IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 June 2023, accepted 12 July 2023, date of publication 25 July 2023, date of current version 14 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298750

APPLIED RESEARCH

Combining Radiological and Genomic TB Portals Data for Drug Resistance Analysis

VY C. B. BUI[1], ZIV YANIV[2], MICHAEL HARRIS[2], FENG YANG[1], KARTHIK KANTIPUDI[2], DARRELL HURT[2], ALEX ROSENTHAL[2], AND STEFAN JAEGER[1]

[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[2]Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Corresponding authors: Stefan Jaeger (stefan.jaeger@nih.gov) and Vy C. B. Bui (vy.bui@nih.gov)

This work was supported in part by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under Interagency Agreement 750119PE080057; in part by the Lister Hill National Center for Biomedical Communications of the National Library of Medicine (NLM), National Institutes of Health; and in part by the Federal Funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, under Bioinformatics and Computational Biosciences Branch (BCBB) Support Services to Medical Science and Computing under Contract HHSN316201300006W/75N93022F00001.

ABSTRACT Tuberculosis (TB) drug resistance is a worldwide public health problem. It decreases the likelihood of a positive outcome for the individual patient and increases the likelihood of disease spread. Therefore, early detection of TB drug resistance is crucial for improving outcomes and controlling disease transmission. While drug-sensitive tuberculosis cases are declining worldwide because of effective treatment, the threat of drug-resistant tuberculosis is growing, and the success rate of drug-resistant tuberculosis treatment is only around 60%. The TB Portals program provides a publicly accessible repository of TB case data with an emphasis on collecting drug-resistant cases. The dataset includes multi-modal information such as socioeconomic/geographic data, clinical characteristics, pathogen genomics, and radiological features. The program is an international collaboration whose participants are typically under a substantial burden of drug-resistant tuberculosis, with data collected from standard clinical care provided to the patients. Consequentially, the TB Portals dataset is heterogenous in nature, with data representing multiple treatment centers in different countries and containing cross-domain information. This study presents the challenges and methods used to address them when working with this real-world dataset. Our goal was to evaluate whether combining radiological features derived from a chest X-ray of the host and genomic features from the pathogen can potentially improve the identification of the drug susceptibility type, drug-sensitive (DS-TB) or drug-resistant (DR-TB), and the length of the first successful drug regimen. To perform these studies, significantly imbalanced data needed to be processed, which included a much larger number of DR-TB cases than DS-TB, many more cases with radiological findings than genomic ones, and the sparse high dimensional nature of the genomic information. Three evaluation studies were carried out. First, the DR-TB/DS-TB classification model achieved an average accuracy of 92.4% when using genomic features alone or when combining radiological and genomic features. Second, the regression model for the length of the first successful treatment had a relative error of 53.5% using radiological features, 25.6% using genomic features, and 22.0% using both radiological and genomic features. Finally, the relative error of the third regression model predicting the length of the first treatment using the most common drug combination varied depending on the feature type used. When using radiological features alone, the relative error was 17.8%. For genomic features alone, the relative error increased to 19.9%. The model had a relative error of 19.0% when both

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

radiological and genomic features were combined. Although combining radiological and genomic features did not improve upon the use of genomic features when classifying DR-TB/DS-TB, the combination of the two feature types improved the relative error of the predictive model for the length of the first successful treatment. Furthermore, the regression model trained on radiological features achieved the best performance when predicting the treatment length of the most common drug combination.

INDEX TERMS Tuberculosis, radiomics, genomics, TB Portals, machine learning, drug resistance.

I. INTRODUCTION

Despite recent reductions in tuberculosis incidence (TB) and mortality, the emergence of drug-resistant Mycobacterium tuberculosis is a critical global health issue. Failure to identify and appropriately treat patients with drug-resistant TB can lead to increased mortality, nosocomial outbreaks, and the spread of drug resistance. The TB Portals program is an international collaboration that currently includes 16 countries under a heavy burden of drug-resistant tuberculosis [1].[1] Though drug-sensitive tuberculosis numbers are diminishing worldwide, and the effectiveness of treatment is above 90%, drug-resistant tuberculosis cases are increasing, and the successful treatment percentage of DR-TB is about 60% [2]. To address this issue, the TB Portals program was established to serve as a research resource with an emphasis on the collection of DR-TB case data. The TB Portals database is an anonymized patient-centric resource containing multi-domain data, including patient images (chest X-rays and computed tomography scans), derived radiological features, socio-economical information, and pathogen single-nucleotide polymorphisms associated with drug resistance. Data is collected during standard patient care, curated, and made available to the research community. The program has collected and processed genome sequences for more than 2200 Mycobacterium tuberculosis (M. tuberculosis) samples as part of this international collaboration, information that is usually not collected as part of standard patient care.

Previous analyses of TB Portals data include country-specific studies of DR-TB molecular evolution [3], genomic evaluation of relapse/reinfection status [4], polyclonal infection among lung resection samples [5], clinical metadata correlation with treatment outcomes [6], and prediction of drug susceptibility from patient imaging data via machine learning [7], [8], [9], [10], [11].

Analyzing relevant multi-modal information extracted from large datasets can be challenging. Many studies have investigated different statistical and computational methods to effectively address the challenges in big, high dimensional datasets [12], [13], [14], [15], [16]. The present study describes the challenges when integrating radiological and genomic information using not only big but also multi-source, cross-domain, and real-world data. In particular, reducing the genomic data to a tractable set of variables to manage its sparse and high dimensional nature is a point of discussion. In addition, a dimensionality reduction of radiological findings is performed by grouping features that share lower-level information into broader classes to make the statistical analysis and machine prediction processes more efficient. Later, an illustration is provided on how machine learning methods can be applied to unbalanced data, considering that the TB Portals data is explicitly biased toward DR-TB cases.

Using TB Portals data, this study applied machine learning to understand the relationship between the host's radiological findings and the pathogen's genomic information. The underlying hypothesis was that the two information sources are complementary and that the radiological information may improve the performance of predictive models that use the pathogen's genetic information. Radiological information consisted of clinical findings identified by a radiologist in frontal chest X-ray imaging. The coarse location of each finding within the lung was also noted (lungs are divided into six regions). Findings included lung cavity (of various sizes), nodules (of various sizes), infiltrate density, presence of mediastinal lymph nodes, presence of calcified nodules, and others. In addition, the annotations at the overall lung level, such as the overall percentage of abnormal volume and percentage of pleural effusion of the involved hemithorax, were also provided. Genomic information captured a detailed breakdown of the pathogen genomic variance, such as gene mutation variants.

Three evaluation studies were conducted. The first investigated the usefulness of radiological, host, and genomic pathogen features for the classification of drug susceptibility, DR-TB or DS-TB. The second investigated the usefulness of these features for regression, estimating the length of the first successful treatment period. The third investigated the treatment period of the most common drug combination. In three studies, not all patient cases included all data elements, which is a common characteristic of real-world data. Therefore, three subsets were defined, consisting of 5935 patients with radiological data, 2161 patients with pathogen genomic data, and 1272 patients with both radiological and genomic data, respectively.

The rest of the paper is organized as follows: First, the radiological and genomic information, feature encoding, and the methodology used to address data imbalance are described. Then, the statistical analysis and machine learning models developed for the three tasks drug susceptibility classification and treatment period prediction are presented. The following

[1] https://tbportals.niaid.nih.gov/

sections report the experimental results and discuss the limitations of working with a real-world heterogeneous dataset. The final section provides a summary of the challenges encountered and the results of the predictive models.

II. METHODS

A. RADIOLOGICAL AND GENOMIC INFORMATION

A dataset of 5935 patients with radiological findings based on chest X-rays provided by the TB Portals program was used. Each patient record included radiological findings from a frontal chest X-ray, as determined by a single radiologist. Radiological findings included chest radiography patterns such as the overall percentage of abnormal volume, the pleural effusion percentage of the involved hemithorax and whether the pleural effusion is bilateral, the presence of mediastinal lymphadenopathy, cavities, infiltrates, calcified and partially calcified or non-calcified nodules, and the presence of other non-TB abnormalities, including affected sextants. The 5935 patients were comprised of 4028 DR-TB cases and 1907 DS-TB cases.

