International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

Tuberculosis Prediction by Machine Learning Techniques

Kuldeep Godiyal[a], Surabhi Pokhriyal[b]
[a]Kuldeep Godiyal, Shivalik College of Engineering, Dehradun- 248197, India [b]Surabhi Pokhriyal, College of Pharmacy, Dehradun- 248197, India kuldeep.godiyal@sce.org.in

Abstract: Tuberculosis is one of the top reasons of death all over the planet. Mycobacterium tuberculosis, bacteria that infects the lungs, is what causes it. For professionals working in the medical field, accurately identifying and timely predicting tuberculosis are major challenges. The course of treatment also varies from patient to patient since occasionally a patient develops drug resistance. Doctors will be given algorithmic support while using machine learning to help them diagnose, treat patients appropriately, and make quicker and better judgments. This paper discusses the many tuberculosis causes and symptoms as well as how accurate and fast prediction and diagnostic investigations have been carried out in recent years with the aid of machine learning (ML) techniques

Keywords: Tuberculosis (TB), Machine Learning, Classification, Prediction techniques

1. Introduction

The bacteria Mycobacterium tuberculosis is the source of the airborne disease tuberculosis, by which the lungs of the patient are affected. The major symptoms of the disease include severe coughing, fever, and chest pains. The TB germs are spread when an infected person cough, sneezes, or spits. Even a slight exposure to these germs is enough to affect a person. Humans having weak immune systems, for example, people affected with diabetes, HIV, or malnutrition, or people who used to consume tobacco, have a very high-risk factor of getting affected.

According to WHO report Tuberculosis contributes to the top 10 causes of death around the world. It is also among the main causes of death associated with antimicrobial resistance and the prominent cause of demise for people infected by HIV disease. In 2017, there were approximately 1.3 million tolls (H.I.V. negative cases) and 300,000 deaths of patients who were HIV positive. In 2017, 10 million people were affected by TB, 5.8 million patients were male, 3.2 million female, and 1 million kids'.People surviving with HIV represent 9% of the overall Tuberculosis cases. India, China, Indonesia, Pakistan, Philippines, South Africa, and Nigeria account for two third of the new cases. Approximately 1.7 billion individuals, about 23% of the worldwide population, are likely to have a latent TB infection, and therefore are in danger of getting active TB infection during their lifetime [1].

Machine learning is an area of computer science that enables computer systems the capability to "learn" (i.e., gradually enhance one's performance on a definite task) with historical data, without being explicitly programmed. ML defines the learning and creation of algorithms that can study data and make predictions on it. ML is a technique used to formulate complex models and algorithms that provide insights for prediction; in commercial practice, this is famously well-known as predictive analytics. These analytics models permit data scientists, engineers, researchers, and analysts to "produce dependable, repeatable conclusions and results" and expose "hidden insights" by learning from trends in the data and historical patterns.

Healthcare is among the top industries that generate a huge amount of data. Machine learning lets programmers build models and rapidly analyze data and provide results, leveraging historical as well as real-time datasets. With the proper use of machine learning, healthcare facilities providers can make decisions based on diagnosing the patient and possibilities of treatment, which can lead to total enhancement of healthcare services. Vital statistics can also be derived from algorithms, real-time data, and cutting-edge analytics concerning the patient's disease, results of lab tests, blood pressure, clinical trial data, family history, and more healthcare practitioners. Machine learning can help in the prediction of the disease in earlier stages so that it can be prevented, rather than diagnosis and following the treatment course. [28].

2. About Tuberculosis

2.1. Symptoms

The most common warning sign of tuberculosis includes cough with sputum and blood sometimes, chest pains, weakness, fever, night sweats, and weight loss. Few individuals are more likely to be prone to the risk of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

generating TB than others, which include young adults, healthcare practitioners who help in the diagnosis and treatment process of TB-infected patients and work close to TB patients, and HIV-positive patients whose immunity system is low[2].

2.2. Diagnosis

Most countries use the sputum smear microscopy test for the diagnosis process where the sputum of the suspected patient is studied under a microscope for the existence of tuberculosis bacteria. This test cannot detect drug-resistant tuberculosis cases. Another test that detects tuberculosis and resistance to rifampicin at the same time which is the most essential tuberculosis medicine is the rapid test XPERT MTB/RIF. This test is now suggested by WHO as the initial test for the recognition of tuberculosis as it can provide results in approximately 2 hours. Multi-drug resistant Tuberculosis (MDR-TB) and HIV-TB co-infection cases are difficult to diagnose and the treatment process for such cases is very complex. Four new tests were suggested by the WHO which comprise a rapid molecular test and 3 tests to detect MDR-TB cases for first-line and second-line TB medicines [16].

