IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION OF TUBERCULOSIS BY DEEP LEARNING MODEL USING GENOMIC DATA

Mr. Pratik Dhote, Mr. Swapnil Dhoke, Mr. Subhash Rathod, Mr. Vinayak Madilwar, [1]Student,

[2]**Prof. Tara Shende** [2]Assistant Professor,

[1]Department Of Artificial Intelligence & Data Science

[1]Wainganaga College Of Engineering & Management, Nagpur, Maharashtra, India.

Abstract: The Tuberculosis (TB) remains a significant global health challenge, especially with the emergence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) strains of Mycobacterium tuberculosis. Early detection of antibiotic resistance is critical for initiating effective treatment and reducing transmission. The Tuberculosis Predictor project addresses this need by leveraging deep learning techniques to predict drug resistance from genomic data, providing a fast, scalable, and accessible diagnostic tool.

This application uses whole-genome sequencing (WGS) data in formats as FASTA to analyze genomic mutations associated with resistance to first-line and second-line anti-TB drugs. A deep learning model, trained on curated genomic datasets, evaluates input sequences and outputs resistance probabilities for multiple antibiotics. The system is designed with a user-friendly web interface, allowing researchers and clinicians to upload genetic data, visualize predictions, and download results. Python, Flask for the backend, and contemporary JavaScript for responsive interactions are used in the platform's construction. It offers dynamic results chart presentation, real-time file confirmation, and secure genetic data management. It is simple to see the likelihood of resistance since the model's predictions are represented by probability bars.

By enabling early, accurate detection of resistance patterns, Tuberculosis Predictor aims to assist healthcare professionals in making informed treatment decisions. The tool contributes to personalized medicine in infectious diseases and supports global TB control efforts through rapid diagnostics. Future enhancements may include expanding to other pathogens and integrating more advanced genomic features for even greater accuracy.

Keywords: Drug Resistance Prediction, Genomic Analysis, Whole Genome Sequencing (WGS), FASTA / VCF File Analysis, MDR-TB / XDR-TB, Web-based Application, Tuberculosis, Drugs, Deep Learning, Predictive models, Convolutional neural network (CNN.)

I. INTRODUCTION

Tuberculosis (TB), caused by Mycobacterium tuberculosis, remains one of the top infectious disease killers globally, despite advancements in medical science. The increasing emergence of drug-resistant strains such as multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) poses a major challenge to global health systems. Traditional drug susceptibility testing methods are slow, labor-intensive, and often require weeks to deliver results. This delay in identifying effective treatment regimens can lead to increased transmission, treatment failure, and higher mortality rates.

With the rise of whole-genome sequencing technologies, a wealth of genomic data is now available for clinical and research use. FASTA, a commonly used format for storing such data, is a standardized text representation of DNA sequences. Analyzing these sequences can reveal mutations in genes like katG, rpoB, and pncA, which are often responsible for resistance to drugs like isoniazid, rifampicin, and pyrazinamide, respectively.

Recent advances in artificial intelligence particularly deep learning have shown great promise in genomics. Deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformers are capable of detecting complex, non-linear patterns in sequential data like DNA. Applying these models to genomic sequences enables rapid, automated prediction of drug resistance, potentially transforming TB diagnostics and treatment strategies.

This project leverages deep learning to build a prediction system that can analyze DNA sequences from FASTA files and determine resistance to multiple TB drugs. The model is trained on labeled sequence data and integrated into a web-based interface where users can upload their files and receive instant predictions. By offering a fast, scalable, and accurate solution, this system aims to support clinicians in making timely decisions, reduce the spread of drug-resistant TB, and advance the use of AI in the field of precision medicine and microbial genomics.

II. REVIEW OF LITERATURE

Studies demonstrate the increasing integration of deep learning in tuberculosis (TB) diagnosis and drug resistance prediction. Singh et al. (2022) reviewed the evolution of machine learning applications in medical diagnostics, particularly focusing on TB, highlighting improved accuracy and early detection. Green et al. (2022) developed a convolutional neural network model to identify key mutations in Mycobacterium tuberculosis associated with antimicrobial resistance, showcasing how genomic data can be leveraged for precise predictions. These studies reflect a growing shift towards AI-driven healthcare, emphasizing the role of bioinformatics and computational modeling in enhancing TB treatment strategies.

