Prediction of Drug Resistance in Tuberculosis

Supriya Suresh[1], B Nayana[2], Bharath J Gowda[3], Nilesh Rajan[4], Suravi H U[5]

[1]Assistant Professor, Computer Science and Engineering, K S School of Engineering and Management,
[2,3,4,5]Student, Computer Science and Engineering, K S School of Engineering and Management,

Abstract

This Machine learning (ML) has emerged as a transformative tool in healthcare, enabling early detection and prediction of complex conditions by analyzing large-scale, heterogeneous datasets. In the context of tuberculosis (TB), ML methods can efficiently interpret demographic, phenotypic, and genomic indicators to generate timely resistance predictions. The integration of Random Forest classifiers, data preprocessing techniques, and web deployment frameworks offers a practical pathway for clinical adoption, especially in low-resource settings.

The process used to evaluate tuberculosis drug susceptibility through conventional methods based on culture techniques operates at a slow pace that requires multiple weeks for conclusive results. The slow process of detection risks providing patients with inadequate treatment which leads to both disease transmission and negative patient health outcomes. The increasing spread of multidrug-resistant TB (MDR-TB) together with extensively drug-resistant TB (XDR-TB) makes diagnosis and treatment more complex. Traditional workflows struggle to process large-scale data and lack the necessary tools to understand complex patterns found among different patient characteristics.

The new solution combines Random Forest model with clinical features that include sputum culture outcomes and GeneXpert data along with mutation profiles. The model demonstrates accuracy scores of over 87% through cross-validation while using SMOTE for handling class imbalance. Healthcare providers gain secure browser-based access to the system through its deployment with Fast API and Streamlit. The early detection system provides healthcare providers with tools to intervene sooner while delivering enhanced tailored care for tuberculosis patients.

Keywords: Clinical Decision Support, Drug Resistance, GeneXpert, Machine Learning Tuberculosis, Random Forest, SMOTE, Tuberculosis.

1. Introduction

The worldwide epidemic of tuberculosis infection from Mycobacterium tuberculosis (MTB) remains a primary source of infectious disease mortality. The number of tuberculosis cases during 2023 reached 10.8 million while 1.25 million people died which positioned TB ahead of COVID-19 in worldwide death statistics. India bears the heaviest burden of TB infection with millions of new cases being reported annually. The standard treatment for tuberculosis infections requires a combination therapy of isoniazid and rifampicin.

Drug resistant tuberculosis forms emerged because of improper antibiotic use and insufficient antibiotic administration leading to MDR-TB and XDR-TB which now require more challenging and expensive treatment approaches. Genetic changes in rpoB katG and inhA genes lead to the establishment of drug resistance. These genetic alterations in rpoB katG and inhA genes make drugs less effective or enable bacteria to persist which causes treatment to be ineffective.

Through the implementation of machine learning in tuberculosis diagnostic procedures the extended time period of 8 weeks for standard DST testing methods can be avoided. The artificial intelligence tools can analyze extensive datasets containing clinical information together with demographic profiles and genetic characteristics. The system predicts clinical outcomes more effectively and faster by generating real-world solutions that match current treatment requirements.

2. Literature Survey

[1]. The investigation proves how ML techniques can successfully forecast drug resistance patterns in Ugandan MTB strains by using SNP and clinical information. The models Logistic regression, XGBoost, and Gradient Boosting reached AUC scores as high as0.96. The research demonstrates show regional differences in population composition when combined with diverse datasets lead to model performance that works effectively for various population group [2].

The research work performed an extensive analysis on 32,000 isolates to discover. Genetic mutations that result in drug resistance remain undisclosed to medical professionals The examination of rpoA, rpoC, and ahpC discovered that compensatory mutations play an important role in enhancing the accuracy of resistance prediction models. The study maintained its focus on genetic data since it did not incorporate any phenotypic or clinical information for analysis.[3].

The research project examined patient treatment adherence by implementing both SVMs and Random Forests to detect essential predictors of non-adherence to TB therapy. The study identified sub-county together with antiretroviral status and treatment support as important variables. The research findings suggest that clinical data obtained through non-genomic methods can improve the accuracy of resistance predictions. Antimicrobial resistance [4]. The implementation of deep convolutional neural networks in this research allowed the system to detect mutations for 13 antibiotics. The research achieved AUC scores beyond 97% for specific drugs by using CNN technology to find resistance patterns that were previously unknown. This method requires extensive computational resources which makes it unsuitable for poor-resource environments.

