A Novel Machine Learning Method for Tuberculosis Drug Discovery

Tiwari, Aditya (School: Adlai E. Stevenson High School)
Lakhanpal, Gaurish (School: Adlai E. Stevenson High School)

Purpose: To create a machine learning model that could identify drugs that could be effective against Multi-Drug Resistant Tuberculosis. Procedure: The Assay Central TB dataset was acquired and read using the RDKit library. The SMILES column, which described the structure of each compound, was isolated and the characters of each SMILES were tallied and stored in a single vector. Each vector was normalized and multiple machine learning models (Random Forest, SVM, Logistic Regression, Neural Networks, K-Nearest Neighbors) were trained to determine whether a given compound was effective against Tuberculosis. The SVM was selected on the basis of its Stratified KFold precision and accuracy and used to identify drugs from the Drug Repurposing Dataset that could be effective against Tuberculosis. Conclusion: The model is both highly precise and accurate (>90%) when it comes to predicting drug effectiveness The model identified Clofazimine, which is currently used to treat leprosy, as a potential tuberculosis treatment. In this state, the model could be used to identify more tuberculosis drugs from a larger dataset, and it could also be adapted to discover drugs that could be effective against diseases like Malaria and HIV/AIDS. It could be potentially improved by disproportionately weighting each class or by using ensemble methods. Further research needs to be conducted on the efficacy of Clofazimine against tuberculosis.