IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SURVEY ON DATA MINING AND INFERENCES IN TUBERCULOSIS MEDICAL DATA

T.Baskar [1] and M.Kannan [2]

[1] Ph.D Research Scholar, Dept of CSA, SCSVMV, Kanchipuram, India.

[2]Dept of CSA, SCSVMV, Kanchipuram, India.

ABSTRACT

Data analysis and inferences from medical data system where the data growth is unpredictable in sizes, physicians and medical researchers can face issue in analysis and handling the data due to its increasing volume and variety. Hence traditional analysis or statistical analysis has become insufficient and its method of knowledge mining that incorporate tasks such as Knowledge extraction, data archaeology, data exploration, data pattern processing, data dredging, information harvesting, and other related techniques of data discovery in databases and data analysis required by the intelligence. Diseases are sometimes difficult to spot due to similarities of symptoms or other reasons. For example, if the diseases are unable to diagnosis because of the ambiguity symptoms, irregular sample collection, and medical errors.

This review article discusses a key element in data processing and aims to elaborate a process discovered for better Knowledge Discovery in sharing special symptoms about the occurrence of other diseases such as – tuberculosis (in short called as TB). Further, by Classification Rule, Decision Tree algorithm, and Prediction Tool to infer the defining characteristics, researchers try to supply information relevance of various test components and hence to discover hidden knowledge, unexpected patterns, and new rules from the database. The aim is to identify a new way of data mining processing which has strong impacting reasons to improve further research in Tuberculosis and Data Mining Techniques.

KEYWORDS:

Tuberculosis, Symptoms, Data Mining, Knowledge Discovery.

INTRODUCTION

This review paper discusses the types of issues solved by data mining techniques in the medical fields to provide indirect support to help medical people to understand the sort issues that data mining can address in diagnosing Tuberculosis and how to interpret the mining results. This generic method describes how to translate the issues into data-mining problems and some common data models that can be used.

The aim of this survey paper is to explain how to choose proper data mining techniques and how to interpret and deploy the results. Although in-depth of knowledge in data mining requires facts to understand data mining technology. Researcher defines a method that helps to collect information about related and different test components. Because some test components are more important than others are, the sufficient order of these components or the right choice of the test components may lead to faster and more secure strategy to find. Consistent in their medical records to have significant difference between those patients and people who were fully investigated for parameters like age, sex, symptoms, and prevalence of unknown HIV infection. There are many techniques that allowed by classification models based on historical data used by realizing a more effective tool for diagnosing tuberculosis and prediction.

Objectives

The key objectives of this research article are:

- To gather statistical information about TB diseases.
- To summarize the various research papers that highlight about TB related data analysis.
- Data Mining remains highly active in this research area rather than other current technological area in medicinal research platform.

LITERATURE REVIEW

Jonathan and Joan [1] diagnose TB disease through trained African giant rats for data processing techniques. Some of the techniques used are decision trees, random forest, and Naive Bayes algorithms and R Language tools. He discovered that the trained African giant hero rats have spotted TB, alternatively in use of microscopy and other analytical tools. The rats inhale sputum samples to detect TB diseases. The information was collected from different APOPO TB training and research facility in Morogoro, Tanzania. The performance of ratsis measured regularly to diagnosis TB process. The sizes of the dataset were 471, in which 133 rats were used. The accuracy of the algorithm was Random forest were 78.82%, decision tree were 78.78% and naive Bayes were 78.71%. In future author suggested implementing this process of using rats to diagnosis TB easily.

Nagabhushanam D et.al [2], TB diagnosis through the info mining technique "Adaptive network-based fuzzy inference system"(ANFIS), a multilayer perception and PART model using equivalent data set. The ultimate result is that, ANFIS is better than other two algorithms. The data set about 667 different patient records and 30 variables were used for the research, the data is collected from private clinics. Ucar et.al [3] also used same ANFIS algorithm, Multilayer Perception and PART model. World Health Organization (WHO) based on standard of Direct Observation of Therapy (DOT) predicates the data collected from private clinics and variables. Both got same result, ANFIS is an accurate compared with others.

Garg et.al [4] surveyed about TB, various data processing algorithms like color segmentation, thresholding and histogram equalization. They used genetic algorithm and neural network to diagnose and prevent TB. [5] Techniques used are Centroid selection based clustering algorithm to enhance the clustering scheme, PCA for extraction, genetic algorithm for optimization and neural network for training and testing purposes. Based on the results, supported accuracy, false acceptance and rejection ratio were found.

T. Asha et.al [6] studied, Tuberculosis co-infected with HIV/AIDS. The medical data were collectedfrom 700 records of TB patients from the town hospital. Totally 12 attributes are used with related symptoms, and test detail of patients are applied in Apriori and Association Rule Mining algorithm. The result shows that there is a connection between one symptom with the opposite and findings of the hidden relationship. This made the process to diagnoses TB patients co-infected with HIV/AIDS as early as possible.

