Advanced Machine Learning for Predicting Drug Resistance in Clinical Isolates of Mycobacterium tuberculosis Complex

Naoufal Sirri*
FEMTO-ST Institute, UMR 6174 CNRS,
University of Franche-Comté,
Belfort, France
naoufal.sirri@univ-fcompte.fr

Christophe Guyeux
FEMTO-ST Institute, UMR 6174 CNRS,
University of Franche-Comté,
Belfort, France
christophe.guyeux@univ-fcompte.fr

Christophe Sola
University Paris-Saclay, Bâtiment Bréguet,
3 Rue Joliot Curie, 91190
Gif-sur-Yvette, France
christophe.sola@universite-paris-saclay.fr

***Abstract*—This Tuberculosis is a major public health problem, and the diagnosis of multidrug-resistant and extensively drug-resistant tuberculosis is a global health priority. This resistance is mainly caused by mutations in genes coding for drug targets or conversion enzymes, but knowledge of these mutations is incomplete. Whole-genome sequencing (WGS) is an increasingly common approach for rapidly characterizing isolates and identifying mutations predictive of drug resistance. However, this technique has not accounted for the evolution of resistance. In contrast, machine learning methods have been widely applied to predict the resistance of Mycobacterium tuberculosis (MTB) to a specific drug in a timely manner and even identify resistance markers. In this study, machine learning approaches were applied to 28,073 MTB isolates that underwent WGS analysis and laboratory drug susceptibility testing (DST) for 10 antituberculosis drugs. Boosting models, such as extreme gradient tree (XGBoost), light gradient tree (LightGBM), and a deep neural network model with a new architecture, were used to predict drug resistance. The different proposed models were fitted distinctly for each drug, with the exploration of the 10 most influential feature classes that were used as input features during training to obtain satisfactory performance. The predictive performance was measured using sensitivity, specificity, the f1 score, the receiver operating characteristic curve (ROC) and the area under the curve (AUC). All three tools reliably predicted drug resistance, with the deep learning model outperforming all existing direct association-based approaches as well as the previously reported machine learning models, with AUCs ranging from 0.97 to 0.99 for 9 drugs. This work demonstrated the power of machine learning as a flexible approach for drug resistance prediction, which can consider a significant number of predictors and summarize their predictive ability, facilitating clinical decision making and detection of single-nucleotide polymorphisms in the era of increasing WGS data generation.**

Keywords-component; Tuberculosis; Drug resistance; Drug sensitivity testing; Whole-genome sequencin; Maching learning; Deep learning

I. INTRODUCTION

Tuberculosis (TB) ranks among the top 10 causes of death globally, highlighting its significant impact on public health [26]. The emergence of drug-resistant TB presents a formidable challenge to TB control efforts worldwide, posing a threat to public health systems. According to the World Health Organization (WHO), the prevalence of multidrug resistance is particularly high among first-line drugs used in TB treatment protocols [25]. These drugs encompass standard therapies such as Isoniazid (INH), Rifampicin (RIF), Ethambutol (EMB), and Pyrazinamide (PZA), along with second-line agents like Streptomycin (STR), Fluoroquinolones (OFX), Moxifloxacin (MOX), Ciprofloxacin (CIP), Kanamycin (KAN), Amikacin (AMK), and Capreomycin (CAP) [14]. The development and widespread adoption of predictive techniques leveraging artificial intelligence and machine learning offer promising avenues for addressing drug resistance in TB treatment. These techniques hold potential not only for enhancing the efficacy of existing therapies but also for facilitating the deployment of novel drugs targeting multidrug-resistant (MDR) or extensively drug-resistant (XDR-TB) strains, such as bedaquiline (BDQ), linezolid (LNZ), and delamanid (DEL). Conventional phenotypic drug susceptibility testing (DST) methods, while effective, suffer from limitations such as complexity, cost, and time requirements. In response, whole-genome sequencing (WGS) has emerged as a rapid and comprehensive approach for identifying drug-resistant TB strains [21]. By analyzing genetic variations, including single nucleotide polymorphisms (SNPs) and insertions or deletions, WGS enables the detection of mutations associated with drug resistance. However, while WGS holds promise for characterizing drug resistance, its utility may be limited by the complexity of TB resistance mechanisms. The genetic underpinnings of drug resistance are multifaceted, involving interactions between various mutations and genes within the Mycobacterium tuberculosis complex (MTC) [22][28]. Notably, resistance to Rifampicin often coincides with resistance to Isoniazid, underscoring the