2.3. Treatment

Tuberculosis is a treatable and curable disease and can be cured within 6 months if the patient obeys the treatment process. The four major drugs used for tuberculosis treatment include Isoniazid (INH), Rifampin (RIF), either ethambutol (EMB), Pyrazinamide (PZA), and streptomycin(SM) [2]. Depending on the process of diagnosis, the patient has been prescribed the drugs and a chart is prepared which outlines the duration and doses. For cases that are expansively drug-resistant an additional drug can be added to the routine i.e. either bedquiline or delamanid[16]. For countries where TB rates are still high, an additional procedure called the directly observed therapy short-term (DOTS) is used for treatment observance and treatment completion for greater efficiency and cost-effectiveness [2].

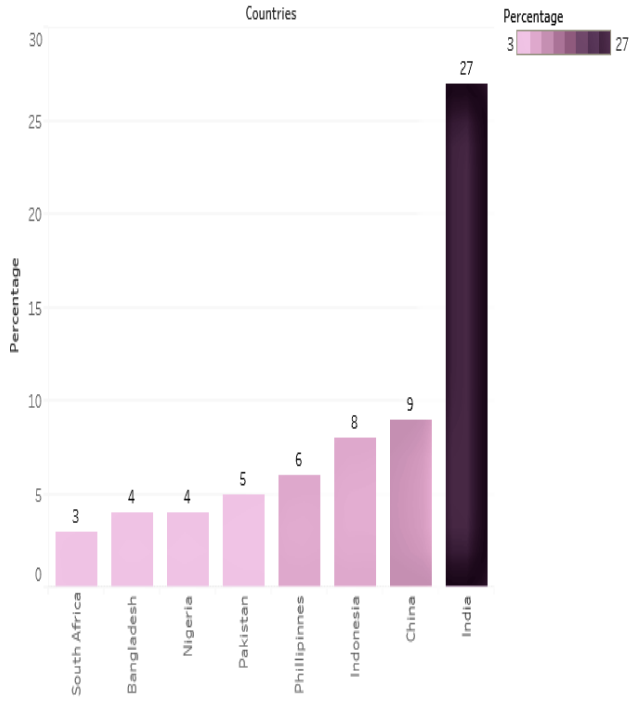

Figure 1: Eight Countries with the maximum number of TB cases worldwide, along with 22 other countries. India has the highest number of TB cases reported worldwide.

3. Approaches

3.1. Supervised ML Techniques

A supervised machine learning technique is applied to a known set of data to predict the outcome of new sets of data [28]. We start by analyzing a known training dataset; the algorithm learns and produces a conditional function to create predictions about the output values by recognizing patterns in data. The system can provide predictive outputs for any new input data after the model is trained appropriately. The Output of the learning algorithm model can be compared with the desired output to find errors to modify the model consequently. Supervised learning is of two types: Classification and Regression. Classification technique is utilized to classify the dataset into different categories based on their values e.g. in the case of the tuberculosis dataset the patient can be categorized as positive test(patients with TB) cases and negative test cases(patients who do not have TB) whereas regression technique is used for numerical values [29]. Examples of supervised ML techniques such as naïve Bayes, support vector machine, logistic regression, decision tree, k-nearest neighbor, and artificial neural networks [15].

3.2. Semi-Supervised ML Techniques

Unsupervised machine learning techniques are those that are used when the datasets that are used for training are neither

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

classified nor labeled in any way. The field of study known as unsupervised learning examines how computers may anticipate an unknown structure based on an unlabeled dataset. This algorithm may not be able to identify the correct solution, but it can examine the data and make inferences across datasets to explain patterns in data that have not been labeled. Clustering is a method of unsupervised machine learning in which the whole dataset is split up into subsets that have similar characteristics.

3.3. Unsupervised ML Techniques

This approach makes use of both labeled and unlabeled datasets to train the algorithm, which enables it to integrate aspects of supervised and unsupervised learning procedures. They use less volume of labeled data and a huge volume of the unlabeled dataset for training the algorithm, which in turn improves the overall accuracy of the system [28]. Typically, semi-supervised techniques are selected when the collected labeled dataset needs trained and appropriate means to train it and learn from it. Otherwise, obtaining unlabeled data generally doesn't need additional resources.

