Pak. J. Biochem. Mol. Biol., 2024, 57(1): 61-67

**Research Paper**

# Pattern Identification of Drug Resistance for Tuberculosis using Machine Learning Techniques

Anam Tariq[1], Omaid Ghayyur [2]and Sahar Fazal[3]

[1]Department of Biosciences, Mohammad Ali Jinnah University, Karachi, Pakistan.

[2]Iqra University, Sector H9, Islamabad Campus, Pakistan.

[3]Capital University of Science and Technology, Islamabad, Pakistan.

**ABSTRACT**

Among all other infectious diseases, the 2nd most infectious disease is Tuberculosis (TB), which is caused by a gram-positive bacterium called *Mycobacterium tuberculosis*. It is the main cause of death in people in developing countries. There are different types of TB. In some people, the disease progression is immediate and is called active TB, but in some patients the bacteria remain in the inactive state and are called inactive TB. Multiple drugs are being used for the treatment of TB but people are becoming resistant to these drugs because every human being has a different response to the drugs. In the health care domain, data mining is used for the processing of massive data. The aims and objectives of this research were to explain some supervised learning algorithms used in finding drug resistance in patients from different regions of Pakistan, and also the type of drug resistance developed in patients from Pakistan. These algorithms include the random forest algorithms, decision tree, Naïve Bayes classifier, and the support vector machine. The results showed that the highest accuracy of results was achieved by the use of the Naïve Bayes classifier, which gave 96.5% accuracy. It was concluded from the research that this research can be expanded by applying other different data mining techniques, like clustering and time series.

**KEYWORDS:** Data Mining; Decision Tree; Machine Learning; Naïve Bayes; Random Forest; Tuberculosis.

## INTRODUCTION

The second most infectious disease is Tuberculosis (TB) among all other infectious diseases, which is leading people to death throughout the world [1]. In developing countries, 95% of deaths are reported because TB is the most common disease there. With the occurrence of the disease, the immediate progression of the disease is observed in patients. Tuberculosis majorly impacts the lungs but also has severe impacts on the spine, kidneys, and the brain of the person. In most of the tuberculosis cases, the bacteria remain inactive, whereas in one-tenth of the cases of tuberculosis, the bacteria become active and cause the infections, which are known as active TB.

*Mycobacterium tuberculosis* (MTB) is the causative agent of TB. Through the air and by inhaling the contaminated droplets, the TB is spreading from one person to the other. When a carrier of TB sneezes, coughs, or spits, and the healthy person inhales that germs that are present in the air, the healthy person can also become infected with TB [2].

The treatment of TB is directly related to the effectiveness of the vaccination or drug regimen. If the bacteria are actively dividing inside the infected person, then the antibiotic drugs for TB are very effective against TB. The drugs first kill the actively dividing bacteria in the extreme stages of the treatment of the TB, and the clinical symptoms get disappear. The first line drugs (FLD) called antibiotics are used in this

phase of treatment. However, the second-line drugs are used for the elimination of those bacteria that are dividing slowly. The success of the treatment of TB has been achieved by the use of the first-line drugs (FLD). The first-line drugs (FLDs) include the pyrazinamide (PZA), isoniazid (INH), rifampin (RIF), streptomycin (SM), and ethambutol (EMB), but for several reasons, these drugs seem to have failed in some TB cases [3].

When the anti-tuberculosis drug is used inaccurately by health care providers, or they use it improperly, or the patients stop taking medicines before the treatment is completed, the resistance is developed in the MTB. A form of TB is developed called the multi-drug resistant tuberculosis (MDR-TB) among patients, in which there is no response to the first-line anti-TB drugs, while they are curable with the help of the second-line anti-TB drugs. Recently, on the clinical datasets, some machine learning techniques have been applied; these techniques were also applied to search the drug resistance [4]. An emerging subdivision of artificial intelligence is machine learning. The major purpose of machine learning is to design systems that can predict and learn based on experience. A training dataset and the hidden patterns in the inputs are detected and used to build the models by machine learning. The missing values are filled, and the dataset is preprocessed. For the prediction, the model uses the new input data and is then tested for accuracy [5]. The prediction of the subsequent cases by the classification task depends on the past information. For the prediction, which depends on the dataset, many classification techniques are used.

Due to the bacterial resistance against the first-line antimicrobial drugs, it is a very big issue for the treatment of Tuberculosis because of the unpredictable response to the antibiotics of each individual. By fetching the clinical data of the patients, the response can be predicted easily. To predict the patterns from the clinical data, a huge amount of data needs to be analyzed. In reviewing the drug resistance, these patterns are very significant. On the basis of the clinical data, there is no strong solution for the prediction of the trends in the MDR-TB. The data of MDR-TB is available, but research on this data is not available, which can be used for clinical data, and also, there is no data on drug response to predict the resistance of drugs for TB.

