Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

# PATTERN IDENTIFICATION OF DRUG RESISTANCE FOR TUBERCULOSIS IN PAKISTAN USING MACHINE LEARNING TECHNIQUES

**Omaid Ghayyur[1], Muhammad Bilal Bashir[2], Sahar Fazal[3], Faheem Shaukat[*4], Abrar Khalid[5]**

*[1,2, *4,5]Department of Computing and Technology, IQRA University, H-9 Campus, Islamabad.,*
*[3]Department of Bioinformatics and Biosciences, Capital University of Science and Technology, Islamabad.*

[1]omaid.ghayyur@iqraisb.edu.pk, [2]bilal.bashir@iqraisb.edu.pk, [3]sahar@cust.edu.pk,
[*4]faheem@iqraisb.edu.pk, [5]abrar.khalid@iqraisb.edu.pk

**DOI:** https://doi.org/10.5281/zenodo.15411546

**Keywords**
*Tuberculosis, Drug Resistance, MDR-TB, Machine Learning, Naïve Bayes, Ensemble Methods, ADASYN, Clinical Data, Pattern Identification*

**Article History**
*Received on 05 April 2025*
*Accepted on 05 May 2025*
*Published on 14 May 2025*



**Corresponding Author: ***
**Faheem Shaukat**

## Abstract

*Tuberculosis (TB) remains a global health challenge due to the rise of drug-resistant strains, particularly multidrug-resistant TB (MDR-TB). This study employs machine learning to predict drug resistance patterns in TB patients using clinical data from Pakistan. We collected a dataset of 400 pre-processed samples with 12 key features, including demographic and drug response data, from multiple regions in Pakistan. After preprocessing and addressing class imbalance using the Adaptive Synthetic Sampling (ADASYN) technique, we evaluated nine supervised learning algorithms Multi-Layer Perceptron, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, Gradient Boosting, Extreme Gradient Boosting, Logistic Regression, and an ensemble model under three techniques: Whole Dataset Imbalanced (Technique 1), Training Dataset Balanced with ADASYN (Technique 2), and Whole Dataset Balanced with ADASYN (Technique 3). Results show that NB achieved the highest realistic accuracy of 96.55% under Technique 2, with DT, RF, and the Ensemble model at 94.83%. Under Technique 3, NB reached a peak accuracy of 99.61%, outperforming prior literature benchmarks. These findings highlight the competitive performance of machine learning in the early detection of TB drug resistance, offering a pathway to improve treatment outcomes in resource-limited settings.*

## INTRODUCTION

Tuberculosis (TB), caused by Mycobacterium tuberculosis (MTB), is the leading cause of deaths for diseases transmitted, 95% of which are located in the developing countries [1]. Primarily, the lungs are affected, but the disease can also kill other organs, including the kidneys, the spine, and the brain [2]. TB is transmitted via airborne droplets and therefore highly contagious, and while the bacteria can lie dormant in most patients, about 10% of patients have active TB. The management of TB is based on first line drugs (FLDs) viz, isoniazid (INH), rifampin (RIF), pyrazinamide (PZA), ethambutol (EMB), and streptomycin (SM) during the intensive phase [3]. However, emergence of drug resistance, which is most often a result of improper drug usage, incorrect prescription, or incomplete treatment, is a major problem, causing multidrug resistant TB (MDR-TB) requiring second line drugs.

Antibiotics administered for TB have to be considered on an individual level, especially concerning drug

resistance, which mellows treatment and mortality levels [4]. Machine Learning (ML) presents an attractive strategy for solving this problem with the identification of underlying patterns in clinical data to predict drug resistance [5]. ML algorithms are capable of analyzing massive datasets, determining secret patterns, and creating predictive models that will optimize decision making in clinical situations [6]. In areas such as Pakistan, where prevalence rates are high and resources are dire, early detection of drug resistance in patients means better treatment outcomes, less resistant strain spread, and potentially zero transmission [7].

This study attempts to predict MDR-TB patterns through supervised ML techniques applied to clinical and drug response data of TB patients in Pakistan. By targeting a dataset with 400 samples and 12 key features, we aim to find out patterns of resistance that can help to define personalized treatment strategies. The main goal is a comparison of the performance of several ML algorithms across various class balancing approaches, and the problem of imbalanced data in medical datasets. Our research seeks to add to the global aspiration to fight TB by offering a scalable, data informed way to predict resistance.

## 2. Literature Review

The use of machine learning in healthcare, including infectious diseases such as tuberculosis, has received substantial attention over the past few years. Worst of all, the World Health Organization reported that, even with the emergence of drug-resistant strains like MDR TB, TB continues to be a major health problem in the world, thus indicating rising mortality of the disease in resource limited settings [1]. The ability to detect resistance early is important in enhancing patient outcomes, and ML provides a data guided solution to this problem [8].