3.4. Reinforcement ML Techniques

It is a learning technique that works together with its environment by generating rewards and errors. A method to reward preferred behavior and punish undesired behavior is recognized. [28] [29]. The desired outcomes are allotted positive values and negative values are allotted to undesired outcomes. The reinforcement learning technique permits the programmer to spontaneously decide the ideal behavior for a particular case to maximize its performance. The simple reward feedback is used by the agent to study which action is correct; this is termed the reinforcement signal [29].

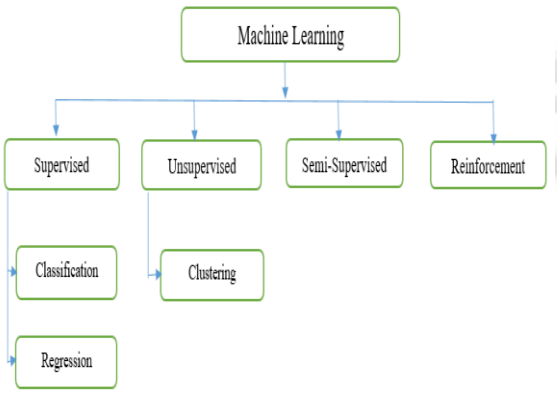

Figure 2: Type of Machine learning technique.

4. Algorithms

4.1. Artificial Neural Networks

Our brains or biological neural networks, which are capable of simultaneously resolving massive amounts of data, are the inspiration for ANN. Through synapses between the dendrites and axons, the signals are transmitted to the neuron's axon terminals. When a certain threshold level is obtained, the neuron gets activated and the signal is transmitted to another neuron through the axon terminal [4, 27].

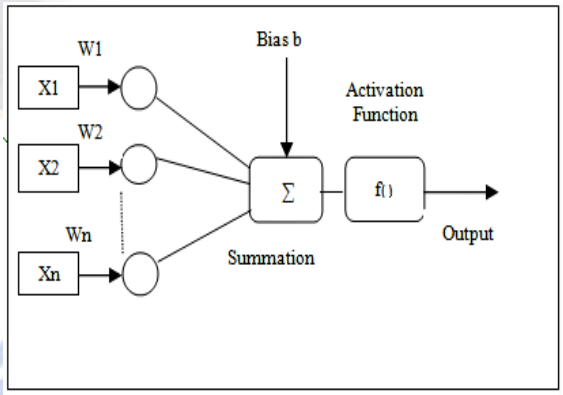

Figure 3: Artificial Neural Networks.

In the diagram X_1, X_2….X_n expresses the input values used and W_1, W_2…W_n expresses the weights associated with the input values and a bias b. The output is represented in binary values 0 or 1 based on the values of $\sum w_n x_n$ being lower than or larger than the threshold value.
There are two cases:
When $\sum w_n x_n$ < threshold then the output is 0.
When $\sum w_n x_n \geq$ threshold then the output is 1.

There are two forms of artificial neural networks: feed-forward neural networks and feedback neural networks, both of which can have one or more layers. In multilayer ANN, there is one more layer called the hidden layer, which receives input from the input layer, and the weighted sum of the layer is transferred to the output layer [18].

4.2. Decision Tree

A decision tree is a structure resembling a tree in which each step or conclusion is represented by a branch. It is used for the classification of data points from the root to the leaf node, where the leaf node represents the outcome or result [3, 18]. The root node is chosen in such a manner that it splits the data into the most effective way and the numbers of steps are minimum to get a possible outcome. A greedy algorithm is the most basic algorithm which is used for the construction of a decision tree in a top-down, divide-and-

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

conquer manner [15]. The accuracy of the forecast is contingent on several different elements. Decision trees are used often in a variety of contexts, including debt collection, automobile insurance, and corporate fraud, to name a few. To construct the yield mapping and anticipate production, it made use of agronomic factors, nitrogen treatment, and weed management via the use of machine learning algorithms such as artificial neural networks and decision trees.