interconnected nature of drug resistance pathways. Given the urgency of addressing drug-resistant TB, there is a pressing need to optimize diagnostic approaches and treatment strategies. Multidrug-resistant (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB) present formidable challenges, necessitating innovative solutions to combat these forms of the disease. By integrating advanced computational techniques, such as machine learning, with genomic data, we aim to develop predictive models capable of rapidly identifying drug-resistant TB strains and guiding personalized treatment interventions. In addition to traditional mutation-based techniques, machine learning (ML) models such as logistic regression (LR), support vector machine (SVM), and random forest (RF) have emerged as powerful tools for identifying drug resistance patterns. While these ML models exhibit comparable performance to variant-based association rules for well-studied medications like INH, RIF, and EMB, they often outperform them for less explored pharmaceuticals such as PZA. However, previous research has been limited by small sample sizes and a lack of comprehensive analysis of potential approaches for classifying drug resistance within the Mycobacterium tuberculosis complex (MTC). For example, a study by [27] examined over 1839 UK bacterial isolates and tested various categorization models for eight different drugs, while [9] investigated the effectiveness of RF using 1397 clinical isolates and a geographically diverse dataset. In these studies, a binary variable was employed to denote the presence or absence of each variant, facilitating the analysis of drug resistance patterns. [28] utilized LR to analyze 161 Chinese isolates and identify new genes associated with drug resistance to seven different drugs. However, the use of limited data can lead to overfitting, highlighting the need for larger and more diverse datasets to validate resistance predictions and develop generalized models. Furthermore, advancements in deep learning techniques, such as convolutional neural networks (CNNs) and hierarchical attentive neural network models (HANNs), have demonstrated superior accuracy and stability compared to traditional methods. For instance, [17] developed a CNN with high accuracy and stability, while [13] proposed a HANN that achieved high accuracy in predicting drug resistance and identifying related genes. Moreover, studies like [10] have leveraged deep convolutional neural networks (MD-CNN and SD-CNN) to effectively predict antibiotic resistance in Mycobacterium tuberculosis (MTB), offering improved sensitivity and specificity compared to existing methods. Additionally, [7] introduced Treesist-TB, a customized ML approach for predicting drug resistance in MTB by extracting and evaluating genomic variants across multiple studies, thereby demonstrating comparable predictive accuracy to widely used tools like TB-Profiler. In this study, we assess the efficacy of two boosting ML models, namely extreme gradient tree (XGBoost) and light gradient machine (LightGBM), alongside a multilayer deep neural network model with a novel, streamlined architecture. Our objective is to classify resistance to four first-line drugs and six second-line drugs using a comprehensive dataset comprising 28,073 Mycobacterium tuberculosis (MTB) isolates from the National Center for Biotechnological Information (NCBI). These isolates exhibit resistance phenotypes for ten antituberculosis drugs, including Isoniazid (INH), Rifampicin (RIF), Ethambutol (EMB), Pyrazinamide (PZA), Streptomycin (STR), Fluoroquinolones (OFX), Ethionamide (ETH), Kanamycin (KAN), Amikacin (AMK), and Capreomycin (CAP). Rather than employing all available features, we conducted a substudy to retrain the models and recalibrate their classification performance. This allowed us to evaluate the impact of highly influential features on classification accuracy, thereby informing the development of a lightweight system for real-time application. Our study underscores the potential of deep learning techniques in the classification of tuberculosis drug resistance. Such techniques have demonstrated remarkable performance across various domains, including computer vision, natural language processing, and speech recognition [5][19][20]. By leveraging complex architectures and large-scale training, deep neural networks enable the extraction of meaningful patterns and the generation of accurate predictions. Our findings reveal that the proposed models outperform previous learning models and achieve comparable or superior classification of drug resistance compared to direct association methods reliant solely on known resistance determinants. We present a novel method for predicting drug resistance in clinical isolates of the Mycobacterium tuberculosis complex using machine learning. Our approach offers rapid and accurate prediction of drug resistance, showcasing its potential to augment traditional drug susceptibility testing and contribute to tuberculosis control efforts. In conclusion, we discuss the clinical implications of our method and outline future research directions to further enhance its utility in combating tuberculosis.

II. MATERIALS AND METHODS

A. *Samples and laboratory phenotyping*

The dataset utilized in this study comprises 28,073 samples randomly selected from the TB-profiler database [24]. Phenotypic information for various drugs, including Isoniazid (INH), Rifampicin (RIF), Ethambutol (EMB), Pyrazinamide (PZA), Amikacin (AMK), Capreomycin (CAP), Ethionamide (ETH), Kanamycin (KAN), Fluoroquinolones (OFX), and Streptomycin (STR), has been meticulously collected and included in the dataset.

B. *DNA sequencing and pre-processing*

The Preprocessing is a pivotal stage in bioinformatics, entailing the cleansing and refining of raw sequencing data prior to downstream analysis. While the specifics of preprocessing may vary based on data type, research objectives, and analysis pipelines, it typically involves several standardized procedures. In a recent investigation into Mycobacterium tuberculosis (MTB) genomics, for instance, the preprocessing pipeline commenced with quality control checks on the raw sequencing data (SRAs) to identify potential errors like low-quality reads, adapter contamination, and sequencing biases. Subsequently, reads underwent trimming to eliminate low-quality bases, adapter sequences, and reads falling below a defined minimum length threshold. Following trimming, reads were aligned to the MTB H37Rv reference genome using BWA-MEM [12], facilitating identification of regions of similarity and divergence. Variant calling ensued, employing tools such as SAMtools [18] and BCFtools, to pinpoint single-

nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) by contrasting aligned reads with the reference genome. Resulting variants were annotated with functional data using tools like SnpSift [4]. Filtering of nucleotide bases based on sequencing and alignment quality scores followed, with a minimum score threshold of 30 from SnpSift, aimed at eliminating potential false positives arising from sequencing errors or alignment artifacts. The high-quality variants obtained from these steps were subsequently used for downstream analyses, such as phylogenetic inference, population genetics, or association studies. In summary, preprocessing in bioinformatics encompasses several standardized processes including quality control, read trimming, read mapping, and variant calling. While specific tools and parameters may vary, the overarching objective remains to obtain high-quality data for subsequent analyses. The final dataset comprised 24,429 SNPs, converted into a binary vector where 0 signifies absence and 1 presence of a mutation in the isolate. Each isolate was assigned a binary resistance/sensitivity label for each drug, obtained from TB-profiler data, with 0 indicating absence and 1 resistance. Notably, more isolates were susceptible than resistant for all drugs, with over 79% susceptible for EMB, 84% for PZA, 68% for INH, and 74% for RIF (refer to Fig. 1-A). Additionally, several isolates exhibited resistance to all four first-line drugs (refer to Fig. 1-B).