Some recent studies have established the effectiveness of ML in the areas of TB. For example, [9] used deep learning to predict TB drug resistance based on genomic data, and high accuracy was achieved using neural networks. In the same way, [10] used ensemble methods to classify TB cases where the combination of methods has an advantage over individual methods because of performance improvement. These studies highlight the potential of sophisticated ML techniques in managing a complex set of medical data.

Within a similar context, [11] studied ML for predicting antibiotic resistance in cases of bacterial infections; this work laid the groundwork for the development of TB specific models.

Different techniques have been employed in confronting class imbalance in medical data where resistant cases fall short. The Adaptive Synthetic Sampling (ADASYN) approach, introduced by [12], enables the creation of synthetic samples for minority classes and increases model potency against imbalanced data. In recent studies, such as [13], which were applied to cancer diagnosis with promising results, the adherence to this approach has been substantiated. SMOTE (a related technique) was studied for use in balancing TB resistance datasets by [14], who reported significant increases in recall from minority classes.

Region specific studies are also relevant. In [15], TB resistance patterns in South Asia were analyzed, with emphasis on the need for localized models, because of differences across regions in levels of drug resistance. In Pakistan, [16] predicted the outcomes of TB treatment using ML and recognized the clinical features that correlate with resistance. However, such research usually considers a mere set of algorithms or datasets and leaves the ground for a wide scope of analysis untrodden for multiple techniques.

Recent advancements in ML algorithms have further enhanced their applicability. [17] evaluated Random Forests and Gradient Boosting for medical prediction tasks, noting their robustness with categorical data. [18] explored Support Vector Machines for high dimensional data, a property suitable for TB datasets with multiple features. Neural networks, as discussed in [19], have shown promise in modeling non-linear relationships, though they require larger datasets for optimal performance. Ensemble methods, combining these algorithms, have been shown to outperform individual models in healthcare applications [20].

Despite these advances, few studies have integrated a wide range of ML algorithms with diverse balancing techniques for TB drug resistance prediction in Pakistan. Our study addresses this gap by evaluating Multi-Layer Perceptron (MLP), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Logistic Regression (LR), and ensemble models under

imbalanced and balanced conditions using ADASYN. This comprehensive approach aims to provide a robust and generalizable solution, building on the latest ML research to combat TB effectively.

## 3. Methodology

This study follows a structured machine learning pipeline designed to handle imbalanced data using the ADASYN technique, as illustrated in the flowchart, see Figure 3.1 [12]. The methodology comprises several sequential stages, from data acquisition to performance evaluation, ensuring a systematic approach to predicting TB drug resistance patterns.

