Abstract—— The NIAID TB Portals program was established as a multi-national collaboration to develop and support open-access information sharing and data harmonization through an online database platform, facilitating and spearheading on-going genomics, bioinformatics and image analysis projects. This image analysis study is motivated by the widely-noted increase in the incidence of drug resistant (DR) Tuberculosis (TB) at first patient visit as a major health concern and a challenge for physicians. The chest X-Ray (CXR) remains the primary screening and diagnostic tool for TB. We used TB Portals collection of CXR data for the analysis of similarity score based on automatically extracted features of images. Further we applied similarity metrics to study the relationships to tuberculosis drug resistance. Our analysis shows possible relationships between image similarity and drug resistance category. Using k-nearest neighbor (KNN) classification we achieved an accuracy of 73% using only histogram of intensities as the image descriptor.

Index Terms—CXR; tuberculosis; similarity; CBIR; resistance; k-nearest neighbor; MDR-TB, XDR, Sensitive

I. Introduction

Tuberculosis (TB) is one of the leading causes of death worldwide. During 2015, more than 10 million people fell ill with TB and 1.8 million died from the disease. World Health Organization statistics indicate that the crisis of multiple drug resistant TB (MDR-TB) detection and treatment continues.

During 2015, of the estimated 580,000 people newly eligible for MDR-TB treatment, only 20% were enrolled. Of cases treated specifically for MDR/XDR-TB in 2013, the success rate was only 52%, and 17% resulted in death. Nearly 10% of MDR-TB cases are extremely drug-resistant TB (XDR-TB), not responding to most second-line treatment drugs, for which treatment success was an even lower 26% [14].

During 2016, a short duration MDR-TB treatment was introduced that includes five drugs for a minimum of 9 to 12 months [13]. This is a long and challenging treatment program that further underscores the needs for rapid diagnosis, improved prognosis and assessment of relapse cases. Nonetheless patients are being cured with these treatment plans. The challenge remains to capture enough clinical cases to optimize treatment interventions and improve outcomes for this segment of the overall TB patient population.

This project is part of TB Portals [4], which is a multinational research program, initiated and supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health. It connects tuberculosis clinical centers in countries with heavy burden of drug-resistant TB with the aim of collecting and making available anonymized clinical data, biological samples and radiological images. TB Portals data contain many important clinical features – age, gender, case definition, etc. -- as well as important clinical categories of outcome and drug resistance. One significant Program emphasis is image analysis study and collaboration. TB Portals Program data were used as a reference