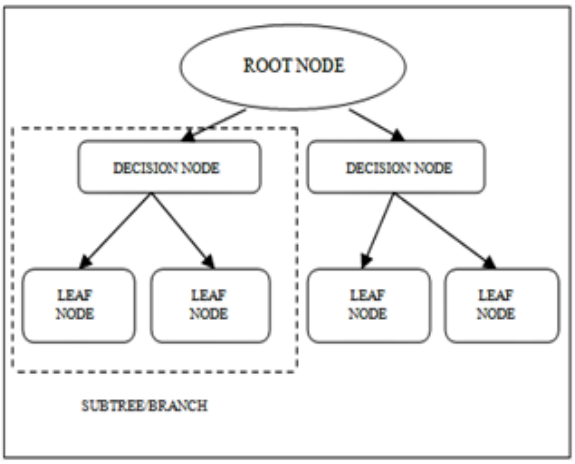

Figure 4: Decision tree.

4.3. Support Vector Machine

A separation hyperplane between the two sets of labeled data is discovered by a support vector machine. The hyperplane's location is found by maximizing the distance between two bordering data points from two labeled classes, which are called support vectors and the hyperplane is called the maximum marginal hyperplane (MMH) [3].

Linear and non-linear datasets both can be classified with a support vector machine, known as the Maximum Margin Classification algorithm [18] [30].

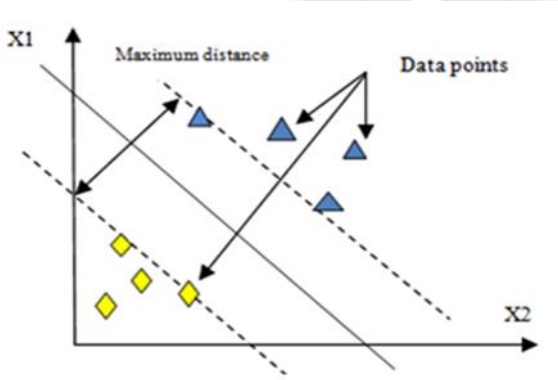

Figure 5: Support vector machine.

4.4. Naïve Bayes

The Bayes theorem, which is used to solve classification problems, is the foundation of the Naive Bayes algorithm. This algorithm is based on the theory that the occurrence of a particular event is independent of the occurrence of another event [4].

$$P(A|B) = [P(B|A)P(A)]/P(B) \quad (1)$$

Where

P (A|B) = probability of occurrence of event A given B

P (B|A) = probability of occurrence of B given A

P (A) and P (B) = probability of A and probability of B.

5. Related work

Different scholars have contributed by studying several different machine learning algorithms like support vector machines, Naïve Bayes algorithms, decision trees, artificial neural networks, and so on, for different chronic diseases like cancer, diabetes, heart diseases, and many more. Timely diagnosis and prediction can be very helpful in the medical field to make better decisions.

In [3], the author evaluated the outcome of 6 machine learning models on a dataset collected from a medical institution in Iran in the year 2005. The data included 35 different parameters and the initial dataset had 6450 different cases out of which 4515 were used for the determination of training and the rest 1935 for testing the models. The decision tree gives the best result for both training and testing datasets as it works well with both continuous and discrete/categorical predictors which are better than other techniques that are decent at handling only continuous variables.

In [5], the main problem discussed is developing bacterial resistance levels to standard antibiotics. The most important 4 standard first-line antibiotics medicines were studied: INH isoniazid (INH), pyrazinamide (PZA), rifampicin (RIF), and ethambutol (EMB).3.6% of all the active TB cases throughout the world are resistant to the frequently used antibiotics rifampicin (RIF) and isoniazid (INH) defined as "Multi-drug resistant Mycobacterium tuberculosis (MDR-MTB)"; the cases nearly doubled in 2011-2012. Though MTB and MDR-MTB can be cured, it is quite difficult to evaluate resistance in a well-timed custom. Earlier used methods like phenotyping by the proportion method, utilizing Lowenstein-Jensen (LJ) solid media needed longer duration and Nucleic acid-based tests provided better results as compared to slow phenotypic methods. A single test can

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

be used by whole genome sequencing (WGS) to predict all unique nucleotide polymorphisms (SNPs) in a given sample. The study was done to determine whether MTB separates were susceptible to or resistant to four common first-line medications. They also conducted a study where misclassified isolates were examined more carefully to provide evidence in the direction of promising mutations for further study.

In [7], 3 models used were the "Adaptive Neuro-Fuzzy Inference System (ANFIS)", partial decision tree, and multilayer perceptron. The dataset used to have 667 patient records each having 30 input values. Using the attribute ranking function, which was implemented with the aid of the information gain ranking filter, the dataset is reduced. This process divided the data set's input variables into 20. The reduced data set is applied to the 3 models and Sensitivity, Specificity, Precision, and Correctness values are calculated. The best performance overall is given by ANFIS.