The TB Portals Steering Committee directed the sequencing and genomic study of M. tuberculosis found in patients to enhance the understanding of the molecular basis of the disease, as genomic variations are known to be related to the resistance of M. tuberculosis to common antibiotics [17], [18], [19], [20], [21], [22]. In addition, genomic information captures a detailed breakdown of the pathogen genomic variance, such as gene mutation variants. Therefore, a dataset of 2161 patients with pathogen genomic information, comprised of 1614 DR-TB cases and 547 DS-TB cases, was used in this study. In addition, genomic analysis results by TB Profiler [23] for patients that have both imaging and genomic data were also included.

Data usage was exempt from local institutional review board review because the data is publicly available from the TB Portals program. The TB Portals program participants are responsible for ensuring compliance with their countries' laws, regulations, and ethical considerations.

B. FEATURE ENCODING

The feature set included 18 radiological features and 43 genomic variants, including nine from the TB Portals genomic pipeline and 34 from TB Profiler. Because the radiological and genomic features were categorical data, they were converted to a numeric format before being used as input to the machine learning models. The categorical features used in this study were all nominal, so one-hot encoding was used to encode these features.

As some of the radiological features referred to the size of clinical findings, e.g., large cavity, large nodule, or high-density infiltrate, all of them were combined into a single class based on the type of finding. This enabled the use of specific domain knowledge to perform dimensionality reduction, reducing the original 25-dimensional feature vector to 18.

The genomic variables in the dataset included the single nucleotide polymorphism (SNP) detected at one nucleotide due to a mutation at that location. As a result, the genomic variable, i.e., gene_snp_mutations, was high dimensional. Each patient could have a combination of SNPs, among 99 possible combinations. To reduce the dimensionality of the sparse one-hot encoded matrix, the SNPs were combined into their unique individual genes. Thus, a total of nine unique genes were obtained after encoding. Next, to encode the genomic features in the TB Profiler data, there were 2218 variables in the TB Profiler data. Each variable had two parts, representing the gene and variant names. The high dimensional variables were reduced by aggregating them at the gene level. To encode an aggregated gene, if there was any variant in the gene, the gene was encoded as one; otherwise, as zero. So, 34 gene names were present after encoding. Therefore, 43 genomic features (nine from TB Portals and 34 from TB Profiler) were used for analysis and prediction.

C. CORRELATION AND STATISTICAL ANALYSIS

The relationship between radiological and genomic features was investigated by computing the Pearson correlation (R). A correlation value $|R| \leq 0.25$ was considered a weak correlation, $0.25 \leq |R| \leq 0.50$ was a mild correlation, and $|R| \geq 0.50$ was a strong correlation. Pearson's chi-squared test measured the statistical significance of radiological and genomic features and the type of resistance. The null hypothesis stated that there is no relationship between the radiological/genomic features and the type of resistance. In addition, a Mann-Whitney U Test was used to measure the statistical significance of differences in the treatment period between radiological and genomic features. The null hypothesis stated that no significant difference in the means of the treatment periods exists for radiological and genomic features. A p-value < 0.05 was considered statistically significant.

D. HANDLING IMBALANCED DATA

The dataset included 4028 DR-TB and 1907 DS-TB cases and was thus biased toward DR-TB. It is known that machine learning is sensitive to the ratio of different classes' sample sizes. The problem is that a model can achieve high accuracy by consistently predicting the majority class. Imbalanced datasets create challenges for predictive modeling, but they are a common and anticipated problem because they are typical of real-world applications. Several approaches have been followed to balance a dataset, such as the down-sampling of the over-represented class and the over-sampling of the under-represented class. Down-sampling removes samples from the over-represented class until the classes have an equal distribution. However, removing samples from the original dataset could result in the loss of useful information. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) [24] is a popular method for synthesizing new samples for the minority class. In this study, a variant of SMOTE referred

to as SVM-SMOTE [25] was used to balance the class distribution.

E. MACHINE CLASSIFICATION OF DRUG-SENSITIVE AND DRUG-RESISTANT TB

In this evaluation study, the radiological findings for chest X-ray images and the gene variants in TB Portals were used to assess their role in discriminating between drug-resistant and drug-sensitive TB. A cohort was formed using the following inclusion and exclusion criteria: cases with available chest X-ray annotations and genomic information, Poly DR and Pre-XDR cases were excluded as there were very few samples, all other drug-resistant classes (MDR non-XDR, Mono DR, XDR) were combined into a single DR-TB class, and only new, no follow-up or relapse patients were used for analysis. Under these conditions, from over 13000 patient cases, cohorts consisting of 5935 patients with radiological findings from chest X-rays, 2161 patients with pathogen genomic information, and only 1272 patients with both radiological and genomic data were identified.

Next, a machine learning classifier, Random Forest [26], was trained to discriminate between DR-TB and DS-TB. Different feature combinations were used to compare the contributions of various features for classifying DR-TB and DS-TB. The dataset included 4028 DR-TB and 1907 DS-TB patients. It was thus biased toward DR-TB. Therefore, the classes were balanced by synthesizing new samples for the minority class. In addition, the feature selection was performed via the Random Forest algorithm's built-in feature importance, which computed the Gini importance or information gain. Random Forest is a set of decision trees in which each tree consists of nodes and leaves. In each node, a selected feature is used to decide how to divide the dataset into two separate groups. The features for internal nodes are selected based on a criterion which is typically Gini impurity or information gain for classification tasks. One can measure how each feature decreases the impurity of the split; the feature with the highest decrease is selected for the internal node. The average decrease over all trees in the forest is then used as the measure of the feature importance. Finally, the classification performance was evaluated based on a five-fold cross-validation.

F. MACHINE PREDICTION OF THE PERIOD OF THE FIRST SUCCESSFUL TREATMENT

Machine learning with radiological and genomic information was used to predict the period of successful treatment. Similar to the first study, the relationship between radiological/genomic features and the treatment period (TP) was investigated. A cohort was formed using the following criteria: cases with available chest X-ray annotations and genomic information, Poly DR and Pre-XDR cases were excluded as there were very few samples, treatment outcome of ''Cured'' was included, and records of treatment periods greater than 30 days were included. As a result, cohorts consisting of 2151 patients with radiological findings from chest X-rays, 1023 patients with pathogen genomic information, and only 818 patients with both radiological and genomic data were identified. Thus, from a big dataset with several thousand patient cases, after applying the selection criteria and due to the fact that the set of patients with genomic data only partially overlaps with the set of patients with radiological data, the usable dataset was reduced to only 818 cases.

Next, a Gradient Boosting regression model was trained to predict the first successful drug regimen length using radiological and genomic features. Gradient Boosting calculates the mean value of the reference values and makes initial predictions [27]. Then, using the predictions, it calculates the gradients, which are the differences between the predicted and actual values. Instead of training a new estimator on the data to predict the target, it trains an estimator to predict the gradients of the initial predictor. This predictor is usually a decision tree. To make predictions, it adds the base estimator's value onto the decision tree's predicted gradient value of the instance. It then calculates the gradients again between the predicted and actual values. This process is repeated until a certain threshold is reached or the gradient difference is minimal.

G. MACHINE PREDICTION OF THE TREATMENT PERIOD OF THE MOST COMMON DRUG COMBINATION

A cohort was formed using the following criteria: cases with available chest X-ray annotations and genomic information, Poly DR and Pre-XDR cases were excluded as there were very few samples, and records of treatment periods greater than 30 days were included. These selection criteria resulted in cohorts comprising 3171 patients with radiological findings, 1426 patients with pathogen genomic information, and 1120 patients with both radiological and genomic data. A long list of drug combinations was found in the TB Portals data: 29 drug combinations to treat DS-TB and 728 drug combinations to treat DR-TB. However, only the most common drug combination (Ethambutol, Isoniazid, Pyrazinamide, Rifampicin) had over a thousand samples. The second (Ethambutol, Isoniazid, and Rifampicin) and third (Bedaquiline, Clofazimine, Cycloserine, Levofloxacin, and Linezolid) most common drug combinations only had 278 and 156 samples, respectively. For other drug combinations, the sample size was less than 50 cases, with some drugs having only one case. Because of this limited sample size, experiments to predict the treatment period were performed only for the most common drug regimen combination. As a result, cohorts consisting of 1296 patients with radiological findings from chest X-rays, 480 patients with pathogen genomic information, and 411 patients with both radiological and genomic data were used.