V. Priyavadana et.al [7] encapsulated techniques on info mining techniques to predict many diseases. Different mining tools were used to predict the accuracy level in several healthcare problems: they're heart disease, cancer, HIV, Tuberculosis, DM, Kidney dialysis, Dengue, IVF and hepatitis C. Different data processing algorithms are implemented for diseases mentioned above. For instance, Cancer — Rules multidimensional language, Tuberculosis — KNN algorithm. Combined effect of more data processing technique is used to improve results.

Muhammad Tahir Khan, et.al [8] discussed some approaches like prediction of TB based on ANN algorithms. The data were collected from TB Units and Health centers and trained on 12,636 records of TB patients, during the years 2016 and 2017 from TB units of Khyber Pakhtunkhwa, Pakistan. ANN based evaluation takes the knowledge based on factors like gender, age, HIV-Status, previous TB history, sample type, signs and symptoms for TB prediction. The Accuracy of ANN to predict the MTB resulted in 94% success rate.

N. Suresh, et.al [9] studied the Random Forest algorithm technique for tuberculosis prediction. The Dataset includes 1250 details collected from the town hospital. They are measured only active and latent TB. The values used to differentiate the diseases were one (1) show positive and high value and zero (0) show negative and low value. The TB with HIV patients are counted in Latent TB, the results wereanalyzed using WEKA Tool (OpenSource Java tool for data classification and analysis in which Classification and Clustering are calculated for Analytical and graphic analysis respectively).

AyeshaSadiya, et.al [10] analyzed how machines learning techniques are used to diagnose Tuberculosis and Pneumonia, Both having identical symptoms at its initial stages. By mistake non TB patients are unnecessarily treated as TB patients, when they treated with TB drugs, they have side effect. They used three algorithms — ID3, Naïve Bayes and Random Forest that fit the dataset and provides the very best accuracy. Within the Dataset, it includes 705 instances, 32 attributes, and a couple of classes. They observed that the technique had improved the performance of all three algorithms viz ID3 with success rate of 93%, Naïve Bayeswith success rate of 92%, Random Forestwith success rate of 97%.

Bukola and Badeji [11] interviewed the staff of the TBL center to spot the danger factors of people who came treatment for Tuberculosis at the TBL centers in Ado Ekiti. The data set of 699 patients data were preprocessed, 10-fold cross-validation technique and Naïve Bayes' classifiers model were used to predict the TB with a minimum of 92% accuracy.

Arnold, et.al [12] predicted the data mining technique to classify the treatment relapses of TB patients. The dataset is applied and tested within the decision tree J48 algorithm using WEKA tool, which identified three significant independent variables (DSSM Result, Age, and Sex only) and 90.39% of the accuracy of the J48 algorithm. The dataset used for analysis was collected from the town Health Office of Cabanatuan.

Jackson A. Killian, et.al [13] observed TB Treatment to improve patient care supported Digital Adh/erence Technologies (DATs) in Mumbai city served by DOTS. They found the missed dosage patients' data of nearly 17,000. The table contains parameters like weight, age, gender, center ID, start and end date of treatment, and also included, treatment is completed, or ongoing, or died, and lost to follow-up. To review the patient, regular observations support on mapping area of patients using phone numbers was used. The decision log and doctor interacts with a patient's data are recorded automatically within the DOTS dashboard. They studied model Decision Focused Learning, LEAP (LstmrEal-time Adherence Predictor), ROC Curve and SMOTE (Synthetic Minority Over-Sampling Technique (Python library)). They improved the performance of 15% to recover the lost to follow-up patients.

Christopher et.al [14] discusses the factors associated with the treatment failure of which patients are at a high risk due to treatment failure. The attributes used for prediction were country, age, sex, education level, employment status, number of daily contacts, sort of resistance, number of X-rays, social risk factors, etc. various sorts of techniques were used: step wise forward selection, step wise backwards elimination, backwards elimination and forward selection, the least Absolute Shrinkage and Selection Operator (LASSO) regression, random forests, support vector machine

(SVM). The performances were measured using (AUC: 0.74), statistical analyses with R (version 3.2.2) and packages gmodels, caTools, MASS, glmnet, randomFores, e10701, mice, gridExtra, ggplot2, and pROC. The database includes 587 patients with median years. Treatment failure happened in additional or less one-fourth of the patients.

HardikManiya et.al [15] studied comparative of two algorithms, Naïve Bayes Classifier and KNN for diagnosing the Tuberculosis disease. To pick the simplest symptoms for classification, a complete of 154 records with 19 variables supported symptoms were collected from SardarGopaldas TB Center, Gujarat, India. The info sets were implemented in C language and WEKA tool. Naïve Bayesian got better results and 78% accuracy compared to KNN.