2.1. EXISTING SYSTEM

Traditional systems for TB drug resistance detection primarily rely on culture-based drug susceptibility testing (DST) and molecular diagnostics like GeneXpert. These methods are often time-consuming, costly, and limited in scope, identifying resistance to only a few drugs. While recent studies have introduced machine learning and deep learning techniques using genomic and imaging data, many models remain research-focused, lack scalability, or require complex bioinformatics tools. Moreover, systems like DeepTB and TB-DROP show promising results but often necessitate advanced infrastructure or are confined to specific data types, limiting their routine use in low-resource clinical environments.

2.2. PROPOSED SYSTEM

The proposed system integrates deep learning models such as CNNs, LSTMs, and Transformers with whole-genome sequencing data in FASTA format to predict TB drug resistance. It automates the identification of resistance-conferring mutations and supports real-time analysis via a user-friendly web platform. Unlike previous models limited to specific inputs or clinical settings, this system is designed to be scalable, accessible, and rapid. It enables healthcare professionals to upload DNA sequences and receive instant predictions, facilitating timely treatment decisions. The solution bridges the gap between raw genomic data and clinical application, offering a practical tool for personalized TB management and public health surveillance.

Table 1: Existing & Proposed Systems According To Their Features

Feature	Existing System	Proposed System
Time to predict resistance	Several days to weeks	A few seconds
Required equipment	Lab-based culture testing	Web application with sequence input
Cost	High (lab reagents, technician time)	Low (single system, open-source tools)
Accessibility	Limited to labs with culture testing facilities	Accessible online, anywhere
Use of AI	Not applicable	Uses deep learning to predict resistance

III. ALGORITHMS USED

Convolution Neural Network (CNN):

A Convolutional Neural Network (CNN) is a class of deep learning algorithms especially effective at recognizing spatial or sequential patterns in data. Originally developed for image processing, CNNs have also proven powerful in analyzing genomic sequences due to their ability to automatically extract meaningful features from raw input data.

In the context of tuberculosis drug resistance prediction, DNA sequences from FASTA files are treated as one-dimensional sequential data. CNNs process this data using layers of filters (also called kernels) that slide across the input sequence. These filters act like pattern detectors, identifying important local features such as mutations or motifs in the genetic code that are associated with drug resistance.

CNNs are particularly well-suited for this task because they do not require manual feature engineering. Instead, they learn the optimal features during training, making the system efficient and adaptable to new genomic patterns. By leveraging CNNs, the model in this project can detect subtle genetic variations linked to TB drug resistance with high accuracy and speed.

IV. METHODOLGY

Methodology adopted for this project effectively combine computational biology and artificial intelligence to tackle the pressing issue of tuberculosis drug resistance. The use of a structured research approach ensured a scientifically valid and results-driven study, while the agile development methodology allowed iterative system development and real-time testing. Data was gathered from authentic genomic databases, processed using reliable bioinformatics tools, and analyzed using sophisticated deep learning techniques. The selection of CNN and BiLSTM models allowed for precise identification of mutation patterns, ensuring accurate classification of drug resistance. Using a variety of criteria to evaluate efficiency ensured robustness and dependability. The implementation of the web-based system using Flask offers real-time interaction, ease of use, and result visualization. Notably, the integration of a downloadable Excel feature enhances usability for clinical and research settings.

4.1PRE-PROCESSING

The preprocessing pipeline begins by parsing FASTA files using Biopython to extract DNA sequences. These sequences are then converted into a numerical format using one-hot encoding, representing each nucleotide as a binary vector. To ensure consistent input size for the deep learning model, sequences are either padded with neutral values or truncated to a fixed length. This standardized format enables efficient and accurate analysis by the neural network.

4.2TRAIN TEST SPLIT

To evaluate model performance, the dataset is divided into training and testing subsets. Typically, 70–80% of the data is used for training the model, while the remaining 20–30% is reserved for testing. This separation ensures that the model learns patterns from one portion and is validated on unseen data to assess generalization. Stratified splitting may be used to preserve class distribution, especially important in imbalanced datasets like those for drug resistance prediction.

4.3TRAINING

Training training involves feeding the preprocessed DNA sequences into the deep learning model to allow it to learn relationships between genetic mutations and drug resistance. The model adjusts its internal parameters weights and biases using optimization algorithms like Adam and a loss function, such as binary cross-entropy, to minimize prediction errors. Training takes place across a number of epochs, with each iteration improving the accuracy of the model. Validation data may also be used during training to monitor for overfitting.

