Automated Drug-Resistant TB Screening: Importance of Demographic Features and Radiological Findings in Chest X-Ray

Feng Yang
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, USA
feng.yang2@nih.gov

Hang Yu
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, USA
hang.yu@nih.gov

Karthik Kantipudi
Office of Cyber Infrastructure and Computational Biology
National Institute of Allergy and Infectious Diseases, National Institutes of Health
Bethesda, MD 20894, USA
karthik.kantipudi@nih.gov

Alex Rosenthal
Office of Cyber Infrastructure and Computational Biology
National Institute of Allergy and Infectious Diseases, National Institutes of Health
Bethesda, MD 20894, USA
alexr@niaid.nih.gov

Darrell E Hurt
Office of Cyber Infrastructure and Computational Biology
National Institute of Allergy and Infectious Diseases, National Institutes of Health
Bethesda, MD 20894, USA
darrellh@niaid.nih.gov

Ziv Yaniv
Office of Cyber Infrastructure and Computational Biology
National Institute of Allergy and Infectious Diseases, National Institutes of Health
Bethesda, MD 20894, USA
zivrafael.yaniv@nih.gov

Stefan Jaeger
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, USA
stefan.jaeger@nih.gov

2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) | 978-1-6654-2471-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/AIPR52630.2021.9762198

***Abstract*— Tuberculosis (TB) is a global disease caused by the bacillus Mycobacterium tuberculosis. In recent years, great progress has been made in care and control of drug-sensitive TB, whereas drug-resistant TB continues to be a worldwide public health problem that takes a heavy toll on both patients and the health care system. Early detection of drug resistance during a patient's first visit is very important because it enables appropriate drug treatment and thus reduces the period of infectiousness. However, discrimination between drug-resistant TB (DR-TB) and drug-sensitive TB (DS-TB) using imaging and readily available demographic data is still an open problem. In this paper, we investigate the possibility of automatic discrimination between DR-TB and DS-TB with demographic data and radiological findings from chest X-rays (CXRs) using machine learning techniques as well as the importance of such features for classifier training. We use a dataset of 1311 DR-TB cases and 1311 DS-TB cases from 10 countries, collected from the NIAID TB Portals program (https://tbportals.niaid.nih.gov). We first perform a two-step preprocessing, which consists of feature quantitation and missing data imputation. Seven demographic features and 25 radiological features are selected from the dataset. Then, we train a random forest (RF) model to evaluate the ability to differentiate between DR-TB and DS-TB. An importance index calculated from the RF model is used to analyze the feature importance with respect to the discrimination task. The importance index from the RF model shows that the top ten important factors for discriminating between DR-TB and DS-TB are: number of daily contacts, BMI, patient type, education, medium density infiltrate, medium density stabilized fibrotic nodules, low ground glass density infiltrate, pleural effusion percentage of hemithorax involved, multiple nodules, small nodules. Ten-fold cross-validation using the RF model shows that automatic discrimination between DR-TB and DS-TB achieves an average accuracy of 75% and an average AUC value of 83%, when using the top ten features. Our study suggests that automatic discrimination between DR-TB and DS-TB with demographic and radiological features is possible.**

Keywords—Tuberculosis (TB), drug resistance, random forest, differentiated diagnosis; demographic features; radiological findings

I. INTRODUCTION

Tuberculosis (TB), caused by the bacillus Mycobacterium tuberculosis, is a serious worldwide health issue with an estimated 10 million people infected and 1.5 million deaths each year [1]. In recent years, great progress has been made in care and control of drug sensitive TB [2], whereas drug resistant TB continues to be a worldwide public health problem [3]. In 2019, there were an estimated 10 million TB cases; approximately half a million cases are resistant to rifampicin, of which 78% are multidrug-resistant TB (MDR-TB) [1]. Drug-resistant TB is a growing public health concern since it requires more complex treatment than drug- sensitive TB and incurs more costs. Early detection of drug resistance is very important, as it helps with decision making, enables appropriate drug treatment, and reduces the period of infectiousness. However, discrimination between drug-resistant TB (DR-TB) and drug-sensitive TB (DS-TB) using imaging and readily available demographic data is still an open problem.

