Prediction of Multi Drug Resistant Tuberculosis using Machine Learning Techniques

R Lokeshkumar, Jothi K R, Anto S, R Kiran kumar, Hari Narayanan

Abstract: Mycobacterium Tuberculosis bacteria is the primary cause for Tuberculosis. TB is one of the main reasons of mortality around the world. Multi Drug Resistant Tuberculosis (MDR-TB) is a type of tuberculosis bacteria which are resistant to anti-TB drugs, drugs like isoniazid (INH) and rifampin (RMP). Different Machine learning approaches has been widely applied to predict MDR TB. Here, we review different Machine Learning Approaches to predict MDR-TB. Different feature estimation methods, execution of distinct machine learning models also have been explored. Additionally, the utilization of the distinctive machine learning system models for distinguishing the dis-functionalities of MDR-TB in the recent decades has been talked about.

Keywords: MDR-TB, Machine Learning, Genome Sequencing, isoniazid (INH) and rifampin (RMP).

I. INTRODUCTION

M Multi Drug Resistant Tuberculosis is caused if a person is resistant to most of the main Tuberculosis drugs like isoniazid(INH) and rifampicin(RMP). It indicates that if the drug is resistant to the body then it won't cure the disease. So other drugs needed to be taken. Mainly there are two ways a person can be affected with MDR-TB. Firstly, if a person don't take drugs according to the prescribed dosage by doctor. This is because the bacteria are resistant the drugs. Another way a person can get MDR-TB if the person get affected by TB Bacteria from a person who is already having MDR-TB. Other factors of MDR TB are HIV infection, TB treatment failure to a prescribed process of dosage and defaulted patients that lapse standard TB treatment. According to WHO, in 2016 around 4,90,000 people in worldwide is affected with MDR-TB also another 1,10,000

people resistant to rifampicin-resistant TB which are eligible

for MDR-TB treatment. Largest number of MDR-TB cases were reported in China, India and Russian Federation.

MDR TB can be treated using second line drugs for minimum of 6 months. If the patient is resistant to rifampin duration can extend upto 18-24 months. MDR TB is treated on the report of sensitivity testing. The suspected MDR-TB who are waiting for sensitivity testing are given SHREZ drugs that is, Streptomycin, isonicotinyl Hydrazine, Rifampicin, Ethambutol and pyrazinamide and moxifloxacin with cycloserine. WHO recommended drugs are the combination of four drugs to which M.Tuberculosis bacteria is susceptible. The drugs are given in an ordered process of five stage groups. The first line drugs are isoniazid, pyrazinamide, and ethambutol. The second group of drugs are fluoroquinolones and high-dose levofloxacin. The third group are injectable drugs which follows an order: capreomycin, kanamycin, and then amikacin. The forth group drugs are called as second line drugs are given in a sequence: thioamides, cycloserine, then aminosalicylic acid. The fifth group drugs include clofazimine, amoxicillin with clavulanate, linezolid, carbapenems, thioacetazone, then clarithromycin.

Nowadays various Machine Learning approaches are performed in this field to find the MDR-TB at its early stages so that it can be cured fast. Machine Learning helps to create a model which is trained with the training dataset containing various attributes. The accuracy of the dataset is found out using the testing dataset. Various machine learning algorithms are being used nowadays. Various machine learning approaches are being performed on the datasets.

There are mainly three types of machine learning algorithms which are Supervised, Unsupervised and Reinforcement Learning Algorithms. Nowadays many deep learning Algorithms such as CNN, RNN etc. are also being widely used to analyze MDR-TB.

A. Supervised Learning

The main objective of supervised learning is to predict one or more output variable y from a given set of input variables x. Output of supervised learning can be of mainly two types either continuous (Regression Problem) or categorical (Classification problem). In Supervised Learning the training dataset contains a set of N attributes as input variables and one or more output variable. Different Supervised Models form a function y(x) which helps to predict the output for a new input value. Various Supervised Algorithms are Decision Tree, SVM, Linear Regression, Logistic Regression, kNN etc.