4.4PICKING THE MODEL

Selecting the right model involves comparing various deep learning architectures like CNNs, LSTMs, and Transformers. Each model is trained on the same dataset, and their performances are evaluated using metrics such as accuracy, precision, recall, and AUC-ROC. The model that achieves the best performance on validation data, while maintaining efficiency and scalability, is chosen for final deployment. Factors such as training time, complexity, and interpretability are also considered during model selection.

V. MODEL BUILDING

In Model building involves designing and implementing a deep learning architecture tailored to analyze genomic sequences for predicting drug resistance in Mycobacterium tuberculosis. The model typically begins with input layers that accept one-hot encoded DNA sequences of uniform length. Convolutional layers are employed to extract local features, identifying patterns such as specific mutations linked to drug resistance. These are followed by pooling layers for dimensionality reduction and dropout layers to prevent overfitting. Fully connected dense layers then interpret the extracted features to produce a binary or multi-class output indicating resistance to specific TB drugs. The model is compiled using an optimizer like Adam and a loss function such as binary cross-entropy. Hyper parameters are fine-tuned to maximize predictive accuracy and generalization.

VI. WORKFLOW

The workflow begins with collecting Mycobacterium tuberculosis genomic data in FASTA format. This data undergoes preprocessing, such as encoding sequences using one-hot or k-mer strategies. The processed sequences are then fed into deep learning models like CNNs, LSTMs, or Transformers to learn mutation patterns associated with drug resistance. The models analyze these patterns to predict resistance against key antibiotics. The predictions are integrated into a user-friendly web interface, enabling healthcare professionals to upload sequence files and receive real-time resistance reports. This automated process aims to expedite and improve accuracy in TB diagnosis and treatment decision-making.