Jiyang Wang, et.al [16] data investigation to predict Tuberculosis stand on four differing types of income groups from 2000 to 2016. They used differing types of methods and techniques: Kruskal-Wallis (KW) test, multivariate analysis, Cuckoo Search (CS) Optimization, Combined Forecasting Method, and Radial Basis Function Neural Networks (RBF). Income classified by gross value (GNI), per capita consistent with the planet Bank. The income classifications were: low-income, lower-middle income, upper-middle income, and high income of four types collected for nearly 17 years. Results of multiple comparison tests like, low vs. high, high vs. upper, there are six pair of groups. The groups got 95% confidence intervals of the mean ranks of tuberculosis. Frequency rate found that lower-middle and low-income groups persons have more Tuberculosis diseases compare to other groups.

Ashwini D. V, et.al [17] surveyed about Tuberculosis with or without HIV co-infected patients' analysis through a special machine learning algorithms. The human Gene system expressed in several diseases, and therefore, the cells react to a specific treatment for the diseases. They analyzed different data processing techniques were: organic phenomenon Analysis in HIV/TB co-infection using KNN algorithm, HIV/TB co-infection analysis using the support Vector Machine, HIV/TB analysis using Gini Co-relation Co-efficient using Pearson coefficient of correlation (PCC) method, and HIV/TB analysis using Artificial Neural Networks. The discussed algorithms were utilized in the medical field to enhance the healthcare of patients.

Sourabh Shastri, et.al [18] study the Bacillus CalmetteGuerin (BCG) a vaccine to prevent tuberculosis and predicted the BCG percentage for subsequent five years (2015 to 2019) found on previous year data of BCG coverage in India. The info collected from Unicef global databases since past nearly 35 years (1980 to 2014). To predict the longer term BCG treatment in India, by using exponential smoothing techniques of your time series analysis.

Sathishkumar R, et.al [19] compared various diseases using the different machine learning algorithms. The names of diseases were: carcinoma, TB, Diabetes, and Heart diseases. The varied machine learning algorithms were: Decision Tree, Naïve Bayes, Neural Network, SVM, Random Forest, and Logistic regression. It's going to help to extend the choice making speed and reduce the error rates.

Vijayalakshmi N, et.al [20] analyses to bring out the main factors causing infertility in women using data processing algorithms. Reasons for infertility in females include sexually transmitted diseases, DNA damages, DM, ovulation problem, tubal blockage, pelvic inflammations because of Tuberculosis, and age-related factors, etc. the whole data set is additionally subject to classify using two different Decision Tree Induction were: (J48 and Random Tree). Association Rule Mining and clustering techniques were also included in extracting information stand on infertility. Comparative study of various algorithms to supply almost similar results, and therefore, the accuracy of prediction is 86%.

Statistical report 2019(WHO) [21] reveals on Tuberculosis: an estimated 10.0 million people fell ill with TB in 2018. Globally, there have been 1.2 million TB deaths among HIV-negative People in 2018. Eight countries accounted for Two thirds of the worldwide total: India (27%), china (9%), Indonesia (8%), Pakistan (6%), etc. In 2018, there have been approximately half 1,000,000 new cases of rifampicin-resistant TB (of which 78% had multidrug-resistant TB). The three countries with the most important share of the worldwide burden were India (27%), china (14%), and Russia (9%). Sources of knowledge estimates TB disease have improved considerably in recent years.

Navneet Walia et.al [22] designs a choice network for tuberculosis diagnose-ability to propose a fuzzy approach. The tuberculosis dataset of 65 patients collected from government health clinic and used nine numbers of the input values fuzzy implication system is made. A fuzzy function takes values between (0, 1) to perform a diagnosis system. The rule deals with the patient symptom and makes choice supported fuzzy rules. The results of Accuracy value is 77%.

Rakhmetulayeva S.B et.al [23] analysis the classification algorithm supported support vector machines (SVM) for determining the effectiveness of treatment on tuberculosis. The basics of preventive diagnosis on tuberculosis patients and develop an optimal treatment course. Different types of methods are used to analyze the possible condition of the patients. There are, Statistical methods of knowledge processing (SMDP), Artificial neural networks(ANN), Nonlinear regression methods(NRM) and Reasoning on the idea of comparable cases(RBSC). Experiment data was collected within the database, which incorporates health status in several periods of the patient. The Gretl open-source software was utilized in this research using which 95% accuracy of result is achieved.

M. Sebban et.al [24] studied a data-mining approach to spacer oligonucleotide(DNA analysis technique) typing of tubercle bacillus may be a suitable model. C4.5 induction algorithm was analyzed and suggested that both negative and positive constraints within the Direct Repeat locus. The principles are different and simpler than those earlier defined by the expert, and therefore, the decision trees use pruning to avoid over fitting. Data processing method to form rules for solving classification tasks from existing or particularly created databases.