B. Unsupervised Learning

In Unsupervised Learning training dataset contains only a set of variables x. One of the most common way of solving unsupervised problems is by using clustering. Clustering groups data in a way that intra-cluster distance is high and inter-cluster distance is low. Some of the unsupervised Algorithms used are k-Means, Gaussian Mixture Model, DBSCAN etc.

C. Reinforcement Learning

Reinforcement Learning (RL) allows model to learn from available variables and tunes itself by evaluating the feedback. RL is used in robotics, inventory management, finance etc. RL is used to find the best possible path in a given condition. In RL the RL agent decides what process should be performed to the given situation. If there is no training dataset, it learns from its experience.

Here, the paper performs some Machine Learning algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Artificial Neural network. Also, we are performing some Ensemble model in Bagging, Boosting and Voting. Various Bagging methods performed are Bagged Decision Trees and Extra Trees. Under Boosting we use AdaBoost and Stochastic Gradient Boosting. Voting Ensemble is used under voting method.

II. RELATED WORK

From Table 1, [1] proposes a hybrid model of Decision tree and CNN. There they use a dataset consisting of CT Scan images, age and gender as the training dataset. Manual distinction between DS and MDR tuberculosis using only CT Scans is not possible. Neural networks, which do not require feature extraction. Convolutional neural network can be used, which only considers the images, on the other hand an architecture can be developed. Creating a feature vector of the CNN's result is one classifier, decision tree has been chosen as the second classifier. The dataset is fed to the model and Martha and Stephan acquired an accuracy of 58.1%. The quality of the classifiers can certainly be increased by improving the preprocessing of the images with the help of medical expertise. Another approach by Michel et al.[2] perform genome sequencing and they uses a deep learning model Multitask wide and deep Neural network(MD-WDNN). Both first line and second line drugs were tested and achieved an accuracy of 97.9% and 93.6% respectively under repeated validation.

Tuberculosis is one of the major issues regarding public health.

Pathology is one important clinical practice for diagnosing TB. Under a microscope to find and confirm TB bacilli is very critical. Because number of bacilli are more and they are small in size, Experienced pathologists finds it as time-consuming and strenuous work, which often leads to low detection rate and false diagnoses. They have investigated the clinical efficacy for acid-fast stained TB bacillus using artificial intelligence (AI)-assisted detection method (TB-AI).

An AI based AI-TB model was proposed by Yan Xiong et al.[3] using CNN to predict TB bacilli. Under a microscope to find and confirm TB bacilli is very critical. Because number of bacilli are more and they are small in size, Experienced pathologists finds it as time-consuming and strenuous work, which often leads to low detection rate and false diagnoses. TB-AI achieved 97.94% sensitivity and 83.65% specificity and it was also diagnosed with microscope which TB-AI is more accuracy in detecting stained TB bacilli. [4]A DNA sequencing data from MTB which contains 1839 bacterial which classifies MTB against 8 anti-TB drugs was used. The model gave -4% increase for isoniazid, rifampicin and ethambutol to 97%. For ciprofloxacin and MDR-TB increased to 96%. Sensitivity of Moxifloxacin and ofloxacin increased to 83% and 81%. For pyrazinamide and streptomycin sensitivity increased to 84% and 87%. The genome sequences including positions outside 23 genes and deep networks for non-linear classification and dimension reduction and also optimizing the number of SPCA/SNMF components can be considered as the future work. Dataset contains 13402 isolates. Pyrazinamide, ciprofloxacin and ofloxacin increased accuracy to 23.11%,15.22% and 10.14%. [5]Logistic Regression and gradient boosting performed better for F1 score by 12.54%, 4.61%, 7.45% and 9.58% for amikacin, moxifloxacin, ofloxacin and capreomycin, respectively. In future deep learning and ensemble methods can be applied to the dataset to increase the accuracy. The number of instances in training and testing dataset contains 4515 and 1935 respectively and has 17 attributes. Various ML algorithms were applied and found that DT(C4.5) was the best algorithm with 74.21% accuracy[6]. A method to uniquely deconvolve out an association of a region with a specific drug. Also in future we can provide a support to the casualty of allele. SVM algorithm applied on AMR-conferred genes. It revealed seven known AMR gene-antibiotic relations absent among 40 alleles determined by pairwise association. Also another four main attributes of the model enabled genetic interaction underlying variable AMR phenotypes[7]. The NIAID [8] supports a global TB research agenda through its extramural and intramural programs. The NIAID extramural portfolio of grants and contracts supports all aspects of TB research, ranging from studying the basic biology of Mycobacterium tuberculosis and its interaction with the host to investigating the various manifestations of TB (pulmonary, extrapulmonary, and latent) in adult and pediatric populations, including HIV-coinfected individuals, to conducting research aimed at developing new health care interventions. [9] Total of 66 MDR TB patients started treatment, from the study population and among them 20 (30%) were resistant to one or more second line drugs including a case of "XDR TB". For treatment only half of the patients