Next, a Gradient Boosting regression model was trained to predict the treatment period of the most common drug combination using radiological and genomic features. Different feature combinations were used to compare the contributions of different features in predicting the span of the treatment

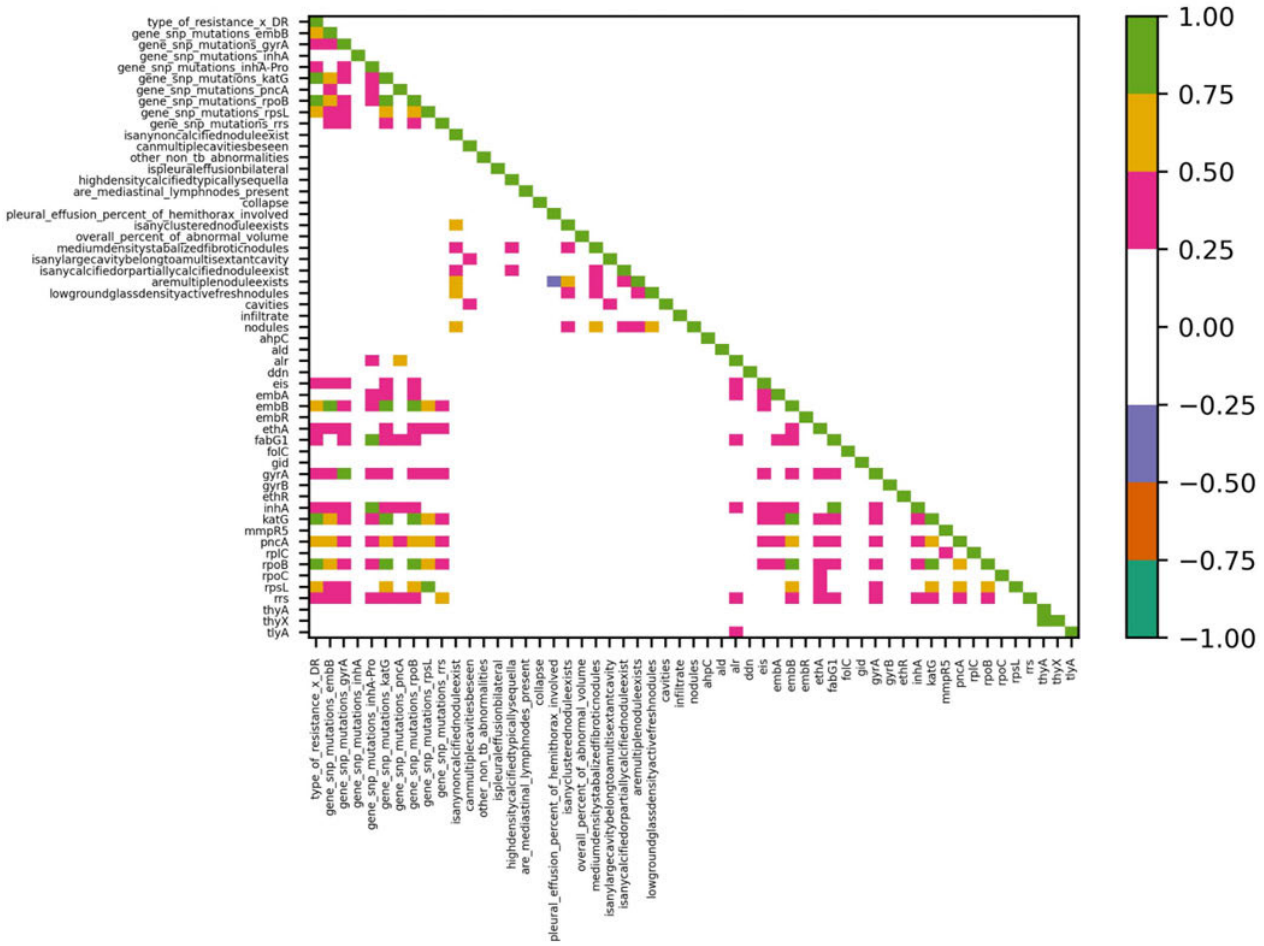

FIGURE 1. The correlation map for radiological and genomic features and drug resistance status (N=1272). Correlation value $|R| \leq 0.25$ is considered a weak correlation, $0.25 \leq |R| \leq 0.50$ is a mild correlation, and $|R| \geq 0.50$ is a strong correlation. Note that genomic features such as gene_snp_mutations_embB and embB are not necessarily distinct, gene_snp_mutations_embB are extracted from the internal TB Portal pipeline and embB came from TB Profiler.

period. The mean, median, and relative percentage errors were used to assess the model performance.

III. RESULTS

In the first study, the relationship between radiological and genomic features and their effect on DR-TB/DS-TB classification was investigated. The correlation was computed between them, and it was observed that all the radiological and genomic features are weekly correlated ($|R| \leq 0.25$), as shown in Figure 1. In Figure 2, all patients with available radiological records (N=5935) were used to compute the correlation between radiological features and DR-TB. Here, all radiological features were weakly correlated with DR-TB ($|R| \leq 0.25$). Similarly, all patients with available genomic records (N=2161) were used to compute the correlation between genomic features and DR-TB. As shown in Figure 3, eight genomic features were mildly correlated with DR-TB ($0.25 \leq |R| \leq 0.50$), and nine genomic features were strongly correlated with DR-TB ($|R| \geq 0.50$). In addition, some genomic features were highly correlated with other genomic features; for example, embB and katG, embB and rpoB, fabG1 and inhA-Pro, fabG1 and inhA, katG and rpoB, rpoB and embB. Next, Pearson's chi-squared test showed that 16 out of 18 radiological features and 29 out of 43 genomic features were statistically significant regarding DR-TB ($p < 0.05$). So, while genomic features showed a strong relationship with DR-TB, as expected, radiological features were statistically significant but weakly correlated with DR-TB, which meant that the association was small but did exist.

Next, a Random Forest algorithm was empirically selected to classify DR-TB vs. DS-TB and evaluated using five-fold cross-validation. This approach was selected because it obtained the best performance among the six machine learning classification approaches that were evaluated. Table 1 summarizes the performances of the evaluated classifiers. The AdaBoost algorithm was configured with a maximum of 50 estimators and a learning rate of 1.0. The Decision Tree classifier employed Gini impurity as the criterion; the minimum number of samples required to split an internal node was 2, and the minimum number of samples at a leaf node was 1. Gaussian Naive Bayes utilized variance