T. W. Rennie et.al [25] discussed the info mining of tuberculosis patient data using multiple correspondence analyses. Data collected from North East London medical care trusts between the years 2002 to 2007 were used. Clustering patterns in Multiple Correspondence Analysis (MCA) output display different associations of variable categories utilized in TB epidemiology. Results suggest this tool to commissioning of TB services.

Brudey et.al [26] analysed the tubercle bacillus complex (MTC) genetic diversity through mining. The spoil go typing database (SpolDB4) for classification, population genetics, and epidemiology. Spoil go typing is the commonly used PCR-based reverse-hybridization blotting technique that assays the genetic diversity of this locus. Data collected from spoligotyping databases (SpolDB4), clarify 1939 shared-types (STs), 39,295 injure from 122 countries, which are divide into 62 clades. An updated my SQL-java-based version is employed. Combined Automatized-Expert Based Classification of Spoil go types algorithms were established.

Result as long as a large-scale abstract structure of the worldwide TB epidemiologic network.

K. R. Lakshmi et.al [27] summarizes various review utilization of knowledge mining techniques for prediction and diagnosis of tuberculosis diseases survivability. The dataset includes 700 records of patients diagnosed with TB obtained from Osmania hospital Hyderabad, Andhra Pradesh, India. The dataset used in ten algorithms and implemented altogether. Finally, PLS-DA was the simplest one compared with other algorithms.

Amit Jain et.al [28] examined the soft fuzzy model for mining aminoalkanoic acid associations in peptide sequences of tubercle bacillus complex. The dataset collected from NCBI for overall total 83,086 patient records. Result through fuzzy set approach found the uncertainties with degree of relationship and patterns on parameters. Also, the physic-chemical properties and secondary structures had been predicted support on amnio acid association patterns.

Orhan Er et.al [29] observed the tuberculosis disease diagnosis using artificial neural networks. Differing types of layers were compared for diagnosis TB. The layers were Multilayer Neural Network (MLNN) and a General Regression Neural Network (GRNN). The patient's reports were taken from Diyarbakir Chest Diseases Hospital from southeast of Turkey was used. They collected 150 samples, which include thirty-eight features. There are, cough, weakness, sputum, RBC, etc. On comparing two algorithms, the simplest results for the classification accuracy were obtained from the MLNN structure with two hidden layers for diagnosing TB disease.

Babu C Lakshmanan et.al [30] studied the info mining with the decision tree to gauge the pattern on effectiveness of treatment for consumption. Clustering and classification techniques are used to find the drug resistance for all drugs or any drug supported age, and weight of the patients. The info collected from National Institute for Research in Tuberculosis (ICMR), Chennai. There have been 1237 patients' records, which contain age, weight, sex, drugs susceptibility test (DST), and treatment group. The results compared with CART and CHAID procedures, the very best predicted rates 97% goes to CART. In order that CART could also be an effective prediction mechanism compared with CHAID.

ManishShukla et.al [31] design a hybrid approach for tuberculosis data classification using the optimal centroid selection based clustering. Tuberculosis dataset taken from local hospital, which incorporates 700 instances and eight attributes for experiment purpose. During initial centroid selection for k-mean clustering this provides better results supported health data.

April Rose C. et.al [32] analyzed the rule-based fuzzy diagnostics decision network for tuberculosis. Algorithm used to found symptoms on tuberculosis patient. Physicians interviewed to classify the intensity of every symptom consistent with its description given. The physician input all essential values of symptoms needed for systems purpose which includes, cough (0-10), fever duration (0-30 days), sputum (0-10), etc. Using symbolic logic, there have been 16 rules for conditions A to E and 323 sets of rules for the aim of the category of tuberculosis the patient has. The pulmonary physicians to scale back the time consumed in making diagnosis using symbolic logic. K. Soundarajan et.al [33], analysis same algorithm to diagnosis TB using fuzzy logic.

Rahul Hooda, et.al [34] examined the potential method for tuberculosis detection using Chest Radiography in deep-learning. Chest X-rays images classified into two categories, one is normal and other abnormal. Convolutional neural network (CNN) methods with Seven convolutional layers and three fully connected layers were used to classify the CXR images. The performances compared with three optimizers namely Adam optimizer, Momentum optimizer, and Stochastic Gradient Descent (SGD) optimizer. Dataset collected from Montgomery County (MC) and Shenzhen dataset. The MC includes 138 CXR images contains 80 normal images and 58 abnormal (TB) images and Shenzhen includes 662 images contains 326 normal and 336 abnormal images. Results supported accuracy obtained using The Adam optimizer is 94.73 attempts to validation accuracy is 82.09%. Compared with other optimizer the Adam optimizer performs best amongst the three.