TITLE	FINDING	FUTURE SCOPE	Publishing Year	Journal/Publisher
Detection of Multidrug-Resistant Tuberculosis Using Convolutional Neural Networks and Decision Trees[1]	The model used for CTScan,Age and Gender dataset are Convolution Neural Network ,Decision Tree and hybrid of both.The accuracy achieved by the model is 0.5810	The models quality can be increased by improving preprocessing of the images with help of medical expertise. Also more features about the dataset can increase the accuracy.	2018	CLEF
Deep learning predicts tuberculosis drug resistance status from genome sequencing data[2]	A Multitask wide and deep neural network(MD-WDNN) model built which has 0.979 accuracy for first line drugs and 0.936 for second-line drugs using repeated cross validation.	Although the phenotype strains used under strict quality control,some resistance mutations produce minimum inhibitory concentrations(MICs) produce low sensitivity and specitivity.Also, third line drugs are not considered which can be used in future.	2018	bioRxiv
Automatic detection of mycobacterium tuberculosis using artificial intelligence[3]	An AI based model is created using CNN to predict TB bacilli where the datset has 45 samples with 30 positive and 15 negative cases. TB-AI model achieved 97.94% sensitivity and 83.65% specificity	More accurate measures can be applied to ncrease the specificity of the model. Different ML models can be tried to get more accuracy.	2018	Juornal of Thoracic Disease
Machine Learning for Classifying Tuberculosis Drug-Resistance from DNA Sequencing Data[4]	A DNA sequencing data from MTB which contains 1839 bacterial which classifies MTB against 8 anti-TB drugs was used. The model gave -4% increase for isonazid,rifampicin and ethambutol to 97%. For ciprofloxacin and MDR-TB increased to 96%.Sensitivity of Moxifloxacin and ofloxacin increased to 83% and 81%.For pyrazinamide and streptomy sensitivityincreased to 84% and 87%.	Specificity of the model can be increased by improving more accuratephenotypic method to test drug resistance. Some drugs out performed others due to dispropotionate numbers, so more datas can be added to increase the efficiency of model.	2018	Bioinformatics
Application of machine learning techniques to tuberculosis drug resistance analysis[5]	Dataset contains 13402 isolates. pyrazinamide, ciprofloxacin and ofloxacin increased accuracy to 23.11%,15.22% and 10.14%.Logistic Regression and Gradient boosting performed better for F1 score by 12.54%, 4.61%, 7.45% and 9.58% for amikacin, moxifloxacin, ofloxacin and capreomycin, respectively.	The genome sequences including positions outside 23 genes and deep networks for non-linear classification and dimension reduction and also optimising the number of SPCA/SNMF components can be considered as the future work.	2018	Bioinformatics
Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course[6]	The number of instances in training and testing dataset contain 4515 and 1935 respectively and has 17 attributes. Various ML algorithms were applied and found that DT(C4.5) was the best algorithm with 74.21% accuracy.	In futuredeep learning and ensemble methods can be applied to the datasetto increase the accuracy.	2013	Scientific Scholar
Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance[7]	SVM algorithm applied on AMR-conferred genes. It revealed seven known AMR gene-antibiotic relations absent among 40 alleles determined by pairwise association. Also another fourmain attributes of the model enabled genetic interaction underlying variable AMR phenotypes.	A method to uniquely deconvolve out an association of a region with a specific drug. Also in future we can provide a support to the casuality of allele.	2018	Nature Communications

Table 1.Comparision of different Approaches for MDR-TB prediction

has stayed for 10 days. The treatment was provided partially under supervision. A major problem is providing injection. Response to treatment could be correctly predicted based on the 6-month results. Treatment outcome was observed only in 37% and a default of 24%. For Adverse reactions, modification of treatment was required only for three patients. Emily et. Al[10] performed a project on MDR-TB treatment and prevention. Ordinary differential equations are used to develop a compartmental model of an adult pulmonary TB epidemic involving a drug-susceptible and a drug-resistant M. tuberculosis strain. Sensitivity analysis is used to reduce relative reductions under primary intervention of MDR screening and Treatment. Partial rank correlation coefficients are computed for effects of other parameters.