Previous works have shown evidence that certain clinical features can potentially aid in identification of DR-TB, such as prior treatment [4]–[8], positive sputum smear microscopy [5], history of drug injection [6], gender [6], [9], and age [7], [8]. Few works have dealt with radiological findings from chest imaging to identify the type of TB, DR-TB or DS-TB. Icksan *et al.* [10] reported that the MDR-TB group are more likely to have large-size lesions than DS-TB group. Wang *et al.* [11] found that

thick-walled multiple cavities (particularly with count ⩾3 and size ⩾30mm) present the most promising radiological sign for MDR-TB with good specificity but at the cost of low sensitivity. Huang *et al.* [12] reported that consolidated nodule number and size can be used to predict the probability of MDR-TB. Flores-Trevino *et al.* [13] found that multiple cavities is a promising predictor for DR-TB. Our previous work [14] found that the number of sextants with abnormalities is useful for discriminating between DR-TB and DS-TB. So far, very few works have been concerned with discriminating between DR-TB and DS-TB in an automated manner. [15]–[17] applied machine learning methods or deep learning methods on chest images to extract features for identifying DR-TB and DS-TB achieving AUC values of 72%, 66% and 85%, respectively.

In this work, we focus on demographic information and radiologist reported findings from patient records. We investigate the possibility of automatic discrimination between DR-TB and DS-TB with demographic and radiological features using machine learning techniques as well as evaluating the importance of such features in classifier training.

II. METHODS

A. Data collection

We use a dataset of 2622 patients, which includes de-identified clinical data and chest X-ray images publicly available from the NIAID TB Portals program [18]. Each patient record is manually annotated with clinical information and radiological findings using the chest X-ray images. Clinical information includes demographic features such as age of onset, gender, patient type (*New, Relapse* or *Failure*), BMI, country of origin, education, employment, number of daily contacts, number of children, and other information such as type of sample (pulmonary or extrapulmonary), prescription drugs, laboratory tests, treatment period, treatment status and outcome. A new case refers to a patient who has never been treated for TB or has taken anti-TB drugs for less than one month. A relapse case refers to a patient who has previously been treated for TB, was declared cured or completed treatment at the end of the most recent course of treatment, and is now diagnosed with a recurrent episode of TB (either a true relapse or a new episode of TB caused by reinfection). A failure case represents a patient who has previously been treated for TB and whose treatment failed at the end of the most recent course of treatment [18]. Radiological findings include chest radiography patterns such as nodules, cavities, infiltrates and collapses, the presence of mediastinal lymphadenopathy, presence of other non-TB abnormalities, the overall percentage of abnormal volume, and the pleural effusion percentage of the hemithorax involved. Due to financial constraints and the size of the TB portals CXR dataset, radiological features are obtained using a single experienced radiologist-reading per image. The whole dataset was annotated by multiple radiologists from the countries contributing data to the program. Consequentially, the radiological annotations are not biased towards a single radiologist. The 2622 patients include 1311 DS-TB and 1311 DR-TB patients, acquired from 10 countries.

B. Feature preprocessing

We perform a two-step preprocessing for demographic and radiological features. It consists of feature quantitation and missing data imputation. Feature quantitation indicates converting text features into numeric features. Missing data for a demographic feature is replaced by the mean value of other non-missing values under the same feature, while missing data for a radiological feature is assigned a special group number. For example, the radiological feature under the category *Overall Percentage of Abnormal Volume* will be assigned four values after feature quantitation and missing data imputation: 1 (0), 2 (<50%), 3 (>50%) and 4 (missing data).

Seven demographic features and 25 radiological features are selected by removing those whose missing data is more than 40% and by removing the country of origin from demographic features. Since almost 80% patients comes from five countries (Belarus, Georgia, India, Ukraine, and Kazakhstan), training on the country of origin may result in biased classification.

C. Random forest classifier

Based on the selected demographic and radiological features, we train a machine learning classifier, a Random Forest (RF) model [19], to discriminate between DS-TB and DR-TB. We illustrate the pipeline of our machine classification in Fig. 1. To compare the contributions of different features for classifying DR-TB vs DS-TB, we train the RF classifiers using different feature combinations.

Fig. 1. Pipeline of the RF-model-based classification beween DR-TB and DS-TB.