C. Feature spaces

Dimensionality reduction plays an important role in machine learning (ML), mainly for a dataset with thousands of features, such as TB data. Moreover, it has been shown to improve classification performance in many applications by reducing unimportant and redundant features [23]. Dimensionality reduction and feature selection are two different approaches used to reduce the number of features in a dataset. Dimensionality reduction techniques such as principal component analysis (PCA) and autoencoders transform the original high-dimensional feature space into a lower-dimensional space, a smaller set of uncorrelated features while preserving the variance of the data. In contrast, feature selection selects a subset of the original features based on some criterion, such as their importance or relevance in the classification task. In this study, feature selection is chosen to define subsets of features because it is a fast and effective way to reduce the number of features according to theirs inputs and in order to maintain interpretability of the features. The selected features were then used to evaluate the performance of different classifiers. In another study the second approach can be for reasons of comparison between the different approaches, but for this study, feature selection was sufficient to achieve good results. Following the work presented by [27] and to evaluate the performance of different classifiers, ten feature sets were considered:

- F1 was the base feature space of all found variants (N = 24,429).
- F2 was the basic feature space of the variants found in the 23 candidate genes (N = 14,418).
- F3 was a subset of F1 selected from a variance threshold (N = 171).
- F4 was a subset of F1 selected from the top K variables based on the $\chi 2$ test for a particular drug.
- F5 was a subset of F1 selected from the XGBoost model for a particular drug.
- F6 was a subset of F1 selected from the LightGBM model for a particular drug.
- F7 was a subset of F2 selected from a variance threshold for a particular drug (N = 120).
- F8 was a subset of F2 selected from the top K variables based on the $\chi 2$ test for a particular drug.
- F9 was a subset of F2 selected from the XGBoost model for a particular drug.
- F10 was a subset of F2 selected from the LightGBM model for a particular drug.

D. Classification methods

Two boosting ML classifiers, XGBoost [2] and LightGBM [15], were examined in this study. Additionally, a multi-layer deep neural network [3] with a novel architecture was considered. This architecture enables the network to directly learn useful rules from the input data and higher-level nonlinear features. It comprises five hidden layers, each containing 100 rectified linear units (ReLU), an output sigmoid layer, batch normalization, and L2 regularization [11] to mitigate overfitting. The network was trained using stochastic gradient descent with the Adam optimizer [8] over 100 epochs. Random initial weights were determined using He normal initialization, and the binary cross-entropy loss function was employed (see Fig. 2). Each algorithm was applied individually to a single drug, utilizing the previously mentioned feature spaces.

E. Training, testing and model evaluation

A split of the dataset was created, allocating 80% for training and 20% for testing. Evaluation metrics included the area under the receiver-operator curve (AUC), sensitivity, specificity, accuracy, precision, F1 score, and true/false positive/negative predicted values. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) were defined accordingly.

$$\textbf{\textit{Accuracy} = (TP + TN) / (TP + TN+ FP+ FN)}$$
$$\textbf{\textit{Sensitivity} = TP / (TP + FN)}$$
$$\textbf{\textit{Specificity} = TN/ (TN + FP)}$$
$$\textbf{\textit{Precision} = TP/ (TP + FP)}$$
$$\textbf{\textit{F1 – score} = 2 * (precision * sensitivity/ (precision +sensitivity))}$$

Experiments were conducted to observe the learning curve of different models, assessing potential overfitting by comparing predictions on training and test data. These steps and experiments aided in selecting the optimal learning models for subsequent optimization. Model selection for each antibiotic was guided by several criteria, prioritized as follows:

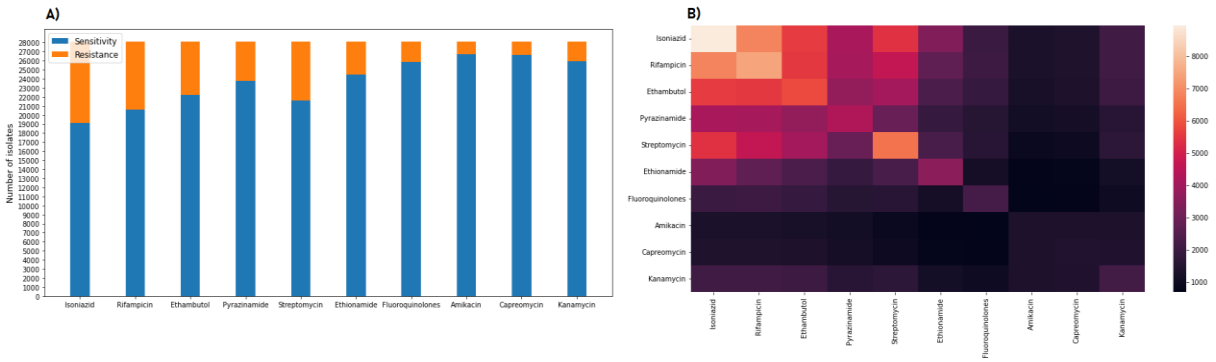

Figure 1-A. The phenotypic profile of the ten drugs; each bar shows the number of isolates that are resistant and susceptible, and 1-B the heatmap quantifying the number of cases of co-occurrence of resistance between drugs. The off-diagonal elements show the co-occurrence of resistance between the different drugs, and the diagonal elements indicate resistance to a single drug.

- Result score, particularly focusing on the best AUC and F1-score.
- Learning curve analysis to ensure generalization on test data and minimize overfitting.
- Feature space, considering the efficiency of models with fewer variables for faster execution.

To ensure the robustness of the proposed models, except for the deep learning model, a selection of hyperparameters was made and applied consistently across all bootstrap experiments. These hyperparameters were determined using Bayesian optimization, specifically Hyperopt. This strategy iteratively explores the hyperparameter space to maximize model performance on a validation set. Compared to traditional grid search methods, Bayesian optimization is more efficient and less computationally demanding. It achieves this by utilizing probabilistic models to predict model performance for given hyperparameters, focusing on promising areas of the hyperparameter space while avoiding irrelevant regions. In contrast, grid search exhaustively explores the entire hyperparameter space, often being time-consuming and computationally intensive, particularly for high-dimensional spaces. This approach facilitated the selection of hyperparameters that yielded optimal accuracy for the models. In the case of the deep learning model, batch normalization layers and L2 regularization were exclusively applied during training to prevent overfitting. Additionally, 20% of the training data was reserved for validation purposes. The training process concluded when the validation loss ceased to improve after 50 epochs.