Adnan Fojnica, et.al [35] study the tuberculosis prediction supported Artificial Neural Network. Data collected from the Clinical Centre University of Sarajevo within the period of two years, which incorporates 1400 patient records. The ANN correctly classified 99.24% patients, out of 1315 patients. Results supported ANN, sensitivity of 99.24% and 100% of specificity in classification of tuberculosis disease.

Ilena, Gabriella, et.al [36], early detection of tuberculosis using Chest X-Ray (CXR) with Computer-Aided Diagnosis. The systems supported image processing to extend image quality were applied homomorphic filter, histogram equalization, median filter, and Contrast-Limited Adaptive Histogram Equalization (CLAHE). The CXR images collected from the Shenzhen Hospital database. Which incorporates 326 normal and 336 abnormal Posterior-Anterior (PA) CXR images. Results found on 76% of accuracy supported CADx. Seelwan Sathitratanacheewin [37] Deep Convolutional Neural Network (DCNN) techniques to diagnosis tuberculosis using CXR data set. Data collected from same hospital mentioned above.

Kamal J. AbuHassan et.al [38] study the automated diagnosis of tuberculosis disease supported plasmonic ELISA and color-based image classification. They develop a mobile-based point-of-care (POC) platform for TB diagnosis. The plasmonic ELISA experiments were supported synthetic biological samples also as real samples. Image segmentation algorithms were coded in MATLAB. Photos were taken at a different distance between the mobile camera and therefore, the ELSA plates. The mobile images converted into the CIELAB color space and implemented the k-means clustering and thresholding algorithms. They need two approaches, one is biosensing mechanism and second is, differentiates the classification performance of various sorts of classifiers. The classifiers were, Bagged trees, boosted trees, fine KNN, cubic SVM, and medium decision tree. The five-fold cross-validation results are 97.2 % accuracy rate got by the bagged trees.

Ferani E. Zulvia et.al [39] discusses the tuberculosis diseases employing a multi-objective gradient evolution-based support vector machine and C5.0 decision tree. The TB Data collected from five hospitals in Palembang city, Indonesia. Dataset contains 374 instances with nine values (age, sex, BCG, etc). Positive TB patients are 187 and 187 non-TB patients, overall 374 patients. They developed a Multi-Objective Gradient Evolution Support Vector Machine (MOGESVM) algorithm to find the simplest experimental results compared with other algorithms.

Awad Ali et.al [40] two machine learning algorithm implements to diagnosis the tuberculosis treatment. The two algorithms were, Naïve Bayes and Bayesian Networks. The data collected from Mahojub, Epidemiology laboratory in Sudan. The dataset includes 54 attributes and 714 instances.

Data classified based on TB treatment, which specified, a = cured, b = complete treatment, c = died, d = failure, e = defaulter, and f = transferred-out. Dataset implemented in both the algorithms, the results were 90.75% for Naïve Bayes and 93.27% on Bayesian Network.

Asia Nesredin [41], effective methods using classification and clustering techniques to diagnosis tuberculosis. Data set contains 7069 patient details from Menelik II hospital, Ethiopia. Researcher apply k-means clustering algorithm with some improvement to separate the TB-positive and negative patients in the dataset. The J48 decision tree and Naïve Bayes classification algorithm implemented to identify model and predict the incidence of TB. Accuracy of the algorithm got the 85.93% to develop for predicting TB diagnosis and used WEKA tools.

From the above review papers, the researcher highlights the outcome and presented in table1, figure1, and figure2 with respect to methodology, tested samples and accuracy rate.

Table 1: Summarize of review papers

Author	Methodology	Samples tested	Samples collected from	Year	Accuracy rate	Remarks
Jonathan and Joan [1]	Random Forest	471	APOPO TB, Tanzania	2019	78.82%	African Giant Rats are used to spot TB
Ucar, et.al [3]	ANFIS	667	Private Clinics	2011	97%	Getting higher Accuracy using ANFIS algorithm
T. Asha, et.al [6]	Apriori and CBA	700	Town Hospital	2012	81.14%	Using Customization rule for getting better accuracy
Adnan Fojnica et.al [35]	ANN	1400	Clinical Centre University of Sarajevo	2016	99%	ANN capable for handling large volume of data
AyeshaSadiya et.al [10]	Random Forest, ID3, and Naïve Bayes	705	Local Hospital	2019	97%	Uses Hybrid algorithm in combination of multiple data processing.
Jackson et.al [13]	DATs, LEAP (Lstm rEal-time Adherence Predictor)	2700	Mumbai City	2019	90%	Uses Machine learning algorithm to process large volume of data
Awad Ali et al[40]	Naïve Bayes	714	Mahojub, Epidemiology laboratory in Sudan	2012	93%	Uses customization for parameter configuration to improve efficiency
Vijayalakshmi N et.al [20]	/J48	575	Research Center, Tirchy	2016	86%	Infertility in women using data processing algorithm
Navneet Walia et.al [22]	Fuzzy Approach	650	Government Health Clinic	2015	77%	Better performance to Diagnose TB using Fuzzy System
Babu C Lakshmanan et.al [30]	CART	1237	ICMR, Chennai	2015	97%	CART unusual Algorithm to predict the TB

