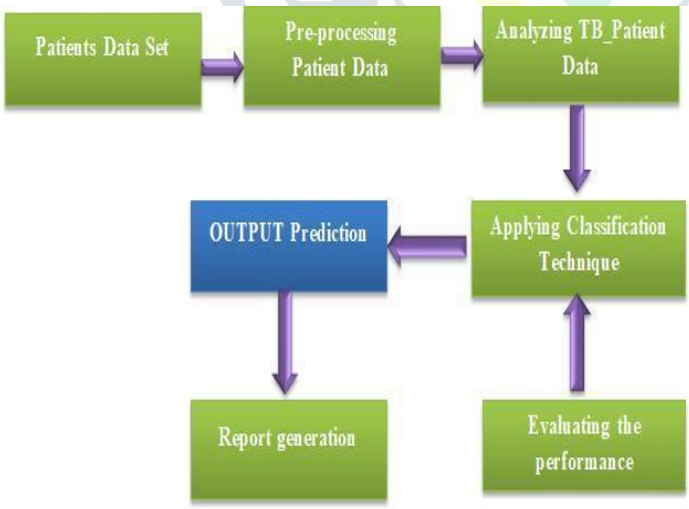
The machines currently used for the purpose of diagnosis and prediction are having side effects over the patient, poor quality, need long hours to perform the diagnosis and to show the result and the other problem is that some machines also need special professionals who can read and define the output pattern of the disease.

The main problem related with algorithms done in previous research works in diagnosis and prediction areas are the tool compatibility problems the tools that used to apply the algorithm are not suitable; the other challenge is the format diversity problem because the data is stored in medical areas are in different formats.so that it should be transformed. The performance of algorithms from data to data and from tool to tool. All techniques do not have same challenge it varies from one to other.

2. ALGORITHM

2.1 **Support Vector Machine**

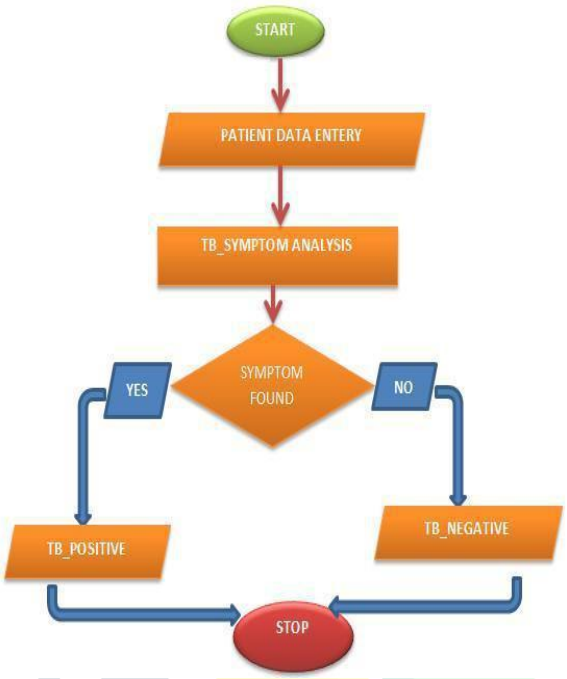
In very recent years when trying to review more papers found SVM frequently used and its accuracy is high when compared to others. Support vector Machine is one of the known classifier algorithm in data mining and machine learning area of field .In our case it helps to predict patient whether he/she has TB or not based on the training data set. In following example blue part on the left side are tuberculosis negative and on the right side are tuberculosis positive



Classification algorithm is going to be applied to analyse data in order to obtain models that are used to characterize data classes. This task works on predicting the status of patients related with the disease in what class they are going to be assigned among predefined classes. It is possible to divide data classification task into two phases. [4]

1. *Learning step*: predetermined set of classes will be constructed. This operation is made by analysing a set of training data. In this step, each patients are assumed to belong a specific, predetermined class
2. *Testing step*. The constructed model is tested by using different set of data if the estimation of accuracy shows an adequate result, then the generated model can be used for classification of new patients sets whose class labels are unknown.[5]

In this study Support vector Machine selected specifically among classification algorithms



ACKNOWLEDGMENT

The first and the most special thanks go to the almighty GOD, thank you for giving me favors in your sight. All I am and all I have, it is because of you GOD. I would also like to thank my advisor Prof.Harshal Shah for advice and continuous support.

REFERENCES

[1] Abraham, T." Application of data mining technology to identify determinant risk factors of Tuberculosis to find their association rules, "Addis Ababa university ICT conference", Ethiopia, vol 1,February,20015,27-54

[2] J. Han and M. Kamber , Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches, "clustering and classification exploration Sanfrancisco conferences" , march 2011, 47-97

[3]Asha.T, S. Natarajan and K.N.B. Murthy, "Diagnosis of Tuberculosis using Ensemble methods", IEEE, Dec 2010, 978-1-4244-5539-3/10.

[4] WHO, "Global Tuberculosis Report 2012," 2012, 34-43

[5] X. Y. Djam and Y. H. Kimbi, "A Decision Support System for Tuberculosis Diagnosis," The Pacific Journal of Science and Technology, 2011,410–425

[6] PDPI, Tuberkulosis Pedoman Diagnosis dan Penatalaksanaan di Indonesia,"Effective utilization of data mining tools conference" august 2012,12-14