3. Methodology

A. Design and Workflow:

The predictive system operates through separate modules in a workflow framework shown in Figure 1 The process begins by accepting user data before conducting initial data exploration which leads to model generation and evaluation. model validation and real-time forecast delivery. The system accepts data through a protected user interface then uses EDA methods to conduct analysis before it transfers the information to a Random Forest classifier.

B. Implementation Details

Data Preprocessing: The system preprocesses patient demographic information with the results from GeneXpert tests and sputum culture information by standardizing and encoding them using one-hot encoding. SMOTE is applied to address class imbalance.

Feature Engineering: The system determines the presence of resistance-associated gene mutations in rpoB and katG along with key clinical indicators such as weight loss and fever and HIV status through domain knowledge together with EDA analysis.

Model Training: The Random Forest classifier receives training through a 70/30 stratified train-test split. The system conducts hyperparameter tuning to determine the optimal number of estimators along with tree depth.

Validation: The model receives testing through metrics that include accuracy, precision, recall, F1-score and ROC-AUC. The classifier achieved a ROC-AUC of 0.94 for rifampicin resistance prediction.

Deployment: A FastAPI endpoint manages incoming data processing while a Streamlit-based GUI provides users with real-time resistance predictions based on their patient data entry.

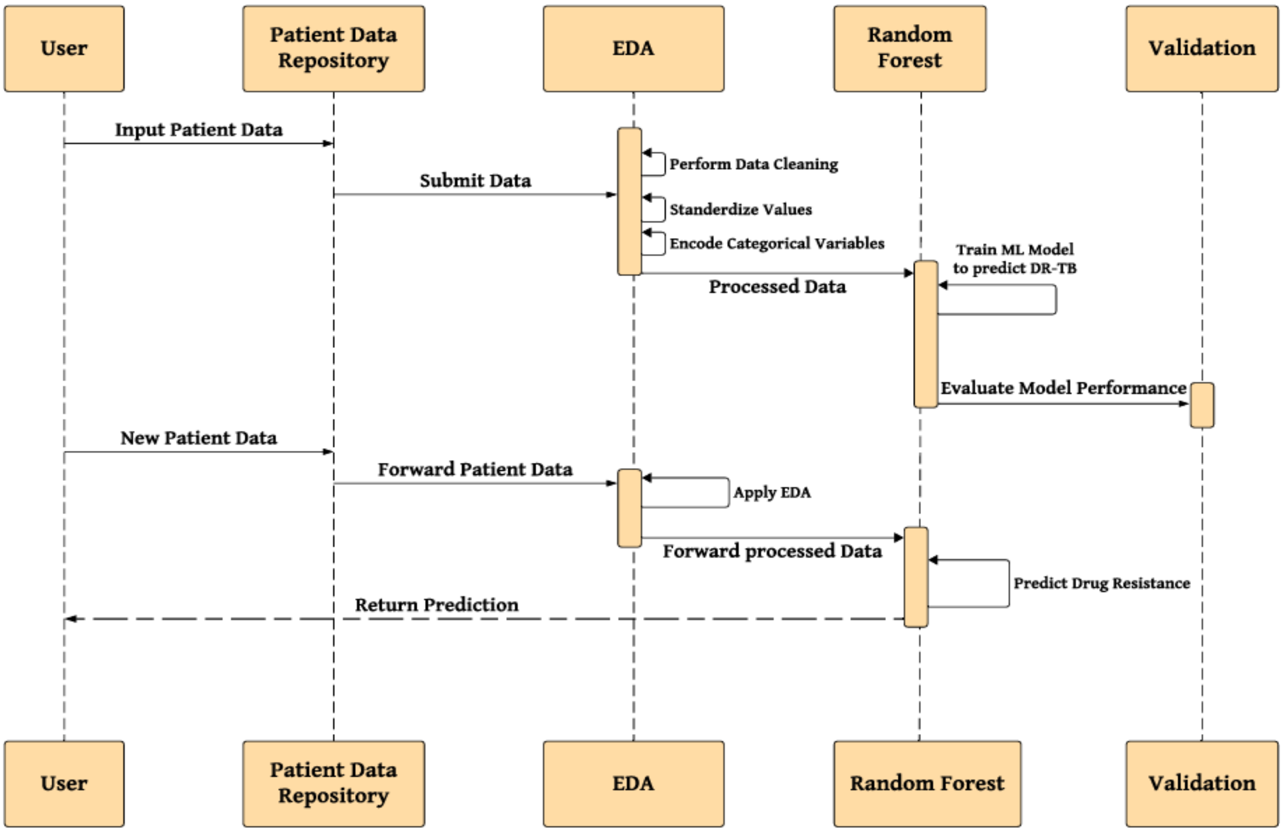

Fig 1: Design of the Project

4. Results and Discussions

The effectiveness of the machine learning model, specifically the Random Forest Classifier, was evaluated using a broad set of performance metrics including accuracy alongside precision as well as recall and F1-score. The model achieved promising results because it delivered strong performance in both accuracy and drug resistance prediction. The classification model evaluation process includes multiple assessment criteria which measure precision, recall, F1-score together with accuracy and ROC-AUC performance. The evaluation metrics deliver a complete assessment of model performance across different class categories. Class 0 (Negative class): The model achieves exceptional performance in making correct class 0 predictions. The precision value stands at 0.9169 which indicates the model makes correct class 0 decisions about 91.69% of the time. The model reaches 0.9535 recall which means it correctly identifies 95.35% of all real class 0 test samples. The model maintains a strong performance across this class through its F1-score of 0.9349 which combines precision and recall metrics.