F. Implémentation details

The deep learning model was developed using the Keras 2.2.4 library with a TensorFlow 2.1.0 backend. The boosting classifiers, XGBoost and LightGBM, were implemented using versions 1.5.2 and 3.3.2 of the xgboost and lightgbm libraries, respectively. Hyperparameter optimization was conducted using hyperopt 0.2.7. The remaining implementations were carried out using Python 3.8.12 and Scikit-Learn 1.0.2. All models were trained on hardware consisting of an i9-9900K processor (CPU) and 32 GB of RAM.

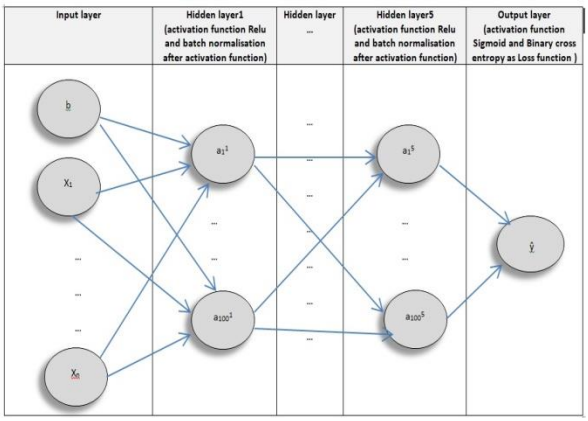

Figure 2. The new architecture of the deeplearning model used in this study

III. RESULTS

A. Phenotypes

The study encompassed 28,073 isolates of M. tuberculosis, representing all major clades of tuberculosis. Each isolate underwent liquid culture-based drug susceptibility testing for up to 10 anti-TB drugs. Fig. 1-A summarizes the phenotypes of the 28,073 MTB strains available for analysis. It illustrates the resistance and susceptibility of the isolates tested for each drug based on the phenotypic DST for the isolates tested for the different drugs. Of the total isolates tested against INH, 8960 (32%) were resistant and 19,113 (68%) were susceptible. For EMB, RIF, and PZA, the number of resistant isolates was only 5828, 7460, and 4274, representing 20%, 26%, and 15% of the total number, respectively. For second-line drugs, the percentage of resistant isolates ranged from 4% to 23% (n=1327 and n=6513), especially for STR, ETH, OFX, AMK, CAP, and KAN, the number of resistant isolates was 6513, 3591, 2180, 1327, 1414, and 2159, representing 23%, 12%, 7%, 4%, 5%, and 7% of the total number, respectively. Table I shows the co-occurrence of pairwise (off-diagonal) and mono (diagonal) resistance. A mono-resistant isolate is defined as an M. tuberculosis resistant to a single drug and susceptible to the others. It shows that INH and RIF cross-resistance is the most frequent (n=6843), representing 24% of resistant MTB. INH and EMB cross-resistance is classified as the second most frequent cross-resistance (n=5590), representing 19% of resistant MTB. INH-AMK cross-resistance was similar to that

TABLE I. SUMMARY OF PAIRWISE RESISTANCE COOCCURENCES FOR THE 10 TESTED TB DRUGS.

Drug	Isoniazid	Rifampicin	Ethambutol	Pyrazinamide	Streptomycin	Ethionamide	Fluoroquinolones	Amikacin	Capreomycin	Kanamycin
Isoniazid	8960	6843	5590	4119	5395	3391	1964	1272	1347	2102
Rifampicin	6843	7460	5508	4106	4633	2726	2002	1272	1346	2096
Ethambutol	5590	5508	5828	3749	4037	2303	1855	1227	1300	2026
Pyrazinamide	4119	4106	3749	4274	2905	1878	1541	1095	1143	1586
Streptomycin	5395	4633	4037	2905	6513	2276	1587	884	954	1668
Ethionamide	3391	2726	2303	1878	2276	3591	1137	722	778	1109
Fluoroquinolones	1964	2002	1855	1541	1587	1137	2180	683	740	982
Amikacin	1272	1272	1227	1095	884	722	683	1327	1325	1325
Capreomycin	1347	1346	1300	1143	954	778	740	1325	1414	1369
Kanamycin	2102	2096	2096	1586	1668	1109	982	1325	1369	2159

of RIF-AMK, representing 4% (n=1272) of resistant MTBs, as well as for INH-CAP which was similar to that of RIF-CAP, representing 4.5% (n=1347). For the other drugs, cases of cross-resistance were much less numerous, as the number of isolates did not exceed the percentage of 8% (n=2276).

B. Classification results

In this section, three models were evaluated and compared in terms of sensitivity, specificity, area under the receiver operating characteristic (AUROC) curve, and F1 scores. Figure 2 illustrates the comparisons of AUC performance for the three classifiers across ten feature sets and using ten drugs. All classifiers demonstrated AUC values of at least 88%. Notably, the AUC values were considerably higher for most drugs except for PZA and ETH, which achieved minimum AUCs of 88% and 93%, respectively, when considering feature set F6 for both drugs. Overall, the deep learning model outperformed the other models for all drugs in terms of AUC, achieving values between 97% and 99% for most drugs, considering feature sets F4 and F8, except for AMK and CAP. This was followed by the XGBoost model, which attained AUC values between 92% and 99% for most drugs, considering feature sets F5 and F7, except for ETH, AMK, and CAP. Lastly, the LightGBM model achieved AUC values between 88% and 98%, considering diverse feature sets (see Table II). Regarding other metrics, the three models demonstrated comparable performance characterized by F1 scores and similar specificities, with high values ranging between 96% and 99%. Sensitivities and accuracies varied across the models. XGBoost exhibited the highest sensitivity for all drugs and the best accuracy of 99%, with accuracies of 97% for AMK and CAP, respectively. However, the sensitivity, accuracy, and F1 score of all models for the drug PZA were the lowest.