III. METHODOLOGY

Previously, there are many approaches to predict MDR-TB using different approaches. Most of the approaches use genome sequencing and some deep learning algorithms like CNN are also being tired for prediction. Here in this paper tries to predict MDR-TB from the dataset which has attributes like Age, Sex, Alcoholic, Tobacco User, HIV Patient, and other attributes that whether the patient taking First line Drugs like Isoniazid, Rifampicin, Pyrazinamide and Ethambutol. Other attributes are whether the patient taking Second Line Drugs which include Group A, Group B, Group C. The outcome of the dataset has four class labels which are Died, Defaulted, Treatment Completed and Cured. The class label Died indicate that whether the patient died causing MDR-TB or not. Defaulted label means the patient has stopped taking the drug prescribed by the Doctor due to some reasons. Treatment Completed indicates that the patient has completed the course of dosage but still the patient is affected with MDR-TB. At last the label Cured denotes the patient has recovered from MDR-TB. Here the paper proposes some machine learning and Ensemble approaches to the dataset. The following algorithms are used for analysis on MDR-TB dataset.

A. Logistic Regression

Logistic model uses logistic function to model binary dependent variable. Parameters of a logistic model is been estimated in regression analysis, using logistic model. It can have binary values like win/lose, pass/fail, true/false; which are indicated using values "0" and "1". Multinomial Logistic Regression can have 3 or more categories without ordering. Multinomial Logistic Regression are mostly used when there are more than three dependent variables are used for prediction. Another type of logistic regression is Ordinal Logistic Regression. Here, there are three or more categories with ordering. Decision Boundary is used to predict which class a data belongs where a threshold is set. Considering the threshold, the estimated probability can be classified into classes. Another type is Decision boundary, where it can be linear or non-linear. By increasing the polynomial order complex decision boundary can be determined. Using Logistic Regression, the Outcome variable is being predicted using other dependent variables.

B. Decision Tree

Decision Tree is used in many fields of machine learning, includes both regression and classification. While performing analysis, a decision tree is used to visually and explicitly represent decision making and decisions. It uses tree model for decision making. Decision trees are made using algorithmic approach which finds ways to split data set using different conditions. It's a predictive modelling tool which covers different areas. It uses supervised learning approach. Decision Trees are non-parametric supervised learning procedure which is used for both regression and classification. Aim of Decision Tree is to create a prototype that assumes the value of a target by training decision rules acquired from the data attributes. Decision rules are mostly in if-else form. As the tree gets deeper, the rules get more complex and model fits.

C. Random Forest

Random Forest is a supervised learning algorithm. Random Forest is used for classification, regression and many other functions by creating a multitude decision tree and the final tree would be mode or mean prediction of individual trees. Random Forest corrects decision trees overfitting to the training set. Using Random Forest, using Random Forest regress or regression tasks can be dealt. It searches for the best attribute among the random subsets instead of searching most important feature. The results would be in very wide diversity and generally the results would be a better model.

D. Support Vector Machine (SVM)

In Support Vector Machine each data item plotted as a point in n-dimensional space. Here each attribute represents value of distinct coordinate. Classification is performed using hyper-plane that separates two classes. Support vectors are coordinates of each observation. SVM is a frontier uses to segregate two classes i.e., hyper-plane and line.

E. K-Nearest Neighbor (KNN)

Classification and regression problems can be handled by KNN algorithm. It is very simple and easy to create and process quickly but it misses shorter paths sometimes due to its greedy nature, which can be found easily by human insight. If last some stages of traversal are comparable in length to the starting stages then traversal is reasonable; if it is larger, then there would be better paths. Another method is to use an algorithm to estimate if path is good or not.