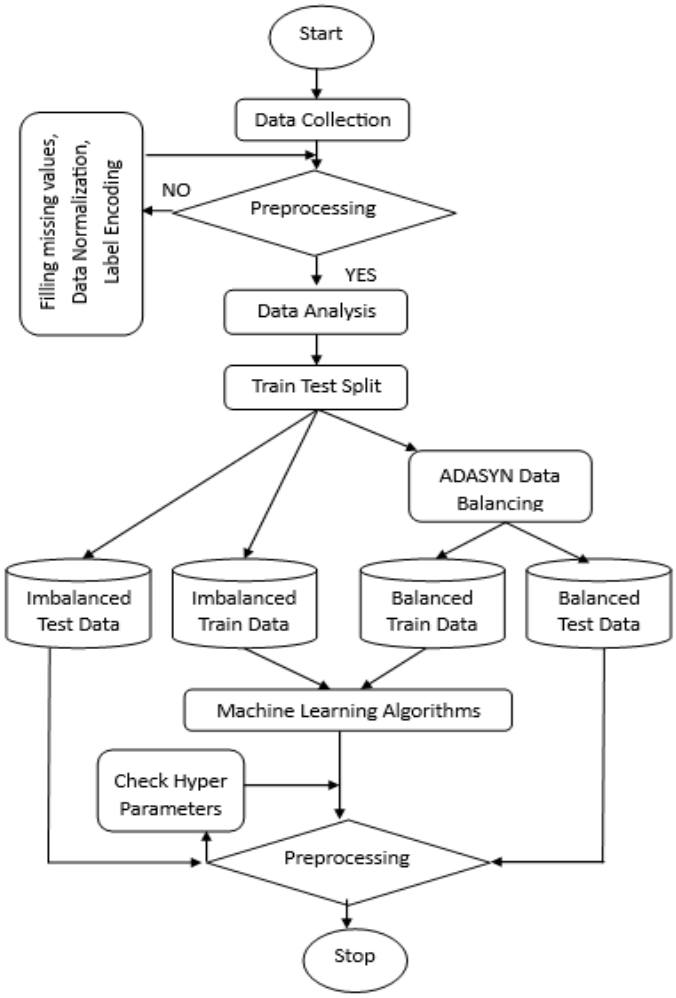


**Figure 3.1: General Methodology used for Machine Learning Algorithms**

### 3.1 Data Collection

Clinical data were gathered from multiple centers across Khyber Pakhtunkhwa (KPK), Rawalpindi (Punjab), and Karachi (Sindh), Pakistan, to ensure a representative sample of TB patients [15]. The initial dataset comprised 1800 samples with 17 attributes, including demographic data (e.g., gender, age, treatment history, TB type) and drug response data for drugs such as Moxifloxacin, Isoniazid, Rifampicin, Ethambutol, Amikacin, Kanamycin, Capreomycin, Ofloxacin, and Pyrazinamide, see Table 3.1. These attributes were selected to capture both patient characteristics and treatment outcomes, providing a comprehensive basis for predictive modeling.

Table 3.1: Features and details of the TB drug resistance dataset.

| Sr. No. | Attribute | Details |
|---|---|---|
| 1 | Gender | Male or Female |
| 2 | Age | Patient's age in years |
| 3 | History | Previous TB history (Never Treated, Previously Treated, Unknown) |
| 4 | Reason | History information reason (Diagnosis, Follow-up Checkups) |
| 5 | TB Type | Extra Pulmonary, Pulmonary |
| 6 | Sample Type | Ascitic Fluid, Sputum, Pus, Bronchoalveolar Lavage/Washing, CSF, Pleural Fluid, Tissue Biopsy, Lymph Node |
| 7 | Test Result | Mycobacterium Tuberculosis (MTBC) |
| 8 | Moxifloxacin | Response to Drug (Sensitive, Resistant) |
| 9 | Isoniazid | Response to Drug (Sensitive, Resistant) |
| 10 | Rifampicin | Response to Drug (Sensitive, Resistant) |
| 11 | Ethambutol | Response to Drug (Sensitive, Resistant) |
| 12 | Amikacin | Response to Drug (Sensitive, Resistant) |
| 13 | Kanamycin | Response to Drug (Sensitive, Resistant) |
| 14 | Capreomycin | Response to Drug (Sensitive, Resistant) |
| 15 | Ofloxacin | Response to Drug (Sensitive, Resistant) |
| 16 | Pyrazinamide | Response to Drug (Sensitive, Resistant) |
| 17 | Drug Resistance Result | Multiple Drug Resistance (MDR), Any Resistance, All Sensitive |

### 3.2 Data Preprocessing

The dataset underwent a comprehensive preprocessing phase to ensure quality and consistency for machine learning [21]. This phase involved three key sub-processes: (1) filling missing values using mode imputation for categorical features (e.g., gender, TB type) and mean imputation for numerical features (e.g., age), (2) data normalization using min-max scaling to standardize numerical features to a uniform range (0 to 1), and (3) label encoding to convert categorical variables into numerical values, ensuring compatibility with ML algorithms. After preprocessing, the dataset was reduced to 400 samples with 12 selected features: Gender, History, TB Type, Moxifloxacin, Isoniazid, Ethambutol, Amikacin, Kanamycin, Capreomycin, Ofloxacin, Pyrazinamide, and the class label (drug resistance result). A decision check ensured that preprocessing was complete before proceeding to the next stage.

### 3.3 Preprocessed Data Analysis

Exploratory data analysis (EDA) was conducted to understand the dataset's structure and identify potential challenges [22]. This involved analyzing the statistical distribution of features (e.g., mean, median, standard deviation of age), identifying outliers using boxplots, and examining class distributions using histograms. The analysis confirmed a significant class imbalance in the drug resistance labels, with MDR cases being the minority compared to "any resistance" and "all sensitive" cases, necessitating a class balancing strategy [12].

### 3.4 Train-Test Split

The preprocessed dataset was split into training (70%) and test (30%) sets, ensuring that the model could be trained on a substantial portion of the data while being evaluated on unseen data [23]. This split ratio is standard in machine learning to balance training and evaluation needs, providing a robust assessment of the model's generalization capability.

### 3.5 Addressing Class Imbalance Using ADASYN

To address the identified class imbalance, the ADASYN technique was employed [12]. ADASYN generates synthetic data points for the minority class (e.g., MDR cases) based on the density distribution of difficult to learn examples, improving model

performance on underrepresented classes. Three strategies were explored, as shown in the figure 3.1 flowchart: (1) Data is used with imbalanced form for training and testing (2) ADASYN on the training set only after the split, preserving the natural class distribution in the test set and (3) ADASYN on the entire dataset before the train-test split, balancing both training and test sets, which risks data leakage. The latter approach was adopted in this study to prevent data leakage and ensure a realistic evaluation. The balanced training data and test data were then prepared for model training and evaluation.