C. Comparison with other studies that apply ML methods

The models proposed in this paper were compared with the results of three recent studies that applied different ML models [1][6] and [16] Specifically, the comparison focused on the maximum reported results for each parameter (sensitivity, specificity, AUC) of each drug in the three studies (see Table III). The results indicated that the deep learning model outperformed all other models for all drugs in terms of AUC, achieving values between 97% and 99%, except for OFX, where it performed slightly worse compared to the gradient boosting-based tree (GBT) algorithm study [6] (0.995 < 0.997). There were clear improvements of 2% for EMB, 4% for STR, 2% for PZA, 1% for INH, 0.5% for RIF, 9% for ETH, and 3% for AMK, CAP, KAN. The XGBoost model also showed promising results, with improvements in AUC compared to previous studies for seven drugs. It increased the AUC by 1% for INH, 1% for EMB, 4% for STR, 6% for ETH, 3% for AMK, and 2% for CAP, KAN. Similarly, the LightGBM model demonstrated enhancements in AUC for 6 drugs compared to previous studies, particularly increasing by 1% for EMB, 3% for STR, 5% for ETH, and 2% for AMK, CAP, KAN. In terms of sensitivity and specificity metrics, the three models generally exhibited better sensitivity scores compared to the other previous studies. However, for specificity, the results were similar for AMK and CAP, where the GBT algorithm study [6] and random forest (RF) [1] achieved 99%, matching the performance of the proposed models. Additionally, the logistical regression L2 (LR L2) study [1] showed better specificity compared to XGBoost and LightGBM.

IV. DISCUSSION

The development of ML techniques is particularly important for improving the prediction of TB resistance, especially in cases where the underlying biological mechanisms are less well understood. This advancement contributes to enhancing the current clinical knowledge base. Table II and Fig. 3 demonstrate that ML techniques generally yield improvements in AUC, sensitivity, and specificity compared to previous studies utilizing different learning models [1][6] and [16]. The findings of this study suggest that sensitivity holds particular importance in this application, as the failure to identify resistance can have severe consequences for patients. In contrast, classical resistance prediction

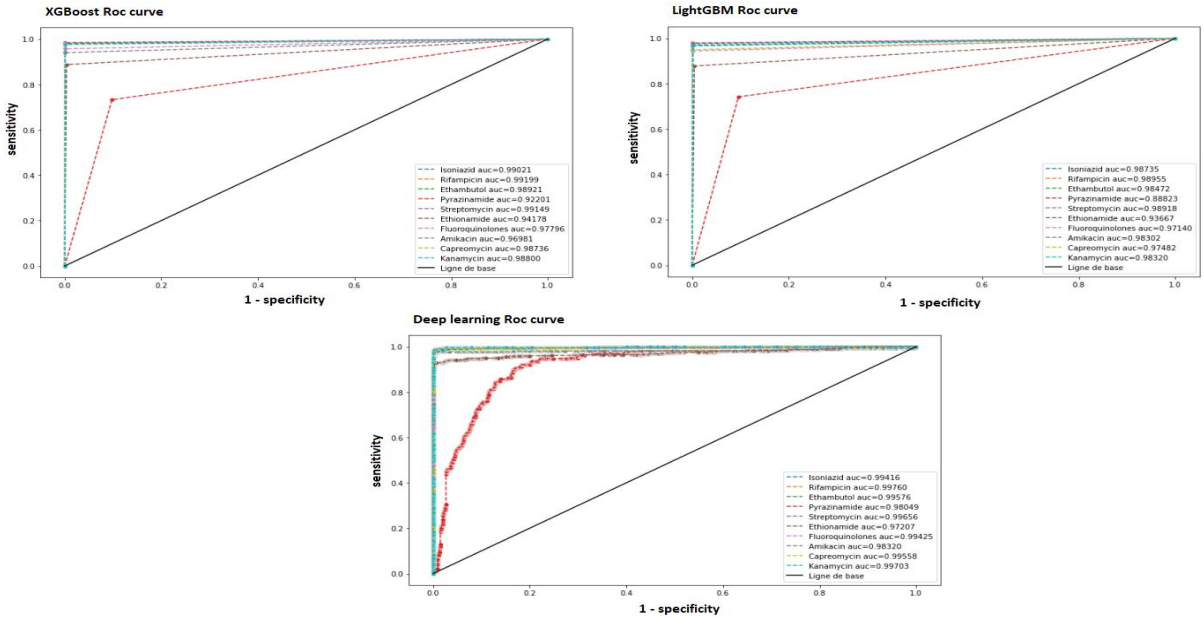

FIGURE 2. ROC PERFORMANCE CURVES OF XGBOOST, LIGHTGBM AND DEEP LEARNING FOR PREDICTING THE DRUG RESISTANCE OF PATIENTS WITH TUBERCULOSIS.

TABLE II. SENSITIVITY, SPECIFICITY, F1-SCORE AND AUC FOR THE THREE MODELS (THE MAXIMUM VALUE PER PREDICTION MEASURE IS BOLDED).