F. Artificial Neural Networks (ANN)

To model extremely complex non-linear functions, we use Artificial Neural Network (ANN). ANN is biologically inspired analytical method. Here, a common architecture named multi-layer perceptron (MLP) with learning by back-propagation algorithm is built. A neural network is a compound of input or output links which has an associated weight. The core phase for predicting the correct class label of input through iterative learning is

adjusting the weights. In classification and prediction tasks, ANN is widely used with high tolerance to noise and the ability to classify unseen patterns.

G. Bagged Decision Tree

Bootstrap Aggregation or Bagging divides the training dataset to multiple subsets. Each sub dataset is applied with decision tree. Bagging reduce the variance and overfitting problem. Bagging can be performed using BaggingClassifer module in sklearn ensemble.

H. Extra Trees

Extra trees are also another type of Bagging model. Here random trees are made from the training sample dataset. The model allows to provide number of tree sets and random state. Extratree Classifier module can be accessed from sklearn.ensemble package.

I. AdaBoost

Boosting performs a sequence of ML models which tries to correct the issues faced by the previous model. The output model predicts according to the accuracy of each sub model. AdaBoost performs weighted instances in dataset.

J. Stochastic Gradient Boosting

One of the most complex Boosting algorithm is Stochastic Gradient Boosting. Gradient Boosting Classifier can imported from sklearn.ensemble package.

K. Voting Ensemble

Voting Ensemble is a simple Ensemble algorithm where multiple machine learning algorithms are used to predict the output. It creates two or more independent from training set. Voting Classifier combines the model and averages the prediction. Voting Classifier can be used from ensemble in sklearn package.

IV. IMPLEMENTATION AND RESULTS

The MDR-TB dataset was performed on the algorithms mentioned above. The dataset was imported and performed preprocessing and data cleaning operations. In data cleaning all the null values and outliers are removed. Also some unimportant attributes are removed. Null values can be occurred by different reasons like user not interested to disclose or while recording mistakes. Also outliers can happen when inputing data values which are out of bound. All these types of data are removed as these values can adversely affect the model.

The preprocessed data is fed to different algorithms that mentioned earlier. The attributes relation in dataset is plotted using heatmap in Figure 1. The dataset is fed to Logisitic Regression (LR) which gave accuracy of 49%. The confusion matrix of LR is in Figure 2.

Decision Tree gave an accuracy of 40 percent with maximum depth upto three and random state was set to 0. The confusion matrix heatmap is given Figure 3.

Random Forest Accuracy on training set is 70%, accuracy on test set is 50%. The maximum depth kept for random forest is three and random state is set to 0. Confusion Matrix for RF is depicted in Figure 4.

For SVM the accuracy on training set is 89%,and the accuracy on test set is 34% .Neural Network Accuracy on training set is 100% and accuracy on test set is 34% and the heatmap obtained is given in figure 5.

KNN algorithm gave an accuracy of 34% for testing. The accuracy for both training and testing set are plotted and shown in Figure 6.

Other than normal ML Algorithms different ensemble methods like Bagging, Boosting and Voting was performed on MDR-TB dataset. Bagging Decision Tree gave an accuracy of 40.4%. Another Bagging algorithm, ExtraTree gave 43.3% accuracy. The Confusion Matrix of ExtraTree is shown in Figure 7.

Boosting algorithms performed in dataset are AdaBoost and Gradient Descent. The accuracy obtained by AdaBoost was the highest. It gave 56.6% accuracy. The confusion matrix of AdaBoost is given in Figure 8. GradientBoost gave an accuracy of 45%.