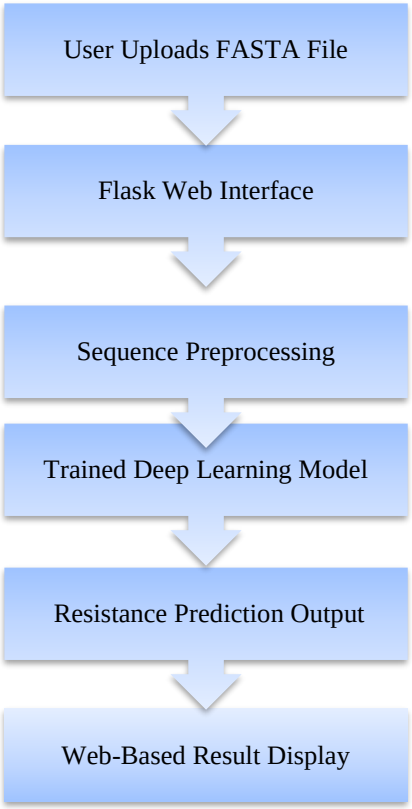


Figure 1: Workflow Diagram

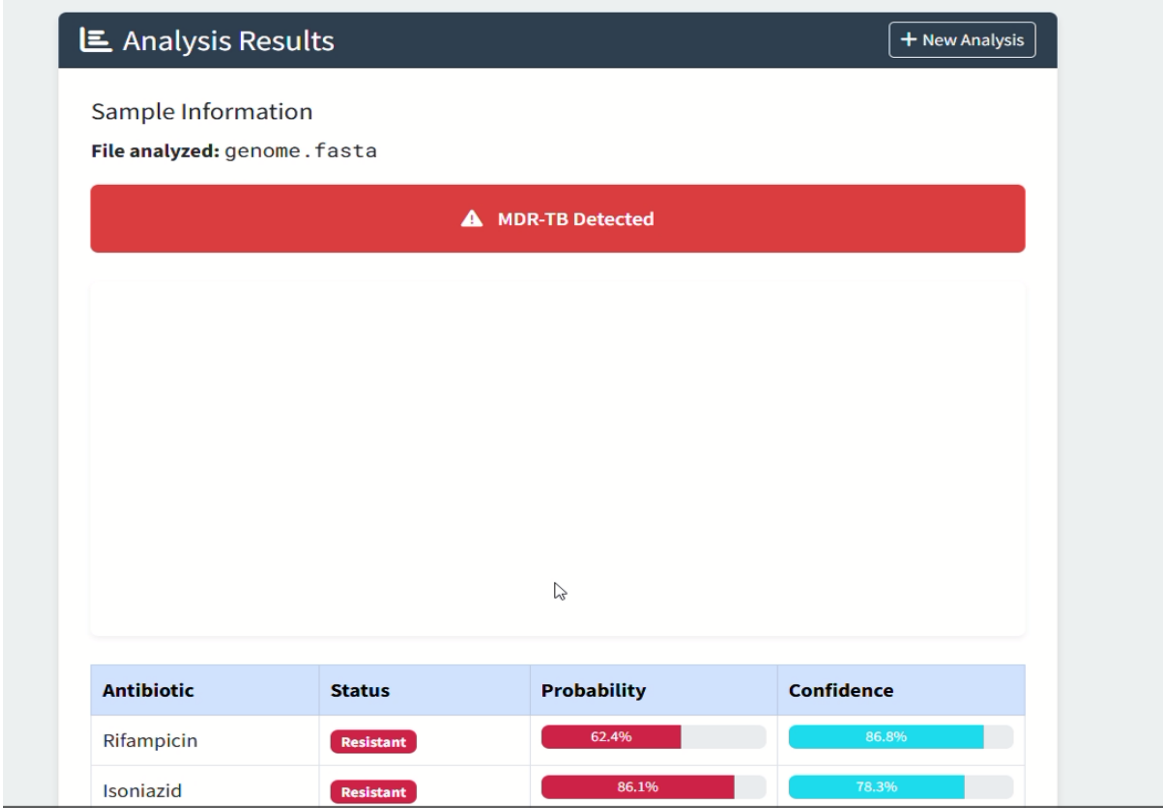
VII. CONCLUSION

This project demonstrates the transformative potential of combining bioinformatics and artificial intelligence for rapid and accurate tuberculosis drug resistance prediction. Utilizing deep learning models such as CNNs and LSTMs, the system enables scalable, cost-effective, and real-time analysis of genomic sequences in FASTA format. This innovative approach significantly reduces diagnostic time compared to traditional culture and molecular methods, facilitating early detection of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB strains. The user-friendly web interface supports clinical decision-making, promotes personalized treatment, and aids in controlling the spread of resistant strains. Continuous advancements, including integration with clinical workflows and explainability features, will further enhance its utility in global TB management and precision medicine.

VIII. FUTURE ENHANCEMENT

The TB drug resistance prediction system focus on increasing its scalability, accuracy, and clinical integration. One key area is integrating real-time clinical workflows, enabling seamless collaboration with hospital information systems for rapid diagnosis. Incorporating multi-modal data, such as clinical, demographic, and radiological information, can improve prediction robustness. Developing a mobile application version will enhance accessibility in remote and resource-limited settings. Additionally, expanding the system to include whole-genome sequencing (WGS) data and incorporating adaptive learning algorithms will allow continuous model updates as new resistance markers emerge. Implementing explainability features, such as highlighting mutations responsible for resistance, will increase clinician trust and interpretability. Further, integrating cloud-based infrastructure can facilitate large-scale deployment and data sharing across regions. Ultimately, these enhancements will make the system more versatile, accurate, and user-friendly, accelerating early detection and personalized treatment strategies for TB worldwide.

IX. OUTPUT



X. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Artificial Intelligence & Data Science, Wainganga College of Engineering and Management, Nagpur, for providing the resources and support necessary for this research. Special thanks to Prof. Tara Shende for her valuable guidance, mentorship, and continuous encouragement throughout the project. The authors also acknowledge the support of Rashtrasant Tukadoji Maharaj Nagpur University (RTMNU) for facilitating this academic endeavor.

XI. REFERRENCES

[1] Ankita Pal, Debasisa Mohanty, (2025) Machine learning-based approach for identification of new resistance associated mutations from whole genome sequences of Mycobacterium tuberculosis, Bioinformatics Advances Deep Learning Techniques

[2] Prasad Mahajan (2022) A Comparative Study of Detection of Tuberculosis using Machine Learning & Deep Learning

[3] Muhteşem Ağıldere, Karaca (2021) Comparative Study for Tuberculosis Detection by Using Deep Learning

[4] N. Bhaskar, P. T. Waghmare and S. V. A, (2023) Deep Learning Framework for Automated Drug Resistance Prediction of Tuberculosis Using Computed Tomography Images," 2023 International Conference on Next Generation Electronics (NEleX)

[5] Jiang Z, Lu Y, Liu Z, Wu W, Xu X, Dinnyés A, et al (2022-23) Drug resistance prediction and resistance genes identification in Mycobacterium tuberculosis based on a hierarchical attentive neural network utilizing genome-wide variants.