Classifier	**XGBoost**					**LightGBM**					**Deep Learning**				
Drug	***Feature set***	***Sensitivity***	***Specificity***	***F1-score***	***AUC***	***Feature set***	***Sensitivity***	***Specificity***	***F1-score***	***AUC***	***Feature set***	***Sensitivity***	***Specificity***	***F1-score***	***AUC***
Isoniazid	F7	**0.99**	0.98	**0.98**	0.990	F7	**0.99**	0.98	**0.98**	0.987	F4	**0.99**	**0.99**	**0.98**	**0.995**
Rifampicin	F5	**0.99**	**0.99**	**0.99**	0.991	F7	**0.99**	**0.99**	0.98	0.989	F4	**0.99**	**0.99**	**0.99**	**0.997**
Ethambutol	F5	**0.99**	0.98	**0.98**	0.989	F3	**0.99**	**0.99**	**0.98**	0.984	F8	**0.99**	**0.99**	**0.98**	**0.993**
Pyrazinamide	F7	**0.90**	0.96	0.85	0.922	F6	0.85	0.96	0.81	0.888	F8	0.87	**0.98**	**0.88**	**0.985**
Streptomycin	F5	**1**	**0.99**	**0.99**	0.991	F3	0.99	**0.99**	0.98	0.989	F4	0.99	**0.99**	0.98	**0.994**
Ethionamide	F4	**0.98**	0.97	**0.92**	0.941	F6	**0.98**	0.97	0.91	0.936	F4	0.93	**0.98**	**0.92**	**0.979**
Fluoroquinolones	F5	**0.99**	**0.99**	0.97	0.977	F8	0.98	**0.99**	0.96	0.971	F4	0.98	**0.99**	**0.98**	**0.995**
Amikacin	F8	**1**	**0.99**	**0.99**	0.996	F8	0.99	**0.99**	0.98	0.988	F7	0.99	**0.99**	0.98	**0.999**
Capreomycin	F8	**0.99**	**0.99**	**0.98**	0.987	F7	**0.99**	**0.99**	**0.98**	0.985	F2	0.98	**0.99**	0.97	**0.992**
Kanamycin	F7	**0.99**	**0.99**	**0.98**	0.988	F4	0.98	**0.99**	0.97	0.983	F8	0.97	**0.99**	0.97	**0.999**

methods, such as direct association (DA), tend to exhibit high specificity but may lack the sensitivity observed in ML methods. The results also indicate that the efficacy of learning methods is closely linked to the input feature space. Several factors may account for this phenomenon, including:

- The existence of mutations associated with resistance, in addition to those reported in the literature.
- The presence of combined patterns of resistance and lineage-related gene depletion.
- The co-occurrence of resistance.

These findings confirm the possibility of additional important mutations beyond those already recognized for tuberculosis resistance classification. In comparison to prior studies [1][6] and [16], the deep learning model improved AUC by 0.05% to 9% for all drugs except OFX, sensitivity and specificity by 4% to 25% and 0.02% to 16%, respectively, for all drugs. While reporting excellent performance in predicting antibiotic resistance, it's essential to assess the models' generalization capability in real-world settings. The alternative ML approaches, XGBoost and LightGBM, also demonstrated enhanced AUC, sensitivity, and specificity for most drugs when applied to all isolates, including those with rarer and previously unobserved variants. This indicates a trade-off between sensitivity and specificity, where an increase in one may result in a decrease in the other. Effective hyperparameter tuning for XGBoost and LightGBM, along with the implementation of a simple yet efficient architecture for deep learning, likely contributed to the dominance of AUC performance for the proposed models. Overall, the proposed

models exhibit strong performance and potential for real-world applications in predicting antibiotic resistance. The interpretable set of inputs used for all drugs, particularly feature sets F4 and F8, yielded significantly better performance, as they are highly correlated. Conversely, feature sets F1, F2, F3, and F6 showed lower prediction performances, mainly due to either the very low number of insignificant variants or the presence of a large number of complex variants lacking generalization. In conclusion, the application of machine learning (ML) techniques in predicting drug resistance in Mycobacterium tuberculosis offers several significant advantages:

- Enhanced accuracy: ML techniques demonstrate superior predictive capabilities compared to traditional methods by leveraging extensive datasets, resulting in more precise predictions.
- Expedited prediction times: ML techniques excel in processing large volumes of data rapidly, enabling healthcare professionals to make timely decisions regarding optimal treatment strategies for patients.
- Improved identification of unknown factors: ML techniques have the potential to identify novel mutations or patterns associated with drug resistance, even in cases where such associations have not been previously documented or studied.
- Personalized treatment plans: ML techniques can aid physicians in tailoring treatment plans based on individual patients' unique genetic profiles and drug resistance characteristics.
- Cost-effectiveness: ML techniques contribute to cost reduction in the treatment of drug-resistant tuberculosis by identifying the most effective drugs and minimizing the need for expensive and potentially harmful trial-and-error approaches.

However, this study had limitations. Reliable prediction of genotypic resistance depends on the quality of raw sequencing data. Variants and small indels in resistance-conferring genes can be accurately called from Illumina raw sequence data if the genes are adequately covered at an acceptable sequencing depth. Additionally, the limited availability of phenotypic resistance data for recently introduced or repurposed drugs, such as bedaquiline or linezolid, hindered the prediction of resistance to these drugs. Furthermore, the study did not address the prediction of multidrug resistance. Attempts to create a multi-label deep learning model were unsuccessful due to several assumptions, including the large number of variants impacting the prediction, inadequacies in the loss function, and challenges in model generalization. Solutions such as dimensionality reduction techniques like autoencoders, customization of loss functions, and the use of overfitting resolution techniques like dropout may address these challenges in future research. Furthermore, it is imperative to acknowledge additional significant limitations:

- Sample size: The study's generalizability may be compromised by a small sample size. Larger, more diverse datasets with comprehensive genomic data are essential to enhance the accuracy of resistance prediction models.
- Data quality: The reliability of predictions made by ML techniques heavily depends on the quality of the data used for model training. Incomplete or biased data can lead to inaccurate predictions.
- Lack of clinical validation: The study solely relied on genomic data for resistance prediction, lacking validation from clinical data. Integrating clinical and genomic data could improve the accuracy of resistance prediction models.
- Platform bias: The study predominantly focused on the Illumina platform, overlooking other sequencing platforms such as PacBio and Nanopore, which could offer additional insights into the accuracy of resistance prediction tools. Platform bias may limit the applicability of the findings to other platforms.
- Technical expertise: The development and implementation of ML techniques require specialized knowledge and technical expertise. Without adequate training, healthcare professionals may encounter challenges in effectively utilizing these tools.
- Ethical considerations: The use of ML techniques for drug resistance prediction raises ethical concerns regarding privacy, data ownership, as well as potential biases and discrimination against certain groups.

Moreover, it is crucial to consider potential implications and limitations in clinical settings:

Potential implications:

- Early detection: ML techniques could facilitate earlier detection of drug-resistant TB, leading to prompt treatment and improved patient outcomes.
- Identification of new patterns: ML models may uncover new mutations or patterns of resistance, offering insights into TB biology and potential drug targets.
- Treatment optimization: ML techniques could optimize treatment regimens by predicting effective drugs for individual patients, thus reducing the use of ineffective or unnecessary medications.

Limitations:

- Resource constraints: Implementing ML models in resource-limited settings may be challenging due to limited access to computing resources and specialized expertise.
- Limited prediction accuracy: ML models may struggle to accurately predict resistance for all drugs or TB strains and may require continuous updating as new mutations emerge.
- Bias and overfitting: ML models may exhibit bias or overfitting if trained on non-representative datasets, leading to inaccurate predictions and potentially harmful treatment decisions.

TABLE III. COMPARISON OF THE PERFORMANCE OF THE PROPOSED CLASSIFIERS WITH THE BEST MODELS STUDIED IN PREVIOUS STUDIES IN TERMS OF SENSITIVITY, SPECIFICITY AND AUC (CELLS ARE COLORED FROM LIGHTEST TO DARKEST FOR THE LOWEST TO HIGHEST PERFORMANCE ACROSS THE 10 DRUGS FOR EACH MODEL, INCLUDING ORANGE FOR AUC, BLUE FOR SENSITIVITY AND GREEN FOR SPECIFICITY).

	XGBoost				LightGBM				Deep Learning				Best models (previous classifiers)			
Drug	***Feature set***	***Sen***	***Spec***	***AUC***	***Feature set***	***Sen***	***Spec***	***AUC***	***Feature set***	***Sen***	***Spec***	***AUC***	***classifier***	***Sen***	***Spec***	***AUC***
INH	F7	0.99	0.98	0.990	F7	0.99	0.98	0.987	F4	0.99	0.99	0.995	logistic regression L2 (LR-L2)	0.89	0.99	0.989
RIF	F5	0.99	0.99	0.991	F7	0.99	0.99	0.989	F4	0.99	0.99	0.997	Deep learning	0.95	0.97	0.994
EMB	F5	0.99	0.98	0.989	F3	0.99	0.99	0.984	F8	0.99	0.99	0.993	logistic regression L2 (LR-L2)	0.93	0.83	0.977
PZA	F7	0.90	0.96	0.922	F6	0.85	0.96	0.888	F8	0.87	0.98	0.985	Deep learning	0.75	0.91	0.961
STR	F5	1	0.99	0.991	F3	0.99	0.99	0.989	F4	0.99	0.99	0.994	logistic regression L2 (LR-L2)	0.87	0.94	0.951
ETH	F4	0.98	0.97	0.941	F6	0.98	0.97	0.936	F4	0.93	0.98	0.979	gradient boosting-based tree (GBT)	0.68	0.93	0.884
OFX	F5	0.99	0.99	0.977	F8	0.98	0.99	0.971	F4	0.98	0.99	0.995	gradient boosting-based tree (GBT)	0.85	0.98	0.997
AMK	F8	1	0.99	0.996	F8	0.99	0.99	0.985	F7	0.99	0.99	0.999	gradient boosting-based tree (GBT)	0.80	0.99	0.964
CAP	F8	0.99	0.99	0.987	F7	0.99	0.99	0.985	F2	0.98	0.99	0.992	Random forest (RF)	0.89	0.99	0.966
KAN	F7	0.99	0.99	0.988	F4	0.98	0.99	0.983	F8	0.97	0.99	0.999	gradient boosting-based tree (GBT)	0.82	0.98	0.968

For readers or practitioners aiming to replicate the methods outlined in this study, several suggestions and considerations should be taken into account. Firstly, the quality of raw sequencing data is paramount for reliable prediction of genotypic resistance. Variants and small indels within resistance-conferring genes can be accurately identified from Illumina raw sequence data if the genes are adequately covered at a sufficient sequencing depth. Secondly, careful consideration of the input feature space is essential, as the efficacy of learning methods is closely tied to it. Factors such as the presence of mutations associated with resistance, combined resistance patterns, lineage-related gene depletion, and co-occurrence of resistance can influence the performance of different input feature spaces. Thirdly, it's crucial to acknowledge the trade-off between sensitivity and specificity. Increasing sensitivity may come at the expense of specificity and vice versa. Therefore, finding the right balance between these two metrics is critical for the success of the predictive model. Finally, the models proposed in this study have demonstrated the robustness of machine learning in predicting drug resistance and identifying underlying mutations. As whole-genome sequencing becomes more commonplace and the need for "big data" analyses grows, these approaches offer scalability and promise for future applications in this field.