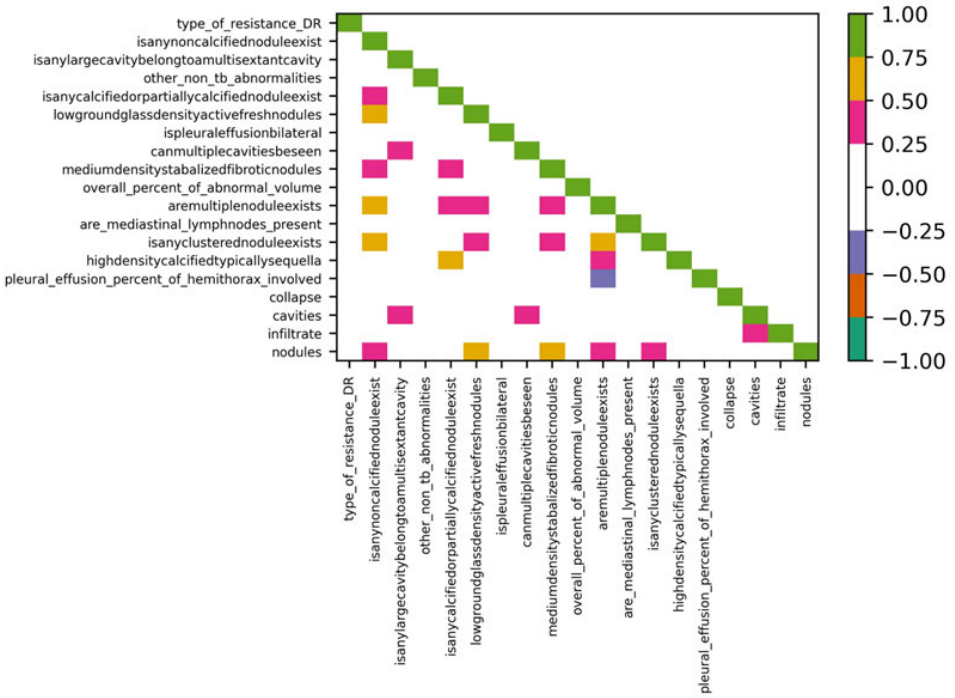

FIGURE 2. The correlation map for radiological features and drug resistance status (N=5935). Correlation value $|R| \leq 0.25$ is considered a weak correlation, $0.25 \leq |R| \leq 0.50$ is a mild correlation, and $|R| \geq 0.50$ is a strong correlation.

TABLE 1. Comparing the performance of various classifiers in distinguishing drug-resistant TB and drug-sensitive TB using radiological and genomic records (N=1272).

Classifiers	Accuracy (%)
AdaBoost	92.0 ± 1.6
Decision Tree	89.4 ± 1.3
Naive Bayes	92.3 ± 1.1
Multi-layers Perceptron	92.1 ± 1.3
Random Forest	92.4 ± 0.8
SVM	91.5 ± 1.2

smoothing set at 1e-9. The Multilayer Perceptron consisted of 100 hidden layer nodes, a ReLU activation function, and employed the Adam optimizer with a learning rate of 0.001. Random Forest incorporated 1000 trees with a maximum tree depth of 10, using Gini impurity as the criterion and requiring a minimum of 2 samples for internal node splitting and 1 sample for a leaf node. SVM employed a regularization parameter of 1.0 and used the radial basis function as the kernel.

As shown in Table 2, the Random Forest classification model achieved an average accuracy of 93.9% using only genomic features (N=2161), whereas the model's average accuracy using only radiological features was 66.7% (N=5935). To compare different features, the model was tested on the same cohort, on all patients with radiological and genomic records (N=1272). The classification model achieved an average accuracy of 92.4% when combining radiological and genomic features, and also 92.4% when using genomic features alone. The model's average accuracy using radiological features alone was 65.0%, as shown in Table 2. This study suggested that the radiological and genomic findings could predict DR-TB, with the genomics features alone providing the best performance, and that adding the radiological features did not diminish the performance of the genomic features but also did not improve it.

In the second study, the relationship between radiological and genomic features was investigated for this cohort, including 2246 patients with radiological features (1130 DR-TB, 1116 DS-TB), 1040 patients with genomic features (672 DR-TB, 368 DS-TB), and only 829 patients with both radiological and genomic features (483 DR-TB, 346 DS-TB). All the radiological and genomic features were weekly correlated, as shown in Figure 4. Similar to the first study, the relationship between radiological and genomic features and the length of the first successful TP were investigated. As shown in Figure 5, three radiological features were mildly correlated with TP ($0.25 \leq |R| \leq 0.50$), including the pleural effusion percentage of the involved hemithorax (R=-0.43), the presence of clustered nodules (R=0.29), and the presence of multiple nodules (R=0.32). In Figure 6, 12 genomic features were mildly correlated with TP ($0.25 \leq |R| \leq 0.50$) and seven genomic features were strongly correlated with

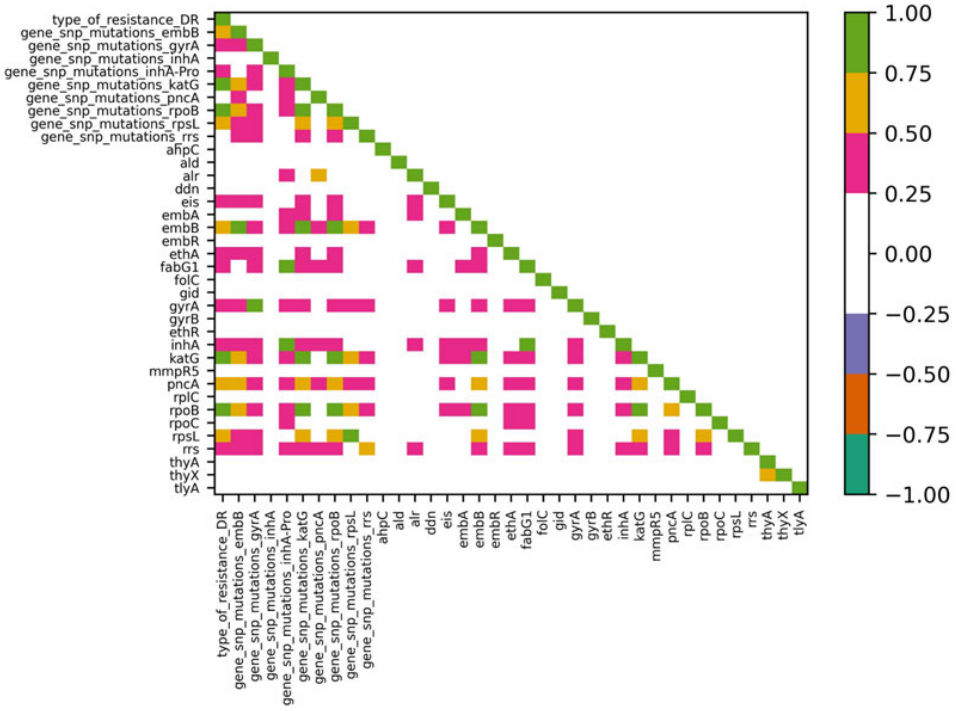

FIGURE 3. The correlation map for genomic features and drug resistance status (N=2161). Correlation value $|R| \leq 0.25$ is considered a weak correlation, $0.25 \leq |R| \leq 0.50$ is a mild correlation, and $|R| \geq 0.50$ is a strong correlation.

TABLE 2. Evaluating the performance of the random forest classifier in distinguishing drug-resistant TB and drug-sensitive TB with different features: using radiological records for 5935 patients, genomic records for 2161 patients, and records with both radiological and genomic information for 1272 patients.

Training features	N	Accuracy (%)	AUC (%)
Radiological features	5935	66.7 ± 1.2	69.5 ± 0.8
Genomic features	2161	93.9 ± 0.7	95.3 ± 0.6
Radiological features	1272	65.0 ± 0.7	70.4 ± 2.4
Genomic features	1272	92.4 ± 0.3	95.5 ± 1.2
Radiological and Genomic features	1272	92.4 ± 0.8	95.7 ± 0.5

TP ($|R| \geq 0.50$). As a result, TP correlated well with most of the genomic features and with only three radiological features. A Mann-Whitney U Test showed that all 18 radiological features and all 42 genomic features were not independent of TP.

Six regression approaches based on machine learning were evaluated for predicting the first successful treatment length (Table 3). Gradient Boosting provided the best performance. Gradient Boosting was configured with 1000 boosting stages. The individual regression estimators had a maximum depth of 4, and a minimum of 5 samples was required to split an internal node. The loss function employed mean squared error, and the learning rate was set to 0.01. The parameters of the other regression models used in the study remained the same as in the first study.

The performance of Gradient Boosting regression was then evaluated using various feature combinations (Table 4). The Gradient Boosting regression model to predict the first successful treatment length had an average relative error of 48.6% using radiological features alone (N=2246), and the model's average relative error using genomic features alone was 25.1% (N=1040).