Class 1 (Positive class): The model performs comparatively weaker on class 1 predictions because the class contains only 59 instances when compared to 301 instances in class 0. The model reaches a precision value of 0.7021 which shows that 70.21% of the predicted class 1 results are actually accurate. The recall value decreases to 0.5593 which shows the model detects 55.93% of all real positive instances. The F1-score reaches 0.6226 which demonstrates an area where the model needs to improve its detection of the less frequent class.

Overall Performance: The model maintains a total accuracy score of 88.89% through which it correctly identifies 89 samples out of 100. Evaluating accuracy by itself proves inadequate for unbalanced datasets so you should evaluate additional performance measures.

Macro and Weighted Averages: The macro average treats both classes equally and yields an F1-score of 0.7787, while the weighted average, while handling imbalanced class distributions, gives a higher F1-score of 0.8837.

ROC-AUC Score: The ROC curve provides a visual depiction of how true positive rate and false positive rate relate to each other. The model receives a ROC-AUC score of 0.826 which indicates the strong ability to separate between the two classes. The model's curve appears clearly above the diagonal random guessing line which demonstrates its superior performance to random classification.

5. Conclusion

The most beneficial approach exists in combining machine learning methods with clinical information and genomic data represents a valuable way to fight against tuberculosis drug resistance. The Random Forest classifier receives its training from high-quality data which enables it to achieve precise predictions for rifampicin and isoniazid. Through this system medical professionals receive immediate decision assistance which results in faster interventions and specific treatment plans.

This method solves the time delay issue of DST results and particularly in places with limited access to laboratory resources. The system's transparent decision process and minimal computational requirements enable its seamless implementation into public health frameworks.

References

1. Xiong, X.-S., Huang, T.-T., Liu, Z.-Z., Li, Z.-K., Wang, L., & Li, F. Identification of Mycobacterium tuberculosis resistance to common antibiotics: An overview of current methods and techniques. Journal of Infection and Drug Resistance. (2024).
2. Sandra Ruth Babirye, Mike Nsubuga. Machine learning based prediction of antibiotic resistance in Mycobacterium tuberculosis clinical isolates from Uganda. BMC Infectious Diseases (2024).
3. Gary Napier, Susana Campino. Large-Scale Genomic Analysis of Mycobacterium tuberculosis Reveals Extent of Target and Compensatory Mutations Linked to Multi-Drug-Resistant Tuberculosis tuberculosis. Scientific Reports (2023).
4. Gichuhi, H. W., & Magumba, M. A machine learning approach to explore individual risk factors for tuberculosis treatment. PLOS Global Public Health, 3(7). (2023).
5. https://doi.org/10.1371/journal.pgph.0002357
6. Sultan, M. E. S., Murray, M. B., & Farhat, M. R. A convolutional neural network identifies mutations of interest to antimicrobial resistance in Mycobacterium tuberculosis. Journal Nature Communications in July (2022)