In [8], the author used the "Artificial immune recognition system (AIRS)" with fuzzy sets for the diagnosis process of tuberculosis. 175 samples with 20 attributes each were included in the collection, which was obtained from the Pasteur Laboratory in northern Iran. The dataset was split in a 7:3 ratio, with 30% of the dataset being utilized for testing purposes and 70% being used for training. The accuracy obtained with this study was 99.14%, with sensitivity and specificity of 87.00%, and 86.12%.

In [10], the author employed an artificial neuro-fuzzy inference system (ANFIS) and k-means clustering for the diagnosis process of TB. MATLAB Fuzzy Logic Toolbox (FLT) was used as the development tool. With the help of fuzzy rules, a dataset was classified into 4 groups: Multiple Drug Resistant tuberculosis (MDR-TB), Extremely Drug Resistant (XXDR-TB), Extensive Drug Resistant (XDR-TB), and HIV-TB.

In [12], Using data from whole genome sequencing, the author applied deep learning to the prediction of multi-drug resistant tuberculosis. The study was conducted for 10 Tuberculosis drugs, classified into first and second-line drugs using 3601 Mycobacterium tuberculosis (MTB) isolates in total, out of which 1228 isolates were multidrug-resistant cases. With average sensitivities and specificities of 92.7% and 92.7% for first-line meds and 82.0% and 92.8%, respectively, for second-line treatments, the Wide and Deep Neural Network (WDNN) surpassed logistic regression and random forest in terms of prediction accuracy.

In [13], a support vector machine is used by the author for the prediction of the disease. The dataset had 150 samples in total, comprising 50 patients' records with tuberculosis and 100 patient's record without tuberculosis. All have 38 parameters and the dataset is divided into two type's i.e. Training and testing datasets. The accuracy achieved with this study was 96.68%.

In [17], to diagnose TB, the author employed a system known as genetic neuro-fuzzy inference. The dataset included 10 patients' medical records from St. Francis Catholic Hospital Okpara-In-Land in Delta State, Nigeria, and it contained 24 parameters. Out of 24 parameters, 13 parameters were used in the evaluation process. The sensitivity and accuracy values obtained were 60% and 70% respectively.

In [22] and [23], the author aimed to develop a real-world, mobile commuting application for the prediction of tuberculosis using an X-ray database. The first phase was to classify the image into two groups: Normal and Abnormal (with TB manifestation). In the second phase, the Images were classified into multiple classes with different types of TB manifestations: cavitation, lymphadenopathy, infiltration, and pleural infusion. The dataset had 4701 X-ray images in total, collected from Dr. Peinado. The study delivered good results in both phases over several iterations.

In [24], Artificial Neural networks were used for prediction. The dataset was obtained from patient epicrisis reports obtained from Diyarbakir chest disease hospital in the south of Turkey. The dataset contains 150 samples with 38 features. The study obtained better results; compared to previous studies conducted using ANN.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

Table 1: Machine Learning Techniques: A Comprehensive view of tuberculosis disease prediction.

Machine Learning Techniques	Author	Year	Sources of Dataset	Accuracy
DT ANN BN LR RBF SVM	Sharareh R et al.	2013	Dataset collected from medical Institute in Iran.	74.21% 57.82% 61.70% 57.82% 53.74% 57.47%
Adaptive Neuro-Fuzzy Inference System(ANFIS) DT Multilayer Perceptron	Tamer Ucar et al.	2011	667 Patient Records from a private clinic	97% 89% 85%
Artificial Immune Recognition System	Shahaboddin Shamshirband et al.	2014	Reports gathered from the Pasteur Laboratory situated in northern Iran regarding Patient epicrisis.	99.14%
SVM	Amani Yahiaoui et al.	2017	Department of Chest diseases of a hospital in Diyarbakir.	96.68%

6. Conclusion

As healthcare is one of the top businesses that generate a great deal of data, the utilization of machine learning techniques to make judgments on treatment options after patient diagnoses may greatly enhance healthcare services as a whole. Vital statistics can also be derived from machine learning algorithms.

Tuberculosis is among the Top reasons for death worldwide and is difficult to cure in most cases; the patient can develop drug resistance which causes further complexity in the treatment process. Many of the cases are from countries with poor resources and developing countries with weaker healthcare facilities. This paper focused on different prediction techniques using healthcare data.

References

[1] Global Tuberculosis Report 2018.