To compare different features, the regression model was tested on the same cohort, including all patients with available radiological and genomic records (N=829). The resulting regression model had a mean absolute error of 162.9 days

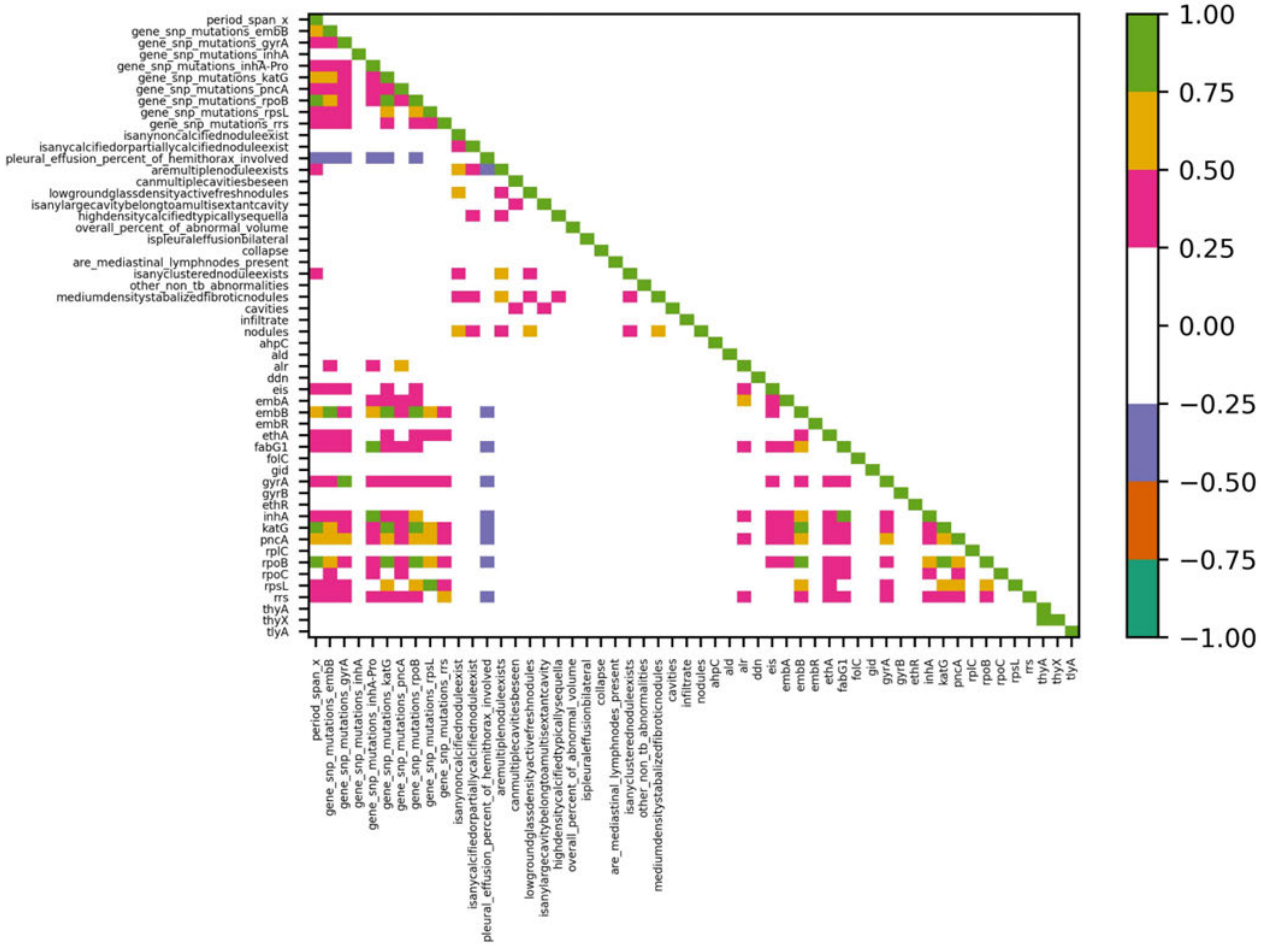

FIGURE 4. The correlation map for radiological and genomic features and treatment period (N=829). Correlation value $|R| \leq 0.25$ is considered a weak correlation, $0.25 \leq |R| \leq 0.50$ is a mild correlation, and $|R| \geq 0.50$ is a strong correlation.

using radiological features, 91.4 days using genomic features, and 76.8 days using both feature types. The model's relative errors were reduced to 22.0% using both radiomics and genomics compared to 53.5% when using only radiological features or 25.6% when using only genomic features. Note that the treatment length could span up to 3 years. Thus, a prediction error of less than three months was encouraging.

Table 5 shows the performance evaluation of different training feature settings when predicting the treatment period of the most common drug combination. In the TB Portal dataset, this was Ethambutol, Isoniazid, Pyrazinamide, and Rifampicin. The regression model had an average relative error of 27.4% using only radiological features (N=1296), and the model's average relative error using only genomic features was 21.2% (N=480).

To compare different features, the regression model was tested on the same cohort, including all patients with available radiological and genomic records (N=411). The regression model achieved a mean absolute error of 34.1 days using radiological features, 37.7 days using genomic features, and 36.3 days using both features. The relative error was 17.8% using the radiological features, 19.9% using the genomic features, and 19.0% using both feature types.

TABLE 3. Comparing the performance of different regression models in predicting the period of the first successful treatment using radiological and genomic records for 829 patients.

Regression Models	Relative Error (%)
AdaBoost	56.3 ± 2.9
Decision Tree	27.4 ± 4.0
Gradient Boosting	22.0 ± 1.6
Multi-layers Perceptron	26.4 ± 3.8
Random Forest	23.4 ± 2.5
SVM	50.0 ± 2.2

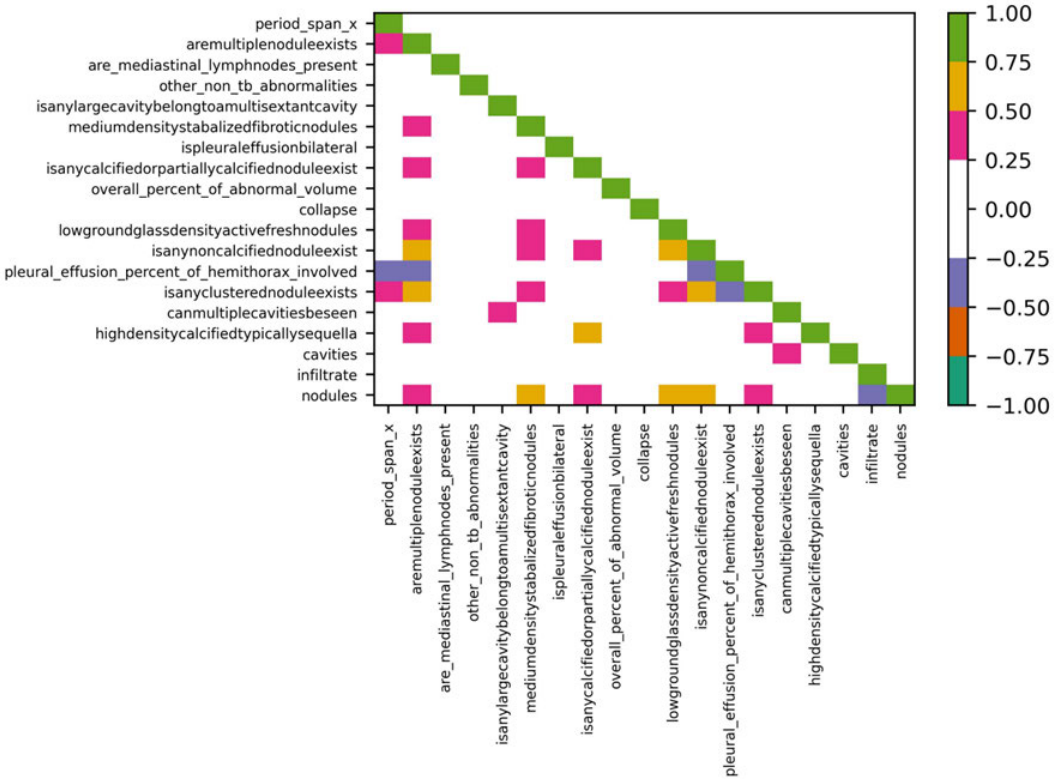

FIGURE 5. The correlation map for radiological features and the first successful treatment period (N=2246). Correlation value $|R| \leq 0.25$ is considered a weak correlation, $0.25 \leq |R| \leq 0.50$ is a mild correlation, and $|R| \geq 0.50$ is a strong correlation.