## MATERIALS AND METHODS

The objective of this research was to predict the identification of patterns of resistance of drugs for TB by using the algorithms of machine learning. To achieve this objective of the research multiple algorithms of the machine learning were used which includes the Artificial Neural Network Multi-Layer Perceptron (MLP) [6], Decision Tree (DT) [7], Random Forest (RF) [8], Support Vector Machine (SVM) [9] and Naïve Bayes [10] on the data set, that are supervised learning algorithms [11]. With a labelled data set many supervised algorithms of machine learning were used previously, including these algorithms, the ANN-MLP was also used for the prediction and the classification. For higher precision, some important characteristics that lead to the other features were also identified in this research. Not all the characteristics may lead to a considerable amount of outcomes, so it may lead to additional areas of different trials for the patients. The sequential chart of the algorithms used in the prediction and classification is given in Figure 1.

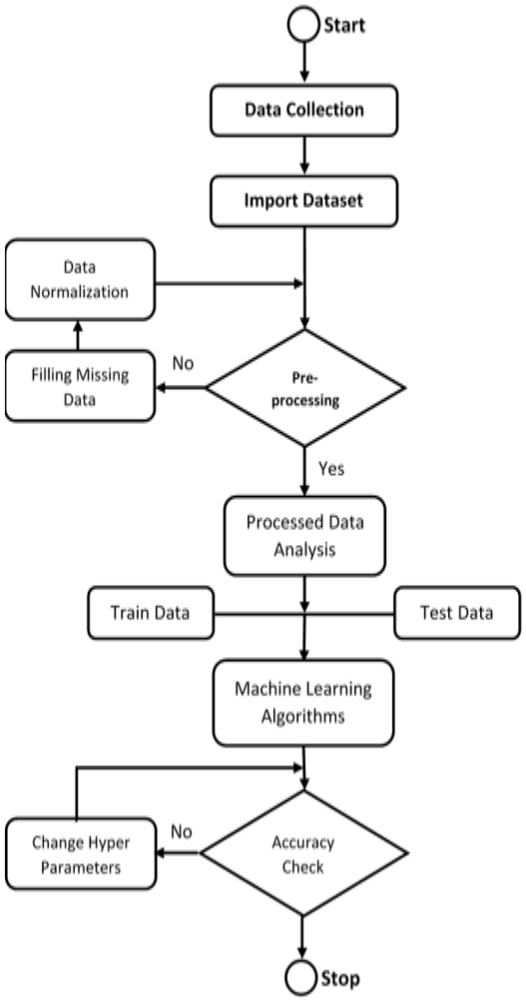


**Figure 1:** Sequential Flowchart of General Methodology used for Machine Learning Algorithms.

In this research, from the multiple centers in the Khyber Pakhtunkhwa (KPK), the real clinical data were collected. From the multiple institutes of Rawalpindi (Punjab) and Karachi (Sindh), similar data were collected. The real data sets consist of the 1800 samples with 17 attributes and 16 features, and one class label. The demographic data contains the gender, location, age, treatment history, disease type, HIV status and the data of response of drugs which includes the history of treatment and identifies that whether the patient is sensitive or resistant to the Isoniazid, Moxifloxacin, Ethambutol, Rifampicin, Kanamycin, Amikacin, Ofloxacin, Capreomycin and Pyrazinamide shown in the table 1. In this research, five different algorithms were used to classify and predict multiple drug resistance and to create a model to achieve the maximum possible correctness for the drug resistance pattern prediction.

A large number of the missing data was included in the real information, and therefore to overcome such problems and to make the predictions accurate, the data is preprocessed. The data was obtained from different sources, which is why the data integration was done before the data preprocessing. To get effective and accurate results, the data was cleaned in such a way that the missing values were filled and the noise was removed from the data. In the data preprocessing, to achieve effective results in the reduction of the data gained, a total 400 rows were selected, which had the complete data of the patients' information. A total 12 significant characteristics were selected which including history, gender, type of TB, INH, MOC, AM,CAP, ETM, KAN, PZA, and OFX that provide important information for the classification and the prediction. Transformation of data was also performed to make the data more comprehensive. In the data transformation, the data is altered from one form to another. It involved the tokenization and the normalization of the data. To convert the categorical data into numerical data, the label encoder was used for the prediction and classification. After the pre-processing of the data to achieve the results with maximum accuracy, the data was then classified and split into a training dataset and a test dataset, which was run on various algorithms.

