Machine Learning Model to Predict TB Bacteria-Resistant Drugs from TB Isolates

Authors : Rosa Tsegaye Aga, Xuan Jiang, Pavel Vazquez Faci, Siqing Liu, Simon Rayner, Endalkachew Alemu, Markos Abebe

Abstract : Tuberculosis (TB) is a major cause of disease globally. In most cases, TB is treatable and curable, but only with the proper treatment. There is a time when drug-resistant TB occurs when bacteria become resistant to the drugs that are used to treat TB. Current strategies to identify drug-resistant TB bacteria are laboratory-based, and it takes a longer time to identify the drug-resistant bacteria and treat the patient accordingly. But machine learning (ML) and data science approaches can offer new approaches to the problem. In this study, we propose to develop an ML-based model to predict the antibiotic resistance phenotypes of TB isolates in minutes and give the right treatment to the patient immediately. The study has been using the whole genome sequence (WGS) of TB isolates as training data that have been extracted from the NCBI repository and contain different countries' samples to build the ML models. The reason that different countries' samples have been included is to generalize the large group of TB isolates from different regions in the world. This supports the model to train different behaviors of the TB bacteria and makes the model robust. The model training has been considering three pieces of information that have been extracted from the WGS data to train the model. These are all variants that have been found within the candidate genes (F1), predetermined resistance-associated variants (F2), and only resistance-associated gene information for the particular drug. Two major datasets have been constructed using these three information. F1 and F2 information have been considered as two independent datasets, and the third information is used as a class to label the two datasets. Five machine learning algorithms have been considered to train the model. These are Support Vector Machine (SVM), Random forest (RF), Logistic regression (LR), Gradient Boosting, and Ada boost algorithms. The models have been trained on the datasets F1, F2, and F1F2 that is the F1 and the F2 dataset merged. Additionally, an ensemble approach has been used to train the model. The ensemble approach has been considered to run F1 and F2 datasets on gradient boosting algorithm and use the output as one dataset that is called F1F2 ensemble dataset and train a model using this dataset on the five algorithms. As the experiment shows, the ensemble approach model that has been trained on the Gradient Boosting algorithm outperformed the rest of the models. In conclusion, this study suggests the ensemble approach, that is, the RF + Gradient boosting model, to predict the antibiotic resistance phenotypes of TB isolates by outperforming the rest of the models.

Keywords : machine learning, MTB, WGS, drug resistant TB
Conference Title : ICMLHM 2024 : International Conference on Machine Learning for Healthcare and Medicine
Conference Location : London, United Kingdom
Conference Dates : May 23-24, 2024