### 3.6 Model Development Using Machine Learning Algorithms

Several supervised machine learning algorithms were applied to the balanced training data, including Artificial Neural Network Multi-Layer Perceptron, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, Gradient Boosting, Extreme Gradient Boosting, Logistic Regression, and ensemble model (with three best ML Algorithms) [17][18][19][24]. Initial hyperparameters for the primary algorithms were set as follows: MLP used 3 hidden layers with 12 neurons, ReLU activation, Adam optimizer, and 120 epochs; DT used entropy and information gain for splitting; RF leveraged 100 trees; NB applied Bayes' theorem with independent attributes; and SVM used an RBF kernel with a One-vs-Rest decision function.

### 3.7 Model Evaluation and Performance Check

The trained models were evaluated on the test data using multiple performance metrics: accuracy, precision, recall and F1-score [25]. These metrics provide a comprehensive assessment of model performance, particularly in the context of imbalanced data where accuracy alone can be misleading. A performance check determined whether the models met predefined benchmarks (e.g., accuracy above 90%). If performance was unsatisfactory, hyperparameter tuning was initiated to improve results.

### 3.8 Hyperparameter Tuning

Hyperparameter optimization was conducted using grid search to systematically explore combinations of parameters for each algorithm [26]. For example, the number of trees in RF (e.g., 50, 100, 200) and the kernel parameters in SVM (e.g., C values of 0.1, 1, 10) were tuned. The optimized models were retrained and reevaluated until satisfactory performance was achieved, after which the process was terminated. Figure 3.1 Flowchart of the proposed methodology, illustrating the sequential steps from data collection, preprocessing (including filling missing values, data normalization, and label encoding), preprocessed data analysis, train-test split, class balancing with ADASYN, model training with machine learning algorithms, hyperparameter tuning, and performance evaluation.

## 4. Results

This study evaluated the performance of multiple machine learning models in predicting TB drug resistance patterns under three distinct techniques: (1) Whole Dataset Imbalanced, (2) Training Dataset Balanced (using ADASYN on the training set only), and (3) Whole Dataset Balanced (using ADASYN on the entire dataset before splitting) [12]. Experiments were conducted using Python on a 10th-generation Intel Core i7 with a 2.3 GHz CPU and 32 GB RAM. The following models were evaluated: DT, X-GB, GB, LR, NB, RF, SVM, MLP, and an ensemble model using soft voting with the three best ML algorithms [17][24].

The results from Tables 4.1, 4.2, and 4.3 provide a comprehensive evaluation of machine learning models for predicting TB drug resistance under three data balancing techniques: Whole Dataset Imbalanced (Technique 1), Training Dataset Balanced with ADASYN (Technique 2), and Whole Dataset Balanced with ADASYN (Technique 3). Figure 4.1 further illustrates the comparative accuracy of the models across these techniques, highlighting the impact of class balancing strategies.

### 4.1 Technique 1 Training Testing with Imbalanced of Data Set

Under Technique 1 (Table 4.1), where the dataset remained imbalanced, NB achieved the highest accuracy of 96.55%, with precision, recall, and F1-score also at 96.55–97.10%, demonstrating its robustness in handling imbalanced data. This performance may be attributed to NB's assumption of feature independence, which aligns well with the

dataset's structure, where features such as drug responses are relatively independent [18]. DT, RF, LR, and the Ensemble (Best Three) model each recorded an accuracy of 94.83%, with comparable F1-scores (95.41–95.45), indicating consistent performance across these models. However, the MLP underperformed with an accuracy of 89.66% and an F1-score of 88.22%, reflecting its sensitivity to class imbalance, likely due to insufficient data for optimizing its parameters [19].