A supervised learning algorithm called Multi-layer Perceptron was used for the prediction and classification. The function used by this algorithm by training on the data, was $f(.): R^m \rightarrow R^o$. in this function, the number of dimensions for the input is represented by m and the number of the

dimensions for the outputs are represented by the o. A set of features was given like $X = x_1$, $x_2$ ...$xm$ and a target $y$, which it can learn for the classification. There can be hidden layer which are one or more than one non-linear layers. Using artificial neural network with a label and the 12 features called multi perceptron model, 91% accuracy has been achieved. For testing and training the data was split with a ratio of 70:30. With 12 neurons, relu activation function with adam optimizer the hidden layers were created and for the model training the total 120 epochs were run.

A classification algorithm works on the numerical as well as categorical data is called a decision tree. For the creation of the tree like structures the decision trees are used. To handle the medical datasets the decision tees are widely used because they are simple and easy in use. In the tree shaped graphs, it is very simple to analyse the data with the help of decision trees. On the three nodes the decision tree model perform the analysis. The 1st node is the main node which is the root node, on the basis of which all the other nodes work. The 2nd node is the interior node which controls different features. The 3rd node is the leaf node which used to show the results of each test. The algorithm of the decision tree used to split the data into two or than two analogous sets which are based on the most significant indicators. With the analysts who have the maximum information gain and the minimum entropy, the data was divided and each feature's entropy was calculated by an equation: $Entropy\ (S) = \sum_{i=1}^{n}$ Pi log2Pi. With the selected features, by using the decision tree an accuracy of the 94.8% has been achieved.

An algorithmic technique of the supervised classification is called the random forest algorithm. Several trees used to create a forest in this algorithm. In random forest the individual tree is letting out the expectations of the class and the class which have the higher votes will turn into the forecast models. The high accuracy is obtained by the more number of trees in the random forest classifier. The three common methodologies used in the random forest algorithm are forest blend RC, forest random input choice RI and the combination of the RC and RI. For the regression tank and the classification tank the random forest is used but it is used for the classification task easily and in a proper way and can overcome the missing values. The results were uncountable because it requires more trees and the large data sets to obtain the predictions. With datasets the accuracy obtained was 95.6% by the random forest algorithm.

A supervised algorithm is the Naïve Bayes classifier. Bayes theorem is used by the Naïve Bayes classifier which is a simple classification. Among the attributes this theorem assumes the strong independence of the Naïve. A mathematical concept to get the probability is the Bayes theorem. The predictors are not related to one another. All the features independently donate to the probability to maximize it. It is possible to work on the Naïve Bayes Model but not to use the methods of Bayesian. Naïve Bayes classifiers were used by many complex real world situations. The general equation of algorithm is shown as $P(X/Y) = \frac{P\left(\frac{Y}{X}\right) \times p(X)}{P(Y)}$. The non-linear and the complicated data is controlled by the Naïve Bayes model and it is an efficient classification algorithm and very easy and simple to apply it. But there is a loss of the accuracy in the results because this model is based on the independence of the class conditionals and the assumptions. 96.5% accuracy was obtained by using the Naïve Bayes model.

Another simple algorithm which is used for the regression and the classification is called the support vector machine and it is used for the classification. The highly preferred algorithm is the support vector machine because, with the less computation power a significant accuracy is obtained. The

algorithm of the support vector aimed to find the hyperplane in the N-dimensional space, the number of features represented by N that classifies the data points. Using the 12 features, 93.90% accuracy was obtained by the algorithms of the classification of the support vector machine. For classification, over rest decision function and the RBF kernel were used.

## RESULTS AND DISCUSSION

The objectives of this research were to find out whether the patients of tuberculosis are developing resistance against the drugs or not, and if the patients are developing resistance against the drugs of tuberculosis then what type of resistance is that. The classification techniques of the supervised machine learning were used to perform this research and this research was carried out by using the different algorithms on the real clinical data of the patients belong to the KPK and other regions of the Pakistan including Artificial Neural Networks Multi-Layer Perceptron, random forest, decision tree, Naïve Bayes algorithm and support vector machine. Using the Python language, the different classifier algorithms were tested in various experiments. The 4$^{th}$ generation Intel Core i7 which had 16 GB RAM and a processor up to 2.3 GHz CPU. The dataset was divided into the test set and the training set. To get accuracy in the results, different techniques of the supervised classification discussed in the methodology were applied and the pre-processing of the data was done. Using Python programming for the test datasets and the training datasets, the different scores of accuracy of results of classification techniques were noted. In the table 2 the scores of the percentage accuracy are shown and in the figure 2 the different algorithms percentage accuracy is shown.