IV. DISCUSSION

The TB Portals program provided a heterogenous, multimodal dataset, with data sources from different countries, containing cross-domain information. Prior studies utilizing TB Portals data had examined drug-resistant molecular evolution on a country-specific level [3], evaluated the genomics of relapse and reinfection status [4], investigated polyclonal infection among lung resection samples [5], explored the correlation between clinical metadata and treatment outcomes [6], and predicted drug susceptibility using machine learning on patient imaging data [7], [8], [9], [10], [11]. Karki et al. presented a thorough study of computational methods for detecting and predicting TB drug resistance [28]. Several of their experiments have either established or pushed the current state-of-the-art in predicting drug resistance by computational means. Their results suggested that detecting DR-TB and predicting treatment outcomes in radiographs and clinical data could be possible to some extent. However, many questions still need to be answered, and this topic remains a subject of research. Predicting drug resistance in a patient early on and administering the appropriate patient-specific drugs would allow more efficient treatment that could save many lives and would thus be a significant breakthrough in the fight against drug-resistant TB [28].

This paper presented the challenges encountered and approaches used to address them while working with a real-world dataset. The goal was to evaluate whether combining radiological features derived from a chest X-ray of the host and genomic features from the pathogen can improve several prediction tasks. To address the issue of significantly imbalanced data, the approach involved synthesizing samples of the minority class. Furthermore, the TB Portals data contained more cases with radiological findings than genomic cases, and the sparse high dimensional nature of the genomic information was also a complication. These challenges were addressed using domain-specific knowledge, combining related features into super-classes, combining SNP information at the gene level, and combining radiological findings based on their type while ignoring their size. In the current work, coarser data classes were used than those available as a means of addressing data sparsity due to the high dimensional feature set. As the TB Portals program continues to collect data, the data is expected to become denser with the acquisition of additional cases.

The current study has some limitations: First, the prediction of the first treatment length is more of a theoretical exercise, evaluating the combination of host and pathogen features, than a practical one. The World Health Organization provides specific guidelines for treatment lengths based on the resistance status of the TB case. For example, sensitive TB is generally treated with a shorter timeline than drug-resistant TB, so knowing the predicted treatment length will not affect decisions with respect to treatment practices. However, this study considered a scenario with unknown

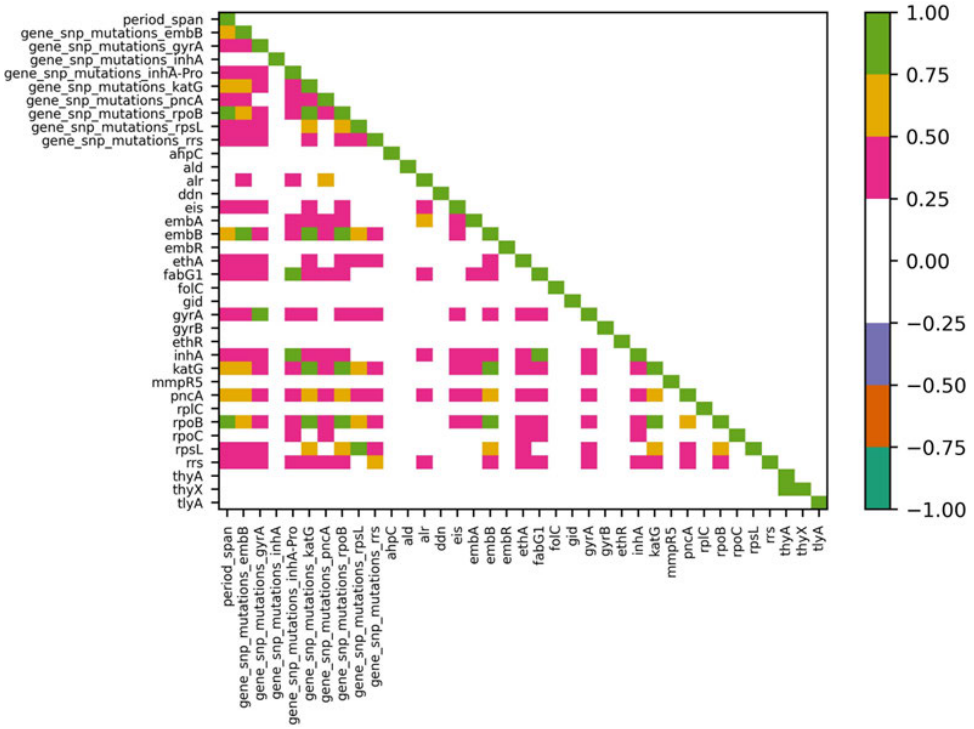

FIGURE 6. The correlation map for genomic features and the first successful treatment period (N=1040). Correlation value $|R| \leq 0.25$ is considered a weak correlation, $0.25 \leq |R| \leq 0.50$ is a mild correlation, and $|R| \geq 0.50$ is a strong correlation.

TABLE 4. Evaluating the performance of the gradient boosting regressor in predicting the period of the first successful treatment with different features: Using radiological records for 2246 patients, genomic records for 1040 patients, and records with both radiological and genomic information for 829 patients.

Training features	N	Mean Absolute Error (days)	Median Absolute Error (days)	Relative Error (%)
Radiological features	2246	144.3 ± 3.2	115.5 ± 3.5	48.6 ± 1.3
Genomic features	1040	93.6 ± 3.9	46.1 ± 2.1	25.1 ± 1.8
Radiological features	829	162.9 ± 6.1	140.0 ± 7.5	53.5 ± 3.2
Genomic features	829	91.4 ± 6.7	43.7 ± 3.8	25.6 ± 2.1
Radiological and Genomic features	829	76.8 ± 3.9	43.0 ± 5.6	22.0 ± 1.6

resistance status. One can make a point that once the genetic features are known, there is a strong indication of the patient's status. Nevertheless, the experiments also included radiological features, which could potentially improve prediction accuracy.

Second, for the genomic features that were used, correlations with radiological features were investigated. A limited genomic feature set was used that correlates well with resistance status. To truly evaluate pathogen genomic association with possible radiology findings, a genome-wide association study with the whole genome is needed.

Regarding the successful treatment length, inconsistencies were observed in the reported treatment lengths, which were a significant source of uncertainty. However, it is important to acknowledge the nuances of the data source. It is real-world data reflecting current clinical care, including the reality of many countries with differing levels of healthcare infrastructure. According to the TB Portals definitions, even for "new"

TABLE 5. Gradient Boosting performance for predicting the length of the first treatment using a single drug combination (Ethambutol, Isoniazid, Pyrazinamide, Rifampicin) with different features: Using radiological records for 1296 patients, genomic records for 480 patients, and records with both radiological and genomic information for 411 patients.

Training features	N	Mean Absolute Error (days)	Median Absolute Error (days)	Relative Error (%)
Radiological features	1296	48.7 ± 2.1	26.9 ± 2.1	27.4 ± 2.4
Genomic features	480	48.2 ± 10.0	17.0 ± 3.6	21.2 ± 6.2
Radiological features	411	34.1 ± 2.5	14.2 ± 1.9	17.8 ± 3.6
Genomic features	411	37.7 ± 5.7	14.1 ± 1.4	19.9 ± 5.9
Radiological and Genomic features	411	36.3 ± 5.0	16.1 ± 2.4	19.0 ± 5.4

cases, it is possible that treatment had already begun one month before the reported dates in the database.