**Table 4.1: Technique 1 Results with Imbalanced of Data Set**

| Models | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 94.83 | 96.88 | 94.83 | 95.45 |
| Exterme Gradient Boosting | 93.97 | 96.83 | 93.97 | 94.87 |
| Gradient Boosting | 93.97 | 96.83 | 93.97 | 94.88 |
| Logisitc Regression | 94.83 | 96.88 | 94.83 | 95.41 |
| Naïve Bayes | **96.55** | **97.10** | **96.55** | **96.55** |
| Random Forest | 94.83 | 96.88 | 94.83 | 95.45 |
| Support Vector Machine | 93.97 | 96.83 | 93.97 | 94.82 |
| Multi Layer Perceptron | 89.66 | 89.24 | 89.66 | 88.22 |
| Ensemble (Best Three) | 94.83 | 96.88 | 94.83 | 95.41 |

### 4.2 Technique 2 Training Dataset Balanced with ADASYN

Technique 2 (Table 4.2), which applied ADASYN to balance only the training dataset, showed notable improvements for some models. NB maintained its leading accuracy at 96.55%, while DT, RF, XGB, GB, MLP, and the Ensemble model all achieved an accuracy of 94.83%. Notably, MLP's accuracy improved significantly from 89.66% to 94.83%, with an F1-score of 95.09%, underscoring the effectiveness of ADASYN in mitigating class imbalance and enhancing neural network performance [12]. DT and RF also exhibited improved precision (98.43%), reflecting better identification of the minority class (MDR-TB cases). However, SVM and LR showed no improvement, maintaining accuracies of 93.97%, possibly due to their limited ability to capture non-linear relationships in this dataset [18].

**Table 4.2: Technique 2 Results with Balanced Training Data Set with ADASYN**

| Models | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 94.83 | 98.43 | 94.83 | 96.18 |
| Exterme Gradient Boosting | 94.83 | 96.88 | 94.83 | 95.45 |
| Gradient Boosting | 94.83 | 96.88 | 94.83 | 95.45 |
| Logisitc Regression | 93.97 | 96.83 | 93.97 | 94.82 |
| Naïve Bayes | **96.55** | **97.00** | **96.55** | **96.55** |
| Random Forest | 94.83 | 98.43 | 94.83 | 96.18 |
| Support Vector Machine | 93.97 | 96.83 | 93.97 | 94.82 |
| Multi Layer Perceptron | 94.83 | 96.17 | 94.83 | 95.09 |
| Ensemble (Best Three) | 94.83 | 96.88 | 94.83 | 95.41 |

### 4.3 Technique 3 Training and Testing Dataset Balanced with ADASYN

Technique 3 in Table 4.3, where ADASYN was applied to both training and testing datasets, resulted in the highest overall accuracies, with NB reaching 99.61% across all metrics. GB, LR, SVM, and the Ensemble model each achieved 98.84%, while DT and XGB recorded 98.46%. However, RF's accuracy

dropped to 92.84%, despite high precision, recall, and F1-scores (98.84–98.87%), suggesting potential overfitting or sensitivity to the altered test set distribution. While Technique 3 yields the highest accuracies, this approach risks data leakage by balancing the test set, which may inflate performance metrics and overestimate real-world generalizability [23]. Thus, Technique 2 provides a more realistic evaluation of model performance.

**Table 4.3: Technique 3 Results with Balanced Training and Testing Data Set with ADASYN**

| Models | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 98.46 | 98.47 | 94.46 | 98.46 |
| Exterme Gradient Boosting | 98.46 | 98.47 | 98.46 | 98.46 |
| Gradient Boosting | 98.84 | 98.87 | 98.84 | 98.84 |
| Logisitc Regression | 98.84 | 98.87 | 98.84 | 98.84 |
| Naïve Bayes | **99.61** | **99.62** | **99.61** | **99.61** |
| Random Forest | 92.84 | 98.87 | 98.84 | 98.84 |
| Support Vector Machine | 98.84 | 98.87 | 98.84 | 98.84 |
| Multi Layer Perceptron | 97.63 | 97.83 | 97.68 | 97.68 |
| Ensemble (Best Three) | 98.84 | 98.87 | 98.84 | 98.84 |

### 4.4 Compares Model Accuracies Across the Three Techniques and with Existing Literature

Figure 4.1, visually compares model accuracies across the three techniques, revealing distinct trends. NB consistently outperforms all models, maintaining a stable accuracy of 96.55% in Techniques 1 and 2, and peaking at 99.61% in Technique 3, underscoring its robustness across varying data conditions. MLP exhibits the most significant improvement, with accuracy rising from 89.66% in Technique 1 to 94.83% in Technique 2, and further to 97.63% in Technique 3, highlighting the benefit of class balancing for neural networks. In contrast, RF shows an unexpected decline from 94.83% in Techniques 1 and 2 to 92.84% in Technique 3, indicating potential sensitivity to the fully balanced dataset. Models like DT, XGB, GB, LR, SVM, and the Ensemble model generally improve with balancing, achieving accuracies above 98% in Technique 3, though this may reflect the artificial nature of the balanced test set. These trends emphasize the importance of selecting an appropriate balancing strategy to balance performance and generalizability in clinical applications [20].