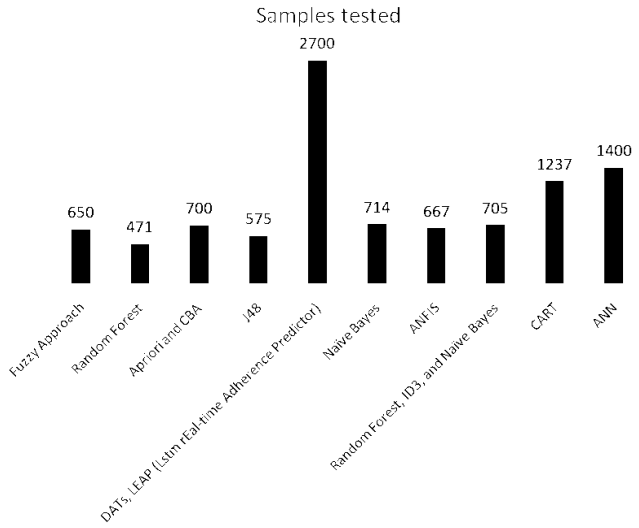

Figure 1: Samples taken for their research work

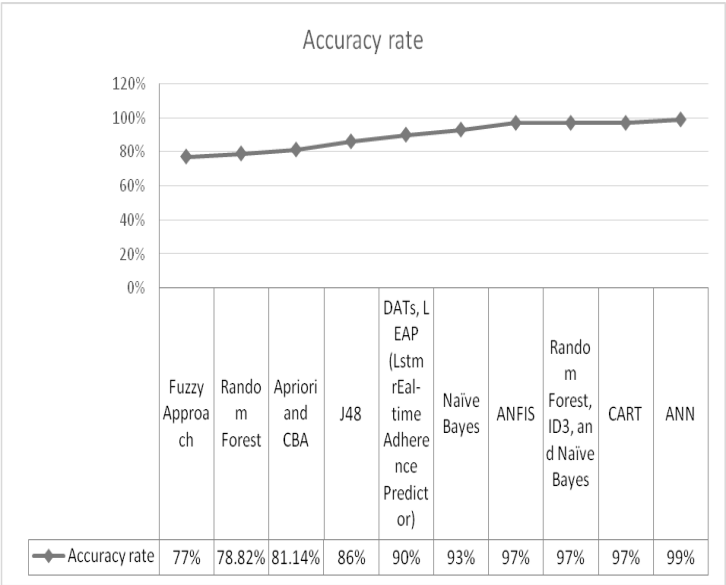

Figure 2: Accuracy Rate

RESEARCH GAP AND SCOPE FOR FURTHER WORK

With the above literature review on the topic, there are some gaps identified in the existing methods and area of improvement by introduction of new method in TB data analysis may highly improve this entire process. Some of the gaps identified are listed below:

- Most of the researchers concentrated on TB, PTB with diagnostic techniques using Decision tree, Naïve Bayes classification, Support Vector Machine (SVM), Artificial Neural Network (ANN) and other classification methodologies.
- Very few researchers have taken the sample population from a specific location (eg:Tamilnadu) to address technical problems and analysis in data gathered from tuberculosis issues.
- Hence, the researcher tries to explore Extra-PTB, HIV co-infected with TB and TB with COVID-19 based on standard methods with more number of attributes.
- Also aims to explore hybrid or novel approach to predict TB disease.

CONCLUSION

This review article addressed Tuberculosis related issues, instead of concentrating on the ability of data mining techniques that are available to infect, tuberculosis related data offers a variety of other tools and techniques that are not been ready to describe the way to use. Also, it gives an opportunity to explain the advantage and benefits of using data processing for the sample data collected from different sources. This study depicts the diagnosis of tuberculosis patients has increased much within several years of investigation. By bridging the gap to develop a new process for TB data analysis, the researcher can travel in-depth analysis by designing an expert system tool to predict TB diseases and to focus on recent medicinal infections, which causes major dominance on society, with newly discovered issues, which created major severity. Further, this work aims to integrate different process such as classification, association and prediction named as expert tool or novel approach.

REFERENCES:

[1] Jonathan, Joan. "Prediction of Factors Influencing Rats Tuberculosis Detection Performance Using Data Mining Techniques." (2019), Spring.