D. Importance measure

Each tree in a RF model is built from a random sample of the data, and not all observations are used to construct a specific tree. The observations that are not used to construct a tree are called out-of-bag (OOB) observations of this tree. In a RF model, each tree is built from a different sample of the original data, so each observation is "out-of-bag" for some of the trees.

Assuming that our RF model includes M decision trees $H=\{h_1, h_2, \ldots, h_M\}$. The importance index of a given predictor X_i is calculated using the following four steps.

Step 1: Use the decision tree h_m to predict its OOB observations. We refer th input matrix as $\mathbf{X}_{OOB}$ (feature matrix), and output matrix as $\mathbf{Y}_m$, then the prediction error $Err1$ can be calculated as the mean square error (MSE) between the predicted values $\mathbf{Y}_m$ and real values $\mathbf{Y}$:

$$Err1 = mean\,(\boldsymbol{Y_m} - \boldsymbol{Y})^2. \qquad (1)$$

Step 2: Permute values for the feature X_i (the ith column of the feature matrix) and use decision tree h_m to predict the OOB observations. Then, the prediction error $Err2$ can be calculated as:

$$Err2 = mean\,(\boldsymbol{Y'}_m - \boldsymbol{Y})^2. \quad (2)$$

Step 3: The importance index of predictor X_i on decision tree h_m is calculated as: $MSE_m = Err2\text{-}Err1$.

Step 4: The importance index of predictor X_i o the RF model is given by:

$$\alpha = \frac{1}{M}\sum MSE_m. \quad (3)$$

$\alpha > 0$ means X_i is important since changing its order makes the error larger; $\alpha = 0$ indicates that the order of X_i is not important since the MSE does not change; $\alpha < 0$ suggests that the variable can have a detrimental impact on the classification since changing its order makes the error smaller (substituting the feature with noise is better than the original feature; hence, the feature is worse than noise).

III. Experimental Results

Figure 1 shows the importance index calculated using Eq. (3) on seven demographic and 25 radiological features. We see that the top ten important factors for classifying DR-TB and DS-TB are: number of daily contacts, BMI, patient type, education, medium density infiltrate, medium density stabilized fibrotic nodules, low ground glass density infiltrate, pleural effusion percentage of hemithorax involved, multiple nodules, small nodules.

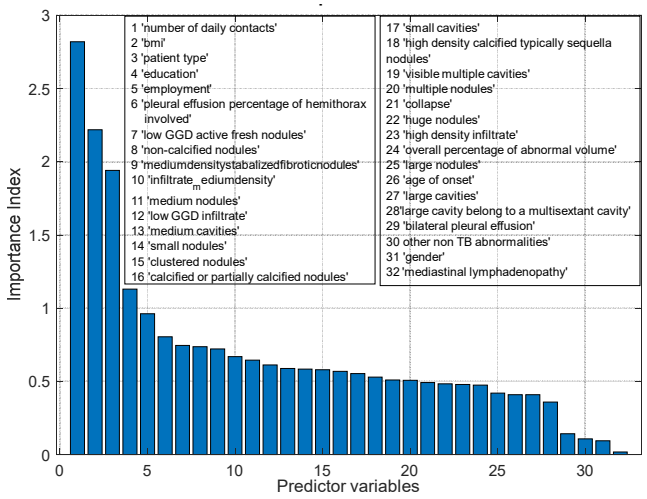

Fig. 2. Importance index of random forest model for the seven demographic features and 25 radiological features. Predictors: 1 - 'number of daily contacts', 2 - 'bmi', 3 - 'patient type', 4 - 'education', 5 - 'employment, 6 - 'pleural effusion percentage of hemithorax involved', 7 - 'low GGD active fresh nodules', 8 - 'non-calcified nodules', 9 - 'medium density stabalized fibroticn odules', 10 - 'infiltrate_medium density', 11 - 'medium nodules',12 - 'low GGD infiltrate', 13 - 'medium cavities', 14 - 'small nodules', 15 - 'clustered nodules', 16 - 'calcified or partially calcified nodules', 17 - 'small cavities', 18 - 'high density calcified typically sequella nodules', 19 - 'visible multiple cavities', 20 - 'multiple nodules', 21 - 'collapse', 22 - 'huge nodules', 23 - 'high density infiltrate', 24 - 'overall percentage of abnormal volume', 25 - 'large nodules', 26 - 'age of onset', 27 - 'large cavities', 28 - 'large cavity belonging to a multisextant cavity', 29 - 'bilateral pleural effusion', 30 - other non TB abnormalities', 31 - 'gender', 32 - 'mediastinal lymphadenopathy'. GGD indicates ground glass density.