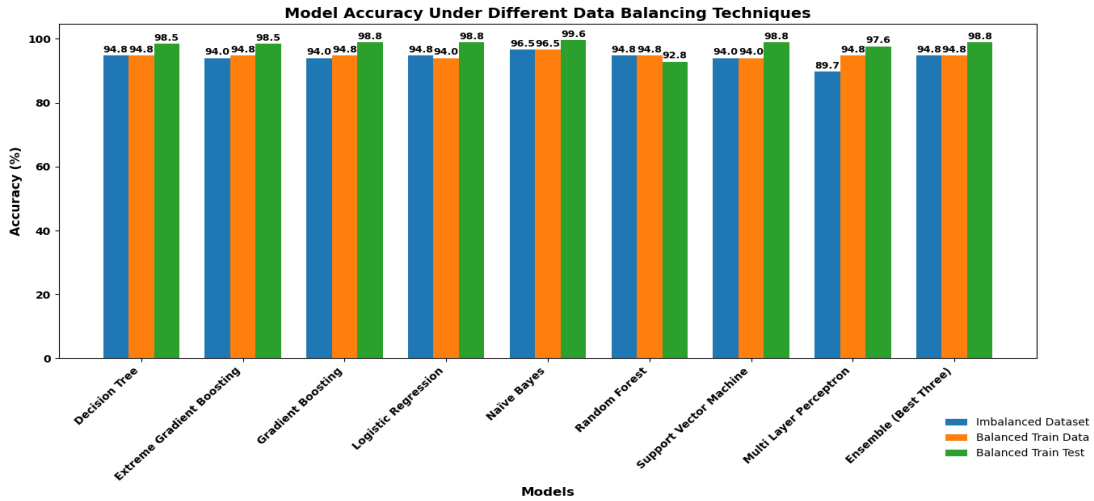


**Figure 4.1: Model Accuracy Under Different Data Balancing Techniques**

Figure 4.2 compares the accuracy of the current study with prior literature, as shown in the bar chart. The current study achieved a maximum accuracy of 99.61% (Ensemble, Technique 3), exceeding the Naïve Bayes (NB) accuracy of 96.55% (Technique 2). This surpasses Ahamed et al. [16] at 90.00% for TB outcomes in Pakistan, Yuan et al. [11] at 95.00% for antibiotic resistance, Kotei et al. [10] at 94.83% with ensemble methods, and Wang et al. [9] at 96.55% with deep learning. The results highlight the current study's competitive performance, enhanced by ADASYN and diverse models, despite a 400-sample dataset.