[2] Nagabhushanam, D., et al. "Prediction of tuberculosis using data mining techniques on Indian Patient's Data." *IJCST* 4.4 (2013).

[3] Uçar, Tamer, and Adem Karahoca. "Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches." *Procedia Computer Science* 3 (2011): 1404-1411.

[4] Garg, Shakshi, and Navpreet Rupal. "A review on tuberculosis using data mining approaches." *International Journal of Engineering Development and Research* 3.3 (2015): 1-4.

[5] Garg, Shakshi, and Navpreet Rupal. "A Data Mining Approach to Detect Tuberculosis Using Clustering and GA-NN Techniques." IJSR, ISSN: 2319-7064,Vol 4,Issue 10,oct-2015.

[6] Asha, T., S. Natarajan, and K. N. B. Murthy. "Data mining techniques in the diagnosis of tuberculosis." *Understanding tuberculosis-global experiences and innovative approaches to the diagnosis*. InTech, 2012. 333-352.

[7] Priyavadana, V., A. Sivashankari, and R. Senthil Kumar. "A comparative study of data mining applications in diagnosing diseases." *International Research Journal of Engineering and Technology* 2.7 (2015): 1046-1053.

[8] Khan, Muhammad Tahir, et al. "Artificial neural networks for prediction of tuberculosis disease." *Frontiers in microbiology* 10 (2019): 395.

[9] Suresh, D. Arulanandam, "A Mining Approach for Detection and Classification Techniques of Tuberculosis Diseases", International Journal of Advanced and Innovative Research, ISSN:2278-7844, Volume 7, Issue 2, 2018.

[10] iyeshaSadiya, Anusha V Illur, Aekhata Nanda, Eshwar Rao, Vidyashree K P, Mansoor Ahmed, " Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8, Issue-6S4, April 2019.

[11] adeji – Ajisafe, Bukola, "Bayesian Classification Model in Predicting Tuberculosis Infection", IOSR Journal of Computer Engineering (IOSR-JCE, e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 20, Issue 4, and Ver. I (Jul - Aug 2018), PP 06-16.

[12] Arnold P. Dela Cruz, Gilbert M. Tumibay, "Predicting Tuberculosis Treatment Relapse: A Decision Tree Analysis of J48 for Data Mining", Journal of Computer and Communications, 7, 2019, 243-251.

[13] Jackson A. Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, BistraDilkina, and Milind Tambe, "Learning to Prescribe Interventions for Tuberculosis Patients Using Digital Adherence Data", In The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), vol 4, iss 8, August 2019, 9 pages.

[14] Christopher Martin Sauer, David Sasson, Kenneth E. PaikI, Ned McCague, LeoAnthonyCeli, Iva´nSa´nchezFerna´ndez, Ben M. W. Illigens, "Feature selection and prediction of treatment failure in tuberculosis", Research Article PLOS| ONE, November 20, 2018.

[15] HardikManiya, Mosin I. Hasan, Komal P. Pate, "Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis", *International Conference on Web Services Computing (ICWSC) and Proceedings published by International Journal of Computer Applications® (IJCA), 2011.*

[16] Jiyang Wang, Chen Wang, and Wenyu Zhang, "Data Analysis and Forecasting of Tuberculosis Prevalence Rates for Smart Healthcare Based on a Novel Combination Model", MDPI Applied Sciences Article, 18 September 2018.

[17] Ashwini D.V and Dr. Seema S, "Machine Learning Approach to Detect Tuberculosis in patients with or without HIV co-infection- A Survey", (IJCSIT) International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol. 6, issue 3, 2015, pp 2574-2578.

[18] Sourabh Shastri, Amardeep Sharma, VibhakarMansotra, Anand Sharma, Arun Singh Bhadwal, Monika Kumari, "A Study on Exponential Smoothing Method for Forecasting", International Journal of Computer Sciences and Engineering, ISSN: 2347-2693, Vol.6, issue 4, Apr 2018,.

[19] Sathishkumar R, Kavitha M, Aravinda hari, Shanmathi mv, "A Study: Predicting The Non-Epidemic Diseases Using Machine Learning Algorithms", International Journal of Pure and Applied Mathematics, ISSN: 1314-3395, volume 118, issue 22, 2018, pp 879-886.

[20] Ms. N. Vijayalakshmi, Ms. M. UmaMaheswari, "Data Mining To Elicit Predominant Factors Causing Infertility in Women", *IJCSMC, Vol. 5, Issue. 8, August 2016, pg.5 – 9.*

[21] World Health Organization, Global tuberculosis report 2019, ISBN 978-92-4-156571-4, 2019.

[22] Navneet Walia, Harsukpreet Singh, Shared Kumar Tiwari and Anurag Sharma, "A Decision Support System For Tuberculosis Diagnosability ", International Journal on Soft Computing (IJSC), Vol.6, No. 3, August 2015.