[2] Fogel, Nicole. "Tuberculosis: a disease without boundaries." *Tuberculosis* 95.5 (2015): 527-531.

[3] Kalhori, Sharareh R. Niakan, and Xiao-Jun Zeng. "Evaluation and comparison of different machine learning methods to predict the outcome of the tuberculosis treatment course." *Journal of Intelligent Learning Systems and Applications* 5.03 (2013): 184

[4] Dande, Payal, and Purva Samant. "Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review." *Tuberculosis* (2017).

[5] Niehaus, Katherine E., et al. "Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis." *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*. IEEE, 2014.

[6] Doshi, Riddhi, et al. "Tuberculosis control, and the where and why of artificial intelligence." *ERJ open research* 3.2 (2017): 00056-2017.

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

[8] Shamshirband, Shahaboddin, et al. "Tuberculosis disease diagnosis using artificial immune recognition system." *International journal of medical sciences* 11.5 (2014): 508.

[9] Yang, Yang, et al. "Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data." *Bioinformatics* 34.10 (2017): 1666-1671

[10] Ashadevi, B., P. Muthamil Selvi, and B. Sasi Revathi. "An Effective Diagnosis of Pulmonary Tuberculosis using K-Means Clustering and ANFIS."(2017)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5797
***Article Received**: 14 August 2022 **Revised**: 18 October 2022 **Accepted**: 12 November 2022*

[11] Ashwini, D. V., and S. Seema. "Machine Learning Approach to Detect Tuberculosis in patients with or without HIV co-infection-A Survey."(2015)

[12] Chen, Michael L., et al. "Deep Learning Predicts Tuberculosis Drug Resistance Status from Whole-Genome Sequencing Data." *bioRxiv* (2018): 275628.

[13] Yahiaoui, Amani, Orhan Er, and Nejat Yumusak. "A new method of automatic recognition for tuberculosis disease diagnosis using support vector machines." *Biomedical Research* 28.9 (2017).

[14] Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems* 14.1 (2008): 1-37.

[15] Kesavaraj, Gopalan, and Sreekumar Sukumaran. "A study on classification techniques in data mining." *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*. IEEE, 2013.

[16] World Health Organization. (2018) Tuberculosis: Fact sheet. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/tuberculosis

[17] Omisore, Mumini Olatunji, Oluwarotimi Williams Samuel, and Edafe John Atajeromavwo. "A Genetic-Neuro-Fuzzy inferential model for diagnosis of tuberculosis." *Applied Computing and Informatics* 13.1 (2017): 27-37.

[18] Godara, Sunila, and Rishipal Singh. "Evaluation of predictive machine learning techniques as expert systems in medical diagnosis." *Indian Journal of Science and Technology* 9.10 (2016).

[19] Amato, Filippo, et al. "Artificial neural networks in medical diagnosis." (2013): 47-58.

[20] Lakhani, Paras, and Baskaran Sundaram. "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks." *Radiology* 284.2 (2017): 574-582.

[21] Chen, Min, et al. "Disease prediction by machine learning over big data from healthcare communities." *IEEE Access* 5 (2017): 8869-8879.

[22] Cao, Yu, et al. "Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities." *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016.

[23] Alcantara, Marlon F., et al. "Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor communities in Peru." *Smart Health* 1 (2017): 66-76.

[24] Er, Orhan, Feyzullah Temurtas, and A. Çetin Tanrıkulu. "Tuberculosis disease diagnosis using artificial neural networks." *Journal of medical systems* 34.3 (2010): 299-302.

[25] Elveren, Erhan, and Nejat Yumuşak. "Tuberculosis disease diagnosis using an artificial neural network trained with genetic algorithm." *Journal of medical systems* 35.3 (2011): 329-332.

[26] Maji, Srabanti, and Srishti Arora. "Decision Tree Algorithms for Prediction of Heart Disease." *Information and Communication Technology for Competitive Strategies*. Springer, Singapore, 2019. 447-454. (2018)

[27] Zakaria, Magdi, AL-Shebany Mabrouka, and Shahenda Sarhan. "Artificial neural network: a brief overview." *neural networks* 1: 2. (2014)

[28] Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic." *Journal of Intelligent Learning Systems and Applications* 9.01 (2017): 1.

[29] Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." *Computational and structural biotechnology journal* 15 (2017): 104-116.

[30] Jain, Divya, and Vijendra Singh. "Feature selection and classification systems for chronic disease prediction: A review." *Egyptian Informatics Journal* (2018).