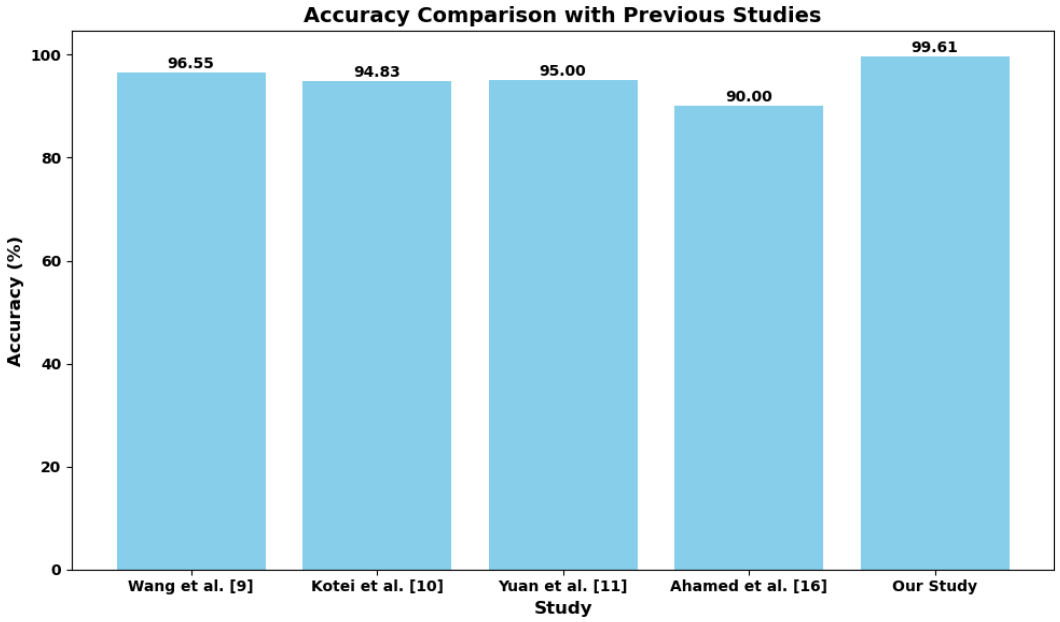


**Figure 4.2: Comparison of Accuracy Results with Existing Literature**

## 5. Discussion

The results demonstrate the effectiveness of machine learning in predicting TB drug resistance, NB achieving the highest accuracy of 96.55% under the Training Dataset Balanced approach (Technique 2), likely due to its feature independence assumption fitting the dataset's structure. RF and DT followed at 94.83%, leveraging their ability to handle mixed data types. In contrast, MLP underperformed at 89.66% in Technique 1, suggesting neural networks need larger datasets. Using ADASYN in Technique 2 improved performance over Technique 1 (Whole Dataset Imbalanced). In Technique 1, models like NB and MLP struggled with class imbalance, a common issue in medical datasets. ADASYN mitigated this in Technique 2, enhancing performance, but Technique 3's 99.61% accuracy risks data leakage, potentially overestimating generalizability.

Additional models like GB, XGB, and ensemble methods offered broader insights. The Ensemble model performed well under Technique 1 (Accuracy: 94.83%, F1-Score: 95.41%), showing ensembles suit complex datasets. Logistic Regression showed lower recall, likely due to its linear nature. The 12 key features, like drug responses to Moxifloxacin and Isoniazid, could reduce testing costs in resource-limited settings like Pakistan. However, the 400-sample dataset limits diversity and may miss regional resistance variations. Future research should use larger, diverse datasets, explore techniques like time series or clustering, and validate models clinically for practical utility.

## 6. Conclusion

This work successfully applied supervised machine learning to Pakistan TB patients to predict drug resistance patterns with maximum accuracy of 96.5-99.6%, resulting from the use of Naïve Bayes under the Training Dataset Balanced approach. Utilizing a dataset of 400 samples with 12 key features, the

presented approach ensures a low requirement for a large number of tests while ensuring high predictive accuracy. The application of ADASYN to handle the class imbalance proved effective, especially the challenge of improving the recall for minority classes, such as MDR-TB cases, that are very important for such clinical applications. The analysis of several models, such as ensemble models, demonstrated the strength of ensemble methods to deal with high dimensional medical data sets.

The findings reveal the latent ability of machine learning to enhance TB treatment outcomes by early determination of resistance, especially in resource deprived settings where diagnostic tools are few. Early detection of resistance patterns will allow clinicians to individualize the approach to patients, minimizing the risk of failure of therapy and spread of resistant strains. However, the use of a relatively small dataset by the study restricts the level of generalization of the study, and future studies should involve larger and heterogeneous datasets that have the capacity of approximating the regional variation on resistance patterns. In addition, the investigation of hybrid models that involve the integration of several algorithms' strengths might contribute to better prediction accuracy. This research lies at the foundation for a scalable, data-driven approach to containing TB and contributes to the worldwide prevention of the spread of this lethal disease.

**References**

[1] World Health Organization, Global tuberculosis report 2023, World Health Organization, Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports

[2] Maphalle, Lehlogonolo N F et al. “Pediatric Tuberculosis Management: A Global Challenge or Breakthrough?.” Children (Basel, Switzerland), vol. 9, no. 8, 2022, p. 1120, doi:10.3390/children9081120

[3] Dorman, S. E. et al., "Four-month rifapentine regimens with or without moxifloxacin for tuberculosis," N. Engl. J. Med., vol. 384, no. 18, pp. 1705–1718, May 2021, doi: 10.1056/NEJMoa2033400

[4] Daftary, A. et al., "Dynamic needs and challenges of people with drug-resistant tuberculosis and HIV in South Africa: a qualitative study," Lancet Glob. Health, vol. 9, no. 4, pp. e479–e488, Apr. 2021, doi: 10.1016/S2214-109X(20)30548-7

[5] Al Meslamani, Ahmad Z et al. “Machine learning in infectious diseases: potential applications and limitations.” Annals of Medicine, vol. 56, no. 1, 2024, p. 2362869, doi:10.1080/07853890.2024.2362869

[6] Shaikh, Juveriya et al. "Skin cancer detection: A review using AI techniques." International Journal of Health Sciences, vol. 6, suppl. 2, 2022, pp. 14339–14346, doi:10.53730/ijhs.v6nS2.8761

[7] Orjuela-Cañón, Alvaro D. et al. "Machine learning in the loop for tuberculosis diagnosis support." Frontiers in Public Health, vol. 10, 2022, p. 876949, doi: 10.3389/fpubh.2022.876949

[8] Wang, Yuhan et al. "Recent Progress in Tuberculosis Diagnosis: Insights into Blood-Based Biomarkers and Emerging Technologies." Frontiers in Cellular and Infection Microbiology, vol. 15, 2025, p. 1567592, doi: 10.3389/fcimb.2025.1567592