V. CONCLUSION

The exploration of XGBoost, LightGBM, and Deep Learning classifiers for TB resistance classification, considering various subsets, has underscored the primary advantage of machine learning algorithms, particularly the approach outlined in this study with a comprehensive feature set. This advantage lies in their ability to discern associations between feature space and resistance prediction, thereby illuminating potential novel drug-associated mutations. Consequently, these developed techniques have significantly enhanced the classification of resistance using genetic data, showcasing their potential in analyzing large, high-dimensional datasets. This potential is especially valuable in scenarios where the underlying biological mechanisms of resistance remain poorly understood for many drugs. The utilization of the most promising model identified in this research (deep learning) for predicting MTB resistance holds promise for improving patient outcomes and mitigating the risk of developing multidrug resistance. However, further investigation is warranted in future studies. This could entail exploring novel concepts such as dimension reduction and devising a deep learning architecture tailored for the prediction of multidrug resistance. Such endeavors would contribute to advancing our understanding and management of drug-resistant TB, ultimately benefiting patient care and public health outcomes.

REFERENCES

[1] Chen, Michael L et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. EBioMedicine, 2019, vol. 43, p. 356-369.

[2] Chen, Tianqi and Guestrin, Carlos. Xgboost: A scalable tree boosting system. In : Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

[3] Cheng, Heng-Tze et al. Wide & deep learning for recommender systems. In : Proceedings of the 1st workshop on deep learning for recommender systems. 2016. p. 7-10.

[4] Cingolani, Pablo et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in genetics, 2012, vol. 3, p. 35.

[5] Coppersmith, Glen and Leary, Ryan and Crutchley, Patrick and Fine, Alex. Natural language processing of social media as screening for suicide risk. Biomedical informatics insights, 2018, vol. 10, p. 1178222618792860.

[6] Deelder, Wouter et al. Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. Frontiers in genetics, 2019, vol. 10, p. 922.

[7] Deelder, Wouter and Napier, Gary and Campino, Susana and Palla, Luigi and Phelan, Jody and Clark, Taane G. A modified decision tree approach to improve the prediction and mutation discovery for drug

resistance in Mycobacterium tuberculosis. BMC genomics, 2022, vol. 23, no 1, p. 46.

[8] Diederik, P Kingma. Adam: A method for stochastic optimization. (No Title), 2014.

[9] Farhat, Maha R et al. Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value. American journal of respiratory and critical care medicine, 2016, vol. 194, no 5, p. 621-630.

[10] Green, Anna G et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. Nature communications, 2022, vol. 13, no 1, p. 3817.

[11] Glorot, Xavier and Bordes, Antoine and Bengio, Yoshua. Deep sparse rectifier neural networks. In : Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011. p. 315-323.

[12] Houtgast, Ernst Joachim and Sima, Vlad-Mihai and Bertels, Koen and Al-Ars, Zaid. Hardware acceleration of BWA-MEM genomic short read mapping for longer read lengths. Computational biology and chemistry, 2018, vol. 75, p. 54-64.

[13] Jiang, Zhonghua et al. Drug resistance prediction and resistance genes identification in Mycobacterium tuberculosis based on a hierarchical attentive neural network utilizing genome-wide variants. Briefings in Bioinformatics, 2022, vol. 23, no 3, p. bbac041.

[14] Jnawali, Hum Nath and Ryoo, Sungweon. First-and second-line drugs and drug resistance. Tuberculosis-current issues in diagnosis and management, 2013, vol. 20, p. 163-80.

[15] Ke, Guolin et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, vol. 30.

[16] Kouchaki, Samaneh et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics, 2019, vol. 35, no 13, p. 2276-2282.

[17] Kuang, Xingyan and Wang, Fan and Hernandez, Kyle M and Zhang, Zhenyu and Grossman, Robert L. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. Scientific reports, 2022, vol. 12, no 1, p. 2427.

[18] Li, Heng et al. The sequence alignment/map format and SAMtools. bioinformatics, 2009, vol. 25, no 16, p. 2078-2079.

[19] Pawar, Karishma and Attar, Vahida. Deep learning based detection and localization of road accidents from traffic surveillance videos. ICT Express, 2022, vol. 8, no 3, p. 379-387.

[20] Pitropakis, Nikolaos and Kokot, Kamil and Gkatzia, Dimitra and Ludwiniak, Robert and Mylonas, Alexios and Kandias, Miltiadis. Monitoring users' behavior: Anti-immigration speech detection on Twitter. Machine Learning and Knowledge Extraction, 2020, vol. 2, no 3, p. 11.

[21] Quan, T Phuong el al. Evaluation of whole-genome sequencing for mycobacterial species identification and drug susceptibility testing in a clinical setting: a large-scale prospective assessment of performance against line probe assays and phenotyping. Journal of clinical microbiology, 2018, vol. 56, no 2, p. 10.1128/jcm. 01480-17.

[22] Safi, Hassan el al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes. Nature genetics, 2013, vol. 45, no 10, p. 1190-1197.

[23] Schleusener, Viola and Koser, Claudio U and Beckert, Patrick and Niemann, Stefan and Feuerriegel, Silke. Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Scientific reports, 2017, vol. 7, no 1, p. 1-9.

[24] Verboven, Lennert and Phelan, Jody and Heupink, Tim H and Van Rie, Annelies. TBProfiler for automated calling of the association with drug resistance of variants in Mycobacterium tuberculosis. Plos one, 2022, vol. 17, no 12, p. e0279644.

[25] WHO, Global Tuberculosis Report 2016. World Health Organization. 2016.

[26] WHO, Global Tuberculosis Report 2020. World Health Organization. 2020.

[27] Yang, Yang el al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics, 2018, vol. 34, no 10, p. 1666-1671.

[28] Zhang, Hongtai el al. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nature genetics, 2013, vol. 45, no 10, p. 1255-1260.