Another simple ensemble method , VotingClassifier was performed which took 3 other algorithms Logistic Regression, SVM and Decision Tree Classifier. It gave an accuracy of 41.6%.
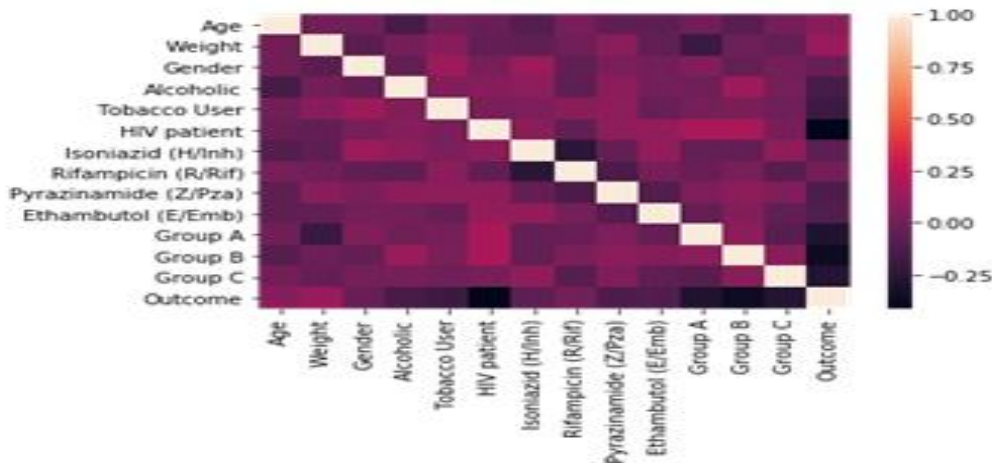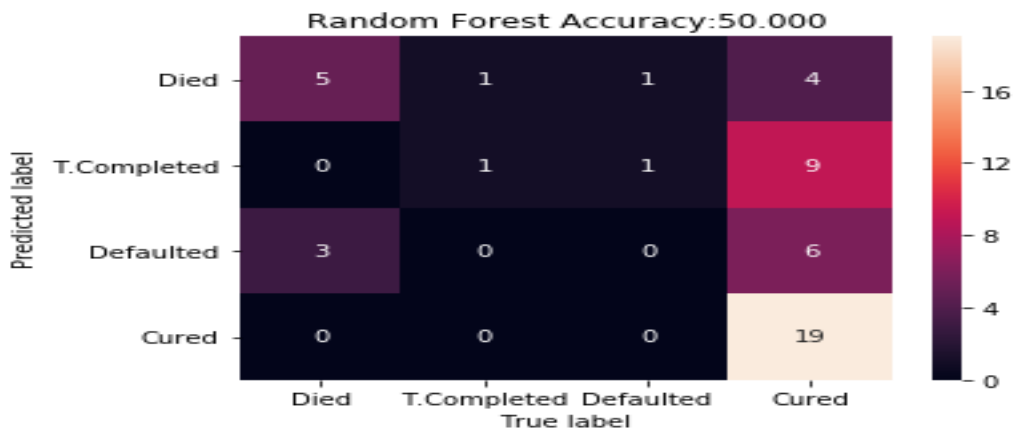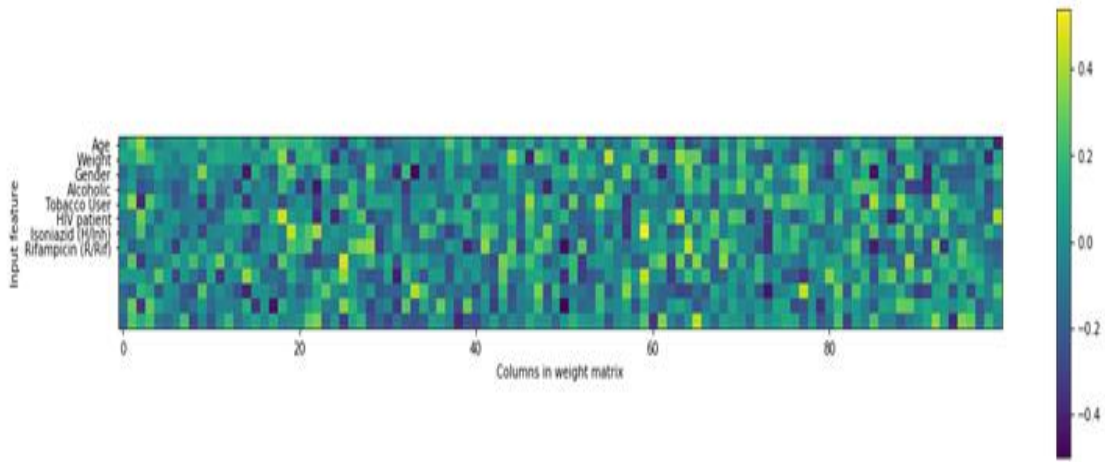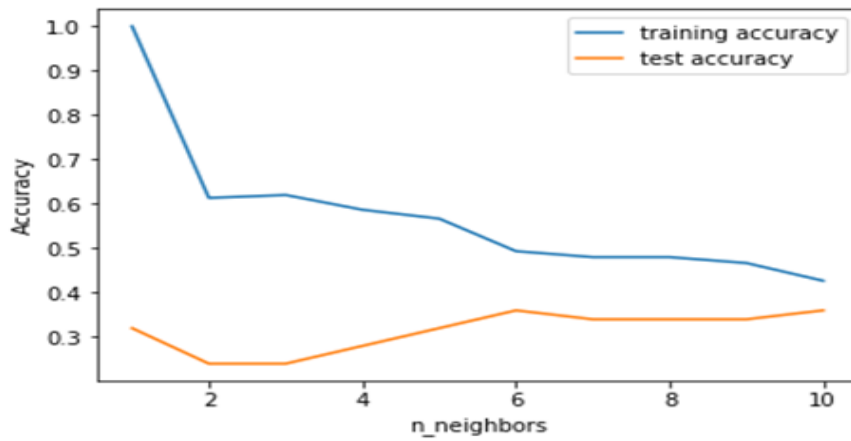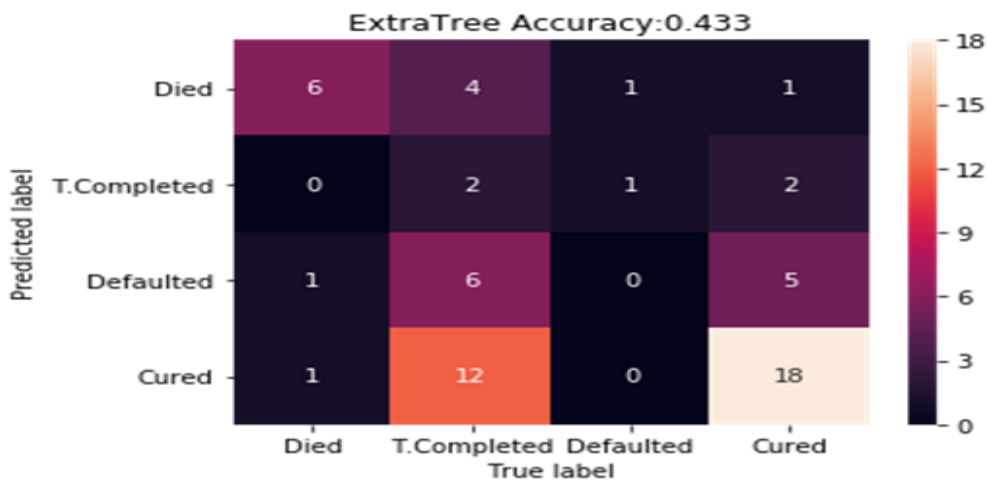