Additionally, the patient cases are not uniformly distributed across data providers. The unequal distribution across providers is present in the first evaluation study as 57.6% of the cases of the selected cohort are from Georgia, 11.5% are from Belarus, 8.8% are from Ukraine, 7.8% are from Kazakhstan, 7.4% and 6.6% are from Romania and Moldova, respectively, and 0.3% are from Azerbaijan. That means the machine classifier learns to discriminate drug-sensitive and drug-resistant TB primarily from these seven countries. Thus, the classification performance is likely to decrease when the trained model is used to identify DR-TB from other countries or when performing classification on a country level. Similarly, when predicting the period of the first successful treatment, the regression model is also not trained based on a uniform distribution across data suppliers. That is, 60% of the patient cases of the selected cohort are from Georgia, 16% are from Belarus, 12% are from Romania, 5% are from Ukraine, 4% are from Moldova, 2% are from Kazakhstan, and 1% are from Azerbaijan. Finally, the machine learning model relies on radiological findings reported by a radiologist, which limits the full automation of the machine learning system.

According to Table 1, the performance of the Naïve Bayes classifier is comparable to the best result obtained by the Random Forest. A Naïve Bayes classifier can perform well if the dataset used for training and testing aligns well with the assumptions of Naïve Bayes in terms of class distributions and feature relationships, in particular statistical independence. Specifically, assuming a normal distribution of features can prove advantageous when dealing with limited or missing data. Therefore, it is plausible for a Naïve Bayes classifier to achieve similar accuracy to Random Forest. Secondly, Naïve Bayes classifiers can be more effective with smaller datasets, while the performance of random forests tends to improve as the dataset size increases. Due to the relatively small number of training samples, deep learning algorithms were not used. Instead, a multilayer perceptron was employed for the purpose of comparing its performance with other classifiers.

The studies confirmed correlations between radiological findings and DR-TB, between gene mutation variants and DR-TB (Figures 2, 3), and between radiological/genomic information and the first successful treatment length (Figures 5, 6). On the one hand, no correlations exist between radiological and genomic information for the cohort extracted to classify DR-TB and DS-TB, as shown in Figure 1. Furthermore, the combination of radiological and genomic information did not improve the accuracy of the classifier compared to the genomic information alone, as shown in Table 2. On the other hand, the combination of these two feature types improves the relative error of the regression model by 3.6%, as shown in Table 4. For predicting the treatment length of the most common drug combination, the model trained on radiological features performed better than the one trained on genomic features, as shown in Table 5. Most of the cases of this cohort are DS-TB, which explains why the genomic information does not contribute to the prediction because the gene mutation variants are more likely to indicate drug-resistant status in TB Portals data.

Therefore, there is value in collecting data from multiple domains to form a complete view of a TB case. Combining data modalities may also be useful for predicting other outcome measures. This requires further investigation.

V. CONCLUSION

TB Portals is a large, multi-source, and cross-domain dataset. This study presented the challenges and methods used when working with this real-world data. Initially, the cohort size

under analysis underwent a substantial reduction due to missing data and the combination of several inclusion/exclusion criteria. Moreover, given the sparsity and high dimensionality inherent in the genomic data, an aggregation technique was employed at the gene level to facilitate the encoding of categorical data for machine learning purposes. Lastly, the issue of imbalanced data arose as TB Portals exhibited a skew towards DR-TB, necessitating the utilization of oversampling techniques to balance the minority class.

To construct the predictive models, the radiological findings and genomic information were incorporated for (1) differentiating between DS-TB and DR-TB, (2) predicting the first successful treatment period, and (3) predicting the treatment period of a specific drug regimen combination. The first model, incorporating solely genomic features, achieved an average accuracy of 92.4% and retained the same performance when radiological features were added. The second model achieved a relative error of 25.6% using only genomic features and reduced it to 22.0% when radiological information was added. Finally, the third model achieved a relative error of 17.8% using radiological features, 19.9% using genomic features exclusively, and 19.0% using both feature types.

ACKNOWLEDGMENT

The authors would like to thank Dr. Gabriel Rosenfeld and Dr. Andrei Gabrielian for their input and feedback.

REFERENCES

[1] A. Rosenthal et al., "The TB portals: An open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis," *J. Clin. Microbiol.*, vol. 55, no. 11, pp. 3267–3282, Nov. 2017.

[2] *Global Tuberculosis Report 2021*, World Health Org., Geneva, Switzerland, 2022.

[3] K. R. Wollenberg, C. A. Desjardins, A. Zalutskaya, V. Slodovnikova, A. J. Oler, M. Quiñones, T. Abeel, S. B. Chapman, M. Tartakovsky, A. Gabrielian, S. Hoffner, A. Skrahin, B. W. Birren, A. Rosenthal, A. Skrahina, and A. M. Earl, "Whole-genome sequencing of Mycobacterium tuberculosis provides insight into the evolution and genetic composition of drug-resistant tuberculosis in Belarus," *J. Clin. Microbiol.*, vol. 55, no. 2, pp. 457–469, Feb. 2017.

[4] K. Wollenberg, M. Harris, A. Gabrielian, N. Ciobanu, D. Chesov, A. Long, J. Taaffe, D. Hurt, A. Rosenthal, M. Tartakovsky, and V. Crudu, "A retrospective genomic analysis of drug-resistant strains of M. Tuberculosis in a high-burden setting, with an emphasis on comparative diagnostics and reactivation and reinfection status," *BMC Infectious Diseases*, vol. 20, no. 1, pp. 1–12, Dec. 2020.

[5] M. Moreno-Molina, N. Shubladze, I. Khurtsilava, Z. Avaliani, N. Bablishvili, M. Torres-Puente, L. Villamayor, A. Gabrielian, A. Rosenthal, C. Vilaplana, S. Gagneux, R. R. Kempker, S. Vashakidze, and I. Comas, "Genomic analyses of mycobacterium tuberculosis from human lung resections reveal a high frequency of polyclonal infections," *Nature Commun.*, vol. 12, no. 1, pp. 1–11, May 2021.

[6] G. Rosenfeld, A. Gabrielian, Q. Wang, J. Gu, D. E. Hurt, A. Long, and A. Rosenthal, "Radiologist observations of computed tomography (CT) images predict treatment outcome in TB portals, a real-world database of tuberculosis (TB) cases," *PLoS ONE*, vol. 16, no. 3, Mar. 2021, Art. no. e0247906.

[7] F. Yang, H. Yu, K. Kantipudi, M. Karki, Y. M. Kassim, A. Rosenthal, D. E. Hurt, Z. Yaniv, and S. Jaeger, "Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features," *Quant. Imag. Med. Surg.*, vol. 12, no. 1, pp. 675–687, Jan. 2022.

[8] F. Yang, H. Yu, K. Kantipudi, A. Rosenthal, D. E. Hurt, Z. Yaniv, and S. Jaeger, "Automated drug-resistant TB screening: Importance of demographic features and radiological findings in chest X-ray," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2021, pp. 1–4.

[9] M. Karki, K. Kantipudi, H. Yu, F. Yang, Y. M. Kassim, Z. Yaniv, and S. Jaeger, "Identifying drug-resistant tuberculosis in chest radiographs: Evaluation of CNN architectures and training strategies," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2964–2967.

[10] M. Karki, K. Kantipudi, F. Yang, H. Yu, Y. X. J. Wang, Z. Yaniv, and S. Jaeger, "Generalization challenges in drug-resistant tuberculosis detection from chest X-rays," *Diagnostics*, vol. 12, no. 1, p. 188, Jan. 2022.

[11] S. Jaeger, O. H. Juarez-Espinosa, S. Candemir, M. Poostchi, F. Yang, L. Kim, M. Ding, L. R. Folio, S. Antani, A. Gabrielian, D. Hurt, A. Rosenthal, and G. Thoma, "Detecting drug-resistant tuberculosis in chest radiographs," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 12, pp. 1915–1925, Dec. 2018.

[12] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.

[13] K. M. Keyes and D. Westreich, "U.K. Biobank, big data, and the consequences of non-representativeness," *Lancet*, vol. 393, no. 10178, p. 1297, Mar. 2019.