To investigate the possibility of automatically differentiating between DR-TB and DS-TB and to evaluate the contribution of specific features, we trained RF models using the following combinations: 1) seven demographic features, 2) 25 radiological features, 3) 32 demographic and radiological features, and 4) top 10 important features. The results in Table 1 show that 1) demographic features have more influence on the RF model than radiological features; 2) the RF classifiers using top 10 features and using 32 features achieve very close performance, with an average AUC value of 83% and an average accuracy of 75%. Figure 2 shows the ROC curves for RF-based classifier using the top 10 features.

Table 1 RF classifier performance with ten-fold cross validation.

RF model features	Performance				
	AUC	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*
7 demog. features	81.09% ±2.52%	72.72% ±1.88%	76.05% ±4.13%	71.39% ±2.22%	72.67% ±1.48%
6 demog. without patient type	77.11% ±2.47%	72.16% ±2.78%	77.33% ±4.11%	66.59% ±3.07%	69.94% ±2.41%
25 radiol. features	64.89% ±3.81%	60.79% ±3.36%	68.65% ±3.68%	52.94% ±5.10%	59.39% ±3.07%
32 features	**82.86% ±3.49%**	75.03% ±3.05%	**78.33% ±5.25%**	69.72% ±4.40%	72.20% ±3.09%
Top 10 features	82.55% ±2.64%	**75.17% ±3.36%**	77.58% ±4.36%	**72.77% ±4.23%**	**74.04% ±3.76%**

Note: demog. indicates demographic, radiol. indicates radiological.

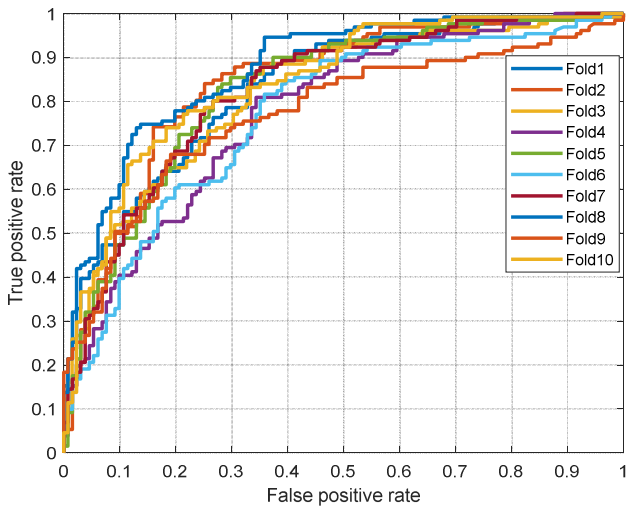

Fig. 3. ROC curves for ten-fold cross validation using random forest classifier based on the top 10 features.

IV. Discussion and Conclusion

In this paper, we investigated the importance of demographic and radiological features in discrimination between DS-TB and DR-TB and the possibility applying machine learning to discrimination between DR-TB and DS-TB by incorporating both features.

We select balanced DR-TB and DS-TB cases to avoid the bias of unbalanced dataset on machine classifier training and to avoid the unpredictable effects of synthetic data from

augmentation methods. It should be noticed that about 80% of the patients come from five countries (Belarus, Georgia, India, Ukraine, and Kazakhstan). That is, our machine classifier learns drug-sensitive and drug-resistant features primarily from five countries, and thus the classification performance will likely decrease when we use it to identify DR-TB from other countries or when we perform a country-level evaluation.

We observe from Table 1 that patient type plays an important role in discriminating between DR-TB and DS-TB, with specificity decreasing around 5% when removing patient type from the training features. This is probably due to the fact that most of the patients with patient types of *Failure* (95%) and *Relapse* (83%) are drug resistant.

Experimental results show that automated discrimination between DR-TB and DS-TB using a RF model achieves an AUC value of 83% and an accuracy of 75% with the top 10 demographic and radiological features. Our study suggests that automatic discrimination between DR-TB and DS-TB is possible by utilizing both demographic features and radiological features.