Figure 1

Figure 2

Figure 3

Figure 4
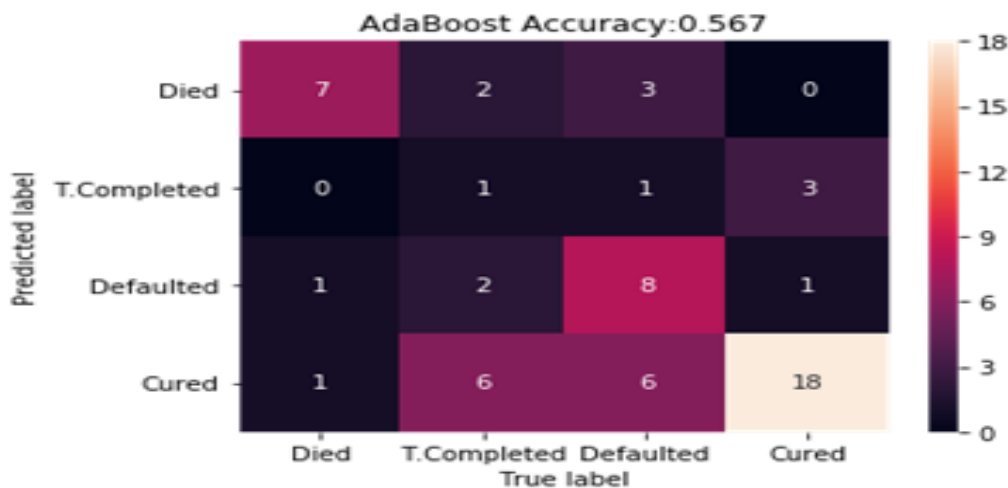
Figure 5

Figure 6

Figure 7

Figure 8

V. CONCLUSION

MDR-TB diagnosis has been one of the serious threat and is increasing in alarming rate. Various approaches are being done to cope with it. As from the analysis we inferred that different types of approaches like genome sequencing, Machine learning approaches for prediction and many voluntary organizations like DOTs are providing different contributions in eradicating MDR-TB. This paper goes through some of the major related works with respect to MDR-TB. Also here an analysis of different Machine learning and Ensemble Methods on MDR-TB dataset. AdaBoost ensemble algorithm gave the highest accuracy of all. As future scope, many more features can be considered so to increase the accuracy. Also, deep learning algorithms can be performed on the dataset.

REFERENCES

1. Anthony S. Fauci and the NIAID Tuberculosis Working Groupa, Multidrug-Resistant and Extensively Drug-Resistant Tuberculosis: The National Institute of Allergy and Infectious Diseases Research Agenda and Recommendations for Priority Research, Field: Tuberculosis Research Centre Experience, The Indian Journal for Tuberculosis, 2017.
2. Erol S. Kavvas, Edward Catoiu, Nathan Mih, James T. Yurkovich , Yara Seif , Nicholas Dillon, David Heckmann, Amitesh Anand, Laurence Yang, Victor Nizet ,Jonathan M. Monk& Bernhard O. Palsson. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance, Nature Communications. 2018.
3. Emily A. Kendall, Andrew S. Azman, Frank G. Cobelens, David W. Dowdy. MDR-TB treatment as prevention: The projected population-level impact of expanded treatment for multidrug-resistant tuberculosis, PLoS One,2017.
4. Lokeshkumar R, Dr. P.Sengottuvelan, A Survey on Preprocessing of Web Log File in Web Usage Mining to Improve the Quality of Data, International Journal of Emerging Technology and Advanced Engineering (IJETAE), Volume 4, Issue 8, PP. 229-234,August 2014, ISSN 2250-2459.
5. Martha Tatusch and Stefan Conrad, Detection of Multidrug-Resistant Tuberculosis Using Convolutional Neural Networks and Decision Trees, CLEF, 2018.
6. Michael L. Chen, Akshith Doddi, Jimmy Royer, Luca Freschi, Marco Schito, Matthew Ezewudo, Isaac S. Kohane, Andrew Beam and Maha Farhat , Deep learning predicts tuberculosis drug resistance status from genome sequencing data,bioRxiv,2018.
7. Samaneh Kouchaki, Yang Yang , Timothy M. Walker , A. Sarah Walker,Daniel J. Wilson , Timothy E.A. Peto, Derrick W. Crook David A. Clifton and CRyPTIC consortium, Application of machine learning techniques to tuberculosis drug resistance analysis,Bioinformatics,2018.
8. Sharareh R. Niakan Kalhori, Xiao-Jun Zeng, Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course. Scientific Scholar,2013.
9. Yan Xiong, Xiaojun B, Ao Hou, Kaiwen Zhang, Longsen Chen, Ting Li, Automatic detection of mycobacterium tuberculosis using artificial intelligence, Juornal of Thoracic Disease,2018
10. Yang Yang, Katherine E. Niehaus, Timothy M. Walker, Zamin Iqbal, A.Sarah Walker, Daniel J. Wilson , Tim E. Peto, Derrick W. Crook,Smith E. Grace, Tingting Zhu, and David A. Clifton. Machine Learning for Classifying Tuberculosis Drug-Resistance from DNA Sequencing Data, Bioinformatics,2018