[14] F. Liu and D. Panagiotakos, "Real-world data: A brief review of the methods, applications, challenges and opportunities," *BMC Med. Res. Methodol.*, vol. 22, no. 1, p. 287, Nov. 2022.

[15] T. Zeng, T. Huang, and C. Lu, "Cross-domain analysis for 'all of us' precision medicine," *Frontiers Genet.*, vol. 12, Jul. 2021, Art. no. 713771.

[16] C. H. Lee and H.-J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, Mar. 2017.

[17] K. A. Cohen et al., "Evolution of extensively drug-resistant tuberculosis over four decades revealed by whole genome sequencing of mycobacterium tuberculosis from KwaZulu-natal, south Africa," *Int. J. Mycobacteriol.*, vol. 4, pp. 24–25, Mar. 2015.

[18] C. A. Desjardins, K. A. Cohen, V. Munsamy, T. Abeel, K. Maharaj, B. J. Walker, T. P. Shea, D. V. Almeida, A. L. Manson, A. Salazar, N. Padayatchi, M. R. O'Donnell, K. P. Mlisana, J. Wortman, B. W. Birren, J. Grosset, A. M. Earl, and A. S. Pym, "Genomic and functional analyses of mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance," *Nature Genet.*, vol. 48, no. 5, pp. 544–551, May 2016.

[19] H. Zhang et al., "Genome sequencing of 161 mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance," *Nature Genet.*, vol. 45, no. 10, pp. 1255–1260, Oct. 2013.

[20] M. R. Farhat et al., "Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis," *Nature Genet.*, vol. 45, no. 10, pp. 1183–1189, Oct. 2013.

[21] N. Casali, V. Nikolayevskyy, Y. Balabanova, S. R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. Bryant, J. Parkhill, S. Nejentsev, R. D. Horstmann, T. Brown, and F. Drobniewski, "Evolution and transmission of drug-resistant tuberculosis in a Russian population," *Nature Genet.*, vol. 46, no. 3, pp. 279–286, Mar. 2014.

[22] T. R. Ioerger, S. Koo, E.-G. No, X. Chen, M. H. Larsen, W. R. Jacobs, M. Pillay, A. W. Sturm, and J. C. Sacchettini, "Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-natal, south Africa," *PLoS ONE*, vol. 4, no. 11, p. e7778, Nov. 2009.

[23] J. E. Phelan, D. M. O'Sullivan, D. Machado, J. Ramos, Y. E. A. Oppong, S. Campino, J. O'Grady, R. McNerney, M. L. Hibberd, M. Viveiros, J. F. Huggett, and T. G. Clark, "Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs," *Genome Med.*, vol. 11, no. 1, pp. 1–7, Dec. 2019.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[25] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2011.

[26] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[27] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[28] M. Karki, K. Kantipudi, B. Haghighi, V. Bui, F. Yang, H. Yu, M. Harris, Y. M. Kassim, D. E. Hurt, A. Rosenthal, Z. Yaniv, and S. Jaeger, "Training data for machine learning to enhance patient-centered outcomes research (PCOR) data infrastructure—A case study in tuberculosis drug resistance," Nat. Library Med., Bethesda, MD, USA, Tech. Rep. ASPE IA 750119PE080057, 2022.

VY C. B. BUI received the B.S., M.S., and Ph.D. degrees in electrical engineering from The Catholic University of America, Washington, DC, USA, in 2012, 2014, and 2020, respectively. From 2016 to 2022, she worked on medical image computing research with the Cardiovascular CT Laboratory, National Heart, Lung, and Blood National Institutes of Health, Bethesda, MD, USA. She is currently with the Applied Clinical Informatics Branch, National Library of Medicine, National Institutes of Health, Bethesda. Her research interests include image processing, computer vision, and deep learning for biomedical image analysis.

ZIV YANIV received the Ph.D. degree in computer science from The Hebrew University of Jerusalem, Israel. He is currently a Senior Imaging Scientist with the Bioinformatics and Computational Biosciences Branch NIAID/NIH and Guidehouse. Over the past 20 years, he has conducted research in image-guided surgical interventions, medical image analysis, and more recently, microscopy image analysis. He believes in the curative power of open research and has been involved in development and leadership of free open source software, including the Image-Guided Surgery Toolkit, the Insight Registration and Segmentation Toolkit, and SimpleITK.

MICHAEL HARRIS received the bachelor's degrees in computer science and in mathematics and the master's degree in applied mathematics from the University of Maryland, College Park. He is currently a Specialist in biomedical informatics with the Bioinformatics and Computational Biosciences Branch (BCBB), Office of Cyber Infrastructure and Computational Biology (OCICB), National Institute of Allergy and Infectious Diseases, which is a part of the National Institutes of Health (NIH). At NIAID, he conducts research in genomics, clinical informatics, and data visualization.

FENG YANG received the B.S. and M.S. degrees from Northwestern Polytechnical University, China, in 2005 and 2007, respectively, and the Ph.D. degree from the National Institute of Applied Science (INSA Lyon), France, in 2011. She joined the Lister Hill National Center for Biomedical Communications (LHNCBC), National Library of Medicine (NLM), in October 2017. She is currently a Research Fellow with NLM, National Institutes of Health (NIH). She was a Principal Investigator and an Associate Professor with Beijing Jiaotong University, China, from 2012 to 2019. Her current research interests include machine learning and artificial intelligence-based biomedical image processing and analysis, and cardiac image processing. She has so far published more than 90 research papers, including 40+ journal articles, one book chapter, and 50+ conference papers. She has been the Special Session chairperson of IEEE ICSP, from 2012 to 2022.

KARTHIK KANTIPUDI received the bachelor's degree from IIT Kharagpur, and the master's degree in data science from the University of St Thomas. He is currently an Imaging Specialist with the Bioinformatics and Computational Biosciences Branch (BCBB), Office of Cyber Infrastructure and Computational Biology (OCICB), National Institute of Allergy and Infectious Diseases, which is a part of the National Institutes of Health (NIH). At NIAID, he conducts research in deep learning and machine learning involving medical images and clinical information.

DARRELL HURT received the B.S. degree (Hons.) in chemistry (computational emphasis, physics minor) from Brigham Young University and the combined M.S./Ph.D. degree in chemistry (biophysical emphasis) from Cornell University, under the supervision of Dr. Jon Clardy. He serves as the Chief of the Bioinformatics and Computational Biosciences Branch (BCBB), Office of Cyber Infrastructure and Computational Biology (OCICB), National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), USA. The Branch provides a wide variety of bioinformatics products and services to the NIAID intramural and extramural community, including the recent formation of the African Centers of Excellence (ACE) in bioinformatics and data science, the first instance of which resides in Bamako, Mali. He did postdoctoral work in lipid signaling and cell trafficking using X-ray crystallography with Dr. James Hurley with NIH.

ALEX ROSENTHAL received the B.S. and M.S. degrees in applied mathematics from Baku State University, Azerbaijan, in 1989, and an executive M.B.A. from Loyola University, Baltimore, in 2010. He has been leading projects on behalf of the NIH since 1994, and currently serves as the NIAID's Chief Technology Officer, where he provides executive leadership to a wide spectrum of clinical informatics, electronic content management, bioinformatics, application development, and information technology activities. He also provides executive oversight to the TB Portals program - leading collaborations between NIAID and multiple countries with MDR/XDR tuberculosis burden.

STEFAN JAEGER received the Diploma degree in computer science from the German Research Center for Artificial Intelligence (DFKI), University of Kaiserslautern, and the Ph.D. degree in computer science from the University of Freiburg. He is currently a Staff Scientist with the Lister Hill National Center for Biomedical Communications, United States National Library of Medicine (NLM), which is a part of the National Institutes of Health (NIH). He has held research positions in AI with the Chinese Academy of Sciences, University of Maryland, University of Karlsruhe, and Daimler, among others. At NLM, he leads a team that uses machine learning to diagnose infectious diseases and applies artificial intelligence and data science for clinical care and education. His other research interests include biomedical image analysis, medical informatics, and traditional medicine.