[23] Rakhmetulayeva S.B, Duisebekova K.S, Mamyrbekov A.M., Kozhamzharova D.K., Astaubayeva G.N., Stamkulova K. "Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis ", Elsevier-Procedia Computer Science, 130, 2018, pp231-238.

[24] M. Sebban, I.Mokrousov, N.Rastogi, C.Sola, "A Data-mining Approch to Spacer Oligonucleotide Typing of Mycobacterium Tuberculosis ", Oxford University, vol. 18, No.2, 2002, pp 235-243.

[25] T. W. Rennie, W. Roberts, "Data Mining of Tuberculosis Patient Data Using Multiple Correspondence Analysis", Cambridge University Press, 137,2009,pp 1699-1704.

[26] Brudey et al. , "Mycobacterium Tuberculosis Complex Genetic Diversity: Mining The Fourth International Spoligotyping Database (SpolDB4) for Classification, Population Genetics and Epidemiology ", BMC Microbiology, 6, 23, 2006, pp 1-17.

[27] K. R. Lakshmi, M. Veera Krishna, S. Prem Kumar, "Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability", I. J. Modern Education and Computer Science, 8,2013,pp 8-17.

[28] Amita Jain, Kamal Raj Pardasani, "Soft Fuzzy Model for Mining Amino Acid Association in Peptide Sequences of Mycobacterium Tuberculosis Complex", Current Science-Research Articles, 110, 4, 25 Feb 2016, pp 603-617.

[29] Orhan Er, FeyzullahTemurtas, A. Cetin Tanrikulu, " Tuberculosis Disease Diagnosis Using Artificial Neural Networks", Springer, 23 Dec 2008.

[30] Babu C Lakshmanan, Valamathi Srinivasan, ChinnaiyanPonnuraja, "Data Mining with Decision Tree to Evaluate the Pattern on Effectiveness of Treatment for Pulmonary Tuberculosis: A Clustering and Classification Techniques ", Scientific Research Journal (SCIRJ), ISSN 2201-2796, 3,6, June 2015, pp 43-48.

[31] Manish Shukla, Sonali Agarwal, " Hybrid Approach for Tuberculosis Data Classification Using Optimal Centroid Selection Based Clustering", IEEE, 2014.

[32] April Rose C. Semogan et al., " A Rule-Based Fuzzy Diagnostics Decision Support System for Tuberculosis", Ninth International Conference on software engineering research, management and Applications, 2011.

[33] K. Soundararajan, Dr. S. Sureshkumar, C. Anusuya, "Diagnostics Decision Support System for Tuberculosis using Fuzzy Logic", International Journal of Computer Science and Information *Technology* and Security(IJCSITS), ISSN: 2249-9555, vol. 2, issue 3, june 2012, pp684-689.

[34] Rahul Hooda, et al., " Deep-learning: A Potential Method for Tuberculosis Detection using Chest Radiography", Interrnational Conference on Signal and Image Processing Applications (ICSIPA), 2017, pp 497-502.

[35] Adnan Fojnica, Ahmed Osmanovic, AlmirBadnjevic, et al. "Dynamical Model of Tuberculosis-Multiple Strain Prediction based on Artificial Neural Network", MECO, 2016, pp 290-293.

[36] Ilena, Gabriella, Kamaraga, Stella A, Setiawan, Agung W, "Early Detection of Tuberculosis using Chest X-Ray (CXR) with Computer-Aided Diagnosis", International Conference on Biomedical Engineering (IBIOMED), 2018, pp 76-79.

[37] SeelwanSathitratanacheewin and KritPongpirul, "Deep Learning for Automated Classification of Tuberculosis-Related Chest X-Ray: Dataset Specificity Limits Diagnostic Performance Generalizability", ACM-class: I.2,6, 2018.

[38] Kamal J. AbuHassan et al, "Automatic Diagnosis of Tuberculosis Disease Based on Plasmonic ELISA and color-based Image Classification", IEEE Research Project – UK and Malaysia, 2017.

[39] Ferani E. Zulvia, R. J. Kuo, E. Roflin, " An Initial Screening Method for Tuberculosis Diseases Using a Multi-objective Gradient Evolution-based Support Vector Machine and C5.0 Decision Tree", IEEE-Computer society, Annual Computer Software and Applications Conference, 2017.

[40] Awad Ali, MoawiaElfaki, Dayang N.A. Jawawi, "Using Naïve Bayes and Bayesian Network for Prediction of Potential Problematic Cases in Tuberculosis", International Journal of Informatics and Communication Technology(IJ-ICT), ISSN: 2252-8776, vol 1, issue 2, December 2012, pp. 63-71.

[41] Asia Nesredin, "Mining Patients' Data for Effective Tuberculosis Diagnosis: The Case of Menelik II Hospital", Addis Ababa University, June 2012.