## CONCLUSION

The major aim of this research was to identify different methods which are significant in the effective and efficient prediction of the resistance of tuberculosis patients of tuberculosis in Pakistan, with patients' history and demographic regions. Our goal was the accurately and efficiently predict with a small number of tests and features. In this research, on the clinical datasets, five classification techniques for supervised learning were applied, and the 12 essential attributes were selected. Best results were achieved by using the Naïve Bayes, Random Forest, and Decision Tree algorithms with the data set of 400 rows, and hypothetically, the accuracy may be affected by large data set, and ANN–MLP accuracy may increase to give better predictions. This research can be expanded by introducing other techniques of data mining, like clustering and time series. To get greater accuracy with the lesser loss, there is a need to apply the combinations of more complex models for the prediction of the resistance of drugs and a pattern study for different regions of Pakistan.

**Table 1. Features and Details of Dataset of TB Drug Resistance.**

| Sr. No. | Attributes | Representation | Details |
|---|---|---|---|
| 1 | Gender | Gender | Male or Female |
| 2 | Age | Age | Patients Age in Years |
| 3 | History | History | Previous TB disease history if known (Never Treated, Previously Treated or Unknown) |
| 4 | Reason | Reason | History Information Reason (Diagnosis, Follow-up Checkups) |
| 5 | Tuberculosis Type | TBType | TB Type (Extra Pulmonary, Pulmonary) |
| 6 | Sample Type | Stype | Patient Sample for Test (Ascitic Fluid, Sputum, Pus, Bronchoalveolar Lavage/Washing, CSF, Pleual Fluid, Tissue Biopsy, Lymph Node) |
| 7 | Test Result | Result | Mycobacterium Tuberculosis (MTBC) |
| 8 | Moxifloxacin | MOX | Response to Drug (Sensitive, Resistant) |
| 9 | Isoniazid | INH | Response to Drug (Sensitive, Resistant) |
| 10 | Rifampicin | RIF | Response to Drug (Sensitive, Resistant) |
| 11 | Ethambutol | ETM | Response to Drug (Sensitive, Resistant) |
| 12 | Amikacin | AM | Response to Drug (Sensitive, Resistant) |
| 13 | Kanamycin | KAM | Response to Drug (Sensitive, Resistant) |
| 14 | Capreomycin | CAP | Response to Drug (Sensitive, Resistant) |
| 15 | Ofloxacin | OFX | Response to Drug (Sensitive, Resistant) |
| 16 | Pyrazinamide | PZA | Response to Drug (Sensitive, Resistant) |
| 17 | Drug Resistance Result | DSTResult | Drug Resistance (Multiple Drug Resistance MDR, Any Resistance, All Sensitive) |

## REFERENCES

[1] A. MacNeil, P. Glaziou, C. Sismanidis, A. Date, S. Maloney, and K. Floyd, "Global epidemiology of tuberculosis and progress toward meeting global targets—worldwide, 2018," *Morb. Mortal. Wkly. Rep.*, vol. 69, no. 11, pp. 281–285, 2020.

[2] N. R. Meier, M. Jacobsen, T. H. M. Ottenhoff, and N. Ritz, "A systematic review on novel Mycobacterium tuberculosis antigens and their discriminatory potential for the diagnosis of latent and active tuberculosis," *Front. Immunol.*, vol. 9, p. 2476, 2018.

[3] Y. Lan, Y. Li, L. Chen, J. Zhang, and H. Zhang, "Drug resistance profiles and trends in drug-resistant tuberculosis at a major hospital in Guizhou Province of China," *Infect. Drug Resist.*, vol. 12, p. 211, 2019.

[4] M. S. Kamal, N. Dey, and A. S. Ashour, "Large scale medical data mining for accurate diagnosis: A blueprint," in *Handbook of large-scale distributed*

*Computing in smart healthcare*, Springer, 2017, pp. 157–176.

[5] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[6] T. Karay\ilan and Ö. K\il\iç, "Prediction of heart disease using neural network," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 719–723.

[7] M. S. P. Kamiński Bogumiłand Jakubczyk, "A framework for sensitivity analysis of decision trees," *Cent. Eur. J. Oper. Res.*, vol. 26, no. 1, pp. 135–159, 2018.

[8] F. Fabris, A. Doherty, D. Palmer, J. P. De Magalhaes, and A. A. Freitas, "A new approach for interpreting random forest models and its application to the biology of ageing," *Bioinformatics*, vol. 34, no. 14, pp. 2449–2456, 2018.

[9] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics-Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[10] C. Gao, Q. Cheng, P. He, W. Susilo, and J. Li, "Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack," *Inf. Sci. (Ny).*, vol. 444, pp. 72–88, 2018.

[11] A. R. Iyanda, O. D. Ninan, A. O. Ajayi, and O. G. Anyabolu, "Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models.," *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, no. 6, 2018.