ACKNOWLEDGMENT

This work was supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under Interagency Agreement #750119PE080057, and by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health. This project has also been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases (NIAID) under BCBB Support Services Contract HHSN316201300006W/HHSN27200002.

REFERENCES

[1] WHO, "Global Tuberculosis Report 2020: Executive summary," 2020.

[2] CDC, "Combating the Global TB Epidemic," 2021. [Online]. Available: https://www.cdc.gov/globalhivtb/who-we-are/about-us/globaltb/globaltb.html.

[3] WHO, "WHO consolidated guidelines ontuberculosisModule 4: TreatmentDrug-resistant tuberculosis treatment," 2020.

[4] A. Faustini, A. J. Hall, and C. A. Perucci, "Risk factors for multidrug resistant tuberculosis in Europe: A systematic review," *Thorax*, vol. 61, no. 2, pp. 158–163, 2006.

[5] B. P. Tembo and N. G. Malangu, "Prevalence and factors associated with multidrug/rifampicin resistant tuberculosis among suspected drug resistant tuberculosis patients in Botswana," *BMC Infect. Dis.*, vol. 19, no. 1, pp. 1–8, 2019.

[6] N. Mdivani *et al.*, "High prevalence of multidrug-resistant tuberculosis in Georgia," *Int. J. Infect. Dis.*, vol. 12, no. 6, pp. 635–644, 2008.

[7] X. Shen *et al.*, "Drug-resistant tuberculosis in Shanghai, China, 2000-2006: Prevalence, trends and risk factors," *Int. J. Tuberc. Lung Dis.*, vol. 13, no. 2, pp. 253–259, 2009.

[8] X. T. Lv, X. W. Lu, X. Y. Shi, and L. Zhou, "Prevalence and risk factors of multi-drug resistant tuberculosis in Dalian, China," *J. Int. Med. Res.*, vol. 45, no. 6, pp. 1779–1786, 2017.

[9] M. R. O'Donnell *et al.*, "Extensively drug-resistant tuberculosis in women, Kwazulu-Natal, South Africa," *Emerg. Infect. Dis.*, vol. 17, no. 10, pp. 1942–1945, 2011.

[10] A. G. Icksan, M. R. S. Napitupulu, M. A. Nawas, and F. Nurwidya, "Chest X-ray findings comparison between multi-drug-resistant tuberculosis and drug-sensitive tuberculosis," *J. Nat. Sci. Biol. Med.*, vol. 9, no. 1, pp. 42–46, 2018.

[11] Y. X. J. Wáng, M. J. Chung, A. Skrahin, A. Rosenthal, A. Gabrielian, and M. Tartakovsky, "Radiological signs associated with pulmonary multi-drug resistant tuberculosis: An analysis of published evidences," *Quant. Imaging Med. Surg.*, vol. 8, no. 2, pp. 161–173, 2018.

[12] X.-L. Huang *et al.*, "Prediction of multiple drug resistant pulmonary tuberculosis against drug sensitive pulmonary tuberculosis by CT nodular consolidation sign," *bioRxiv*, 2019.

[13] S. Flores-Treviño *et al.*, "Clinical predictors of drug-resistant tuberculosis in Mexico," *PLoS One*, vol. 14, no. 8, 2019.

[14] F. Yang *et al.*, "Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features," *Quant. Imaging Med. Surg.*, vol. 0, no. 0, pp. 1–16, 2021.

[15] V. Kovalev *et al.*, "Utilizing radiological images for predicting drug resistance of lung tuberculosis," in *Proceedings - International Congress on Computer Assisted Radiology and Surgery*, 2015, no. JUNE, pp. S129–S130.

[16] S. Jaeger *et al.*, "Detecting drug-resistant tuberculosis in chest radiographs," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 12, pp. 1915–1925, 2018.

[17] M. Karki *et al.*, "Identifying Drug-Resistant Tuberculosis in Chest Radiographs : Evaluation of CNN Architectures and Training Strategies," in *Proceedings - 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2021)*, 2021, pp. 1–4.

[18] 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, 2017.

[19] A. Liaw and M. Wiener, "Classification and Regression with Random Forest," *R News*, vol. 2, pp. 18–22, 2002.