[9] Wang, Yaxi et al. “A deep learning model for predicting multidrug-resistant organism infection in critically ill patients.” Journal of Intensive Care, vol. 11, no. 1, 2023, p. 49, doi:10.1186/s40560-023-00695-y

[10] Kotei, Evans et al. "Ensemble technique coupled with deep transfer learning framework for automatic detection of tuberculosis from chest X-ray radiographs." Healthcare, vol. 10, no. 11, 2022, doi: 10.3390/healthcare10112335

[11] Yuan, Kevin et al. "Machine learning and clinician predictions of antibiotic resistance in Enterobacterales bloodstream infections." Journal of Infection, vol. 90, no. 2, 2025, p. 106388, doi: 10.1016/j.jinf.2024.106388

[12] He, H. et al. "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239

[13] Zakariah, Mohammed et al. "Machine learning-based adaptive synthetic sampling technique for intrusion detection." Applied Sciences, vol. 13, no. 11, 2023, p. 6504, doi: 10.3390/app13116504

[14] Chawla, N. V. et al., "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953

[15] Kumar, Rakesh et al. "Tuberculosis in South Asia." Biomedical Journal of Scientific & Technical Research, vol. 40, no. 4, 2021, pp. 32543–32545, doi: 10.26717/BJSTR.2021.40.006496

[16] Ahamed Fayaz, Shaik et al. "Machine learning algorithms to predict treatment success for patients with pulmonary tuberculosis." PLOS ONE, vol. 19, no. 10, 2024, p. e0309151, doi: 10.1371/journal.pone.0309151

[17] Nadkarni, Shantaram B. et al. "Comparative study of random forest and gradient boosting algorithms to predict airfoil self-noise." Engineering Proceedings, vol. 59, no. 1, 2023, p. 24, doi:10.3390/engproc2023059024

[18] Shi, Guang et al. "Efficient Support Vector Machine Toward Medical Data Processing." Proceedings of Seventh International Congress on Information and Communication Technology, 2022, doi: 10.1007/978-981-19-1610-6_66

[19] Miotto, Riccardo et al. "Deep learning for healthcare: review, opportunities and challenges." Briefings in Bioinformatics, vol. 19, no. 6, 2018, pp. 1236–1246, doi:10.1093/bib/bbx044

[20] Naderalvojoud, Behzad et al. "Improving machine learning with ensemble learning on observational healthcare data." AMIA Annual Symposium Proceedings, vol. 2023, 2024

[21] Kale, Arati K et al. "Data pre-processing technique for enhancing healthcare data quality using artificial intelligence." International Journal of Scientific Research in Science and Technology, 15 Jan. 2024, pp. 299–309, doi: 10.32628/ijsrst52411130

[22] Dhany, Hanna Willa et al. "Exploratory Data Analysis (EDA) methods for healthcare classification." Journal of Intelligent Decision Support System (IDSS), vol. 6, no. 4, 2023, pp. 209–215

[23] Bichri, Houda et al. "Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets." International Journal of Advanced Computer Science & Applications, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150235

[24] Abdulqader, Hozan Akram et al. "A Review on Decision Tree Algorithm in Healthcare Applications." The Indonesian Journal of Computer Science, vol. 13, no. 3, 2024, doi: 10.33022/ijcs.v13i3.4026

[25] Owusu-Adjei, Michael et al. "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems." PLOS Digital Health, vol. 2, no. 11, 2023, doi: 10.1371/journal.pdig.0000290

[26] Monica et al., "A Survey on Hyperparameter Optimization of Machine Learning Models," 2024 2nd Int. Conf. on Disruptive Technologies (ICDT), Greater Noida, India, 2024, pp. 11–15, doi: 10.1109/ICDT61202.2024.10489732

[27] Sharma, Vinayak et al. "Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images." Intelligent Medicine, vol. 4, no. 2, 2024, pp. 104–113, doi: 10.1016/j.imed.2023.06.001

[28] Keser, Sinem Bozkurt et al. "A gradient boosting-based mortality prediction model for COVID-19 patients." Neural Computing and Applications, vol. 35, no. 33, 2023, pp. 23997–24013, doi: 10.1007/s00521-023-08997

[29] Mahajan, Palak et al. "Ensemble Learning for Disease Prediction: A Review." Healthcare (Basel, Switzerland), vol. 11, no. 12, 2023, p. 1808, doi:10.3390/healthcare11121808

[30] Hwang, Eui Jin et al. "AI for detection of tuberculosis: Implications for global health." Radiology: Artificial Intelligence, vol. 6, no. 2, 2024, p. e230327, doi: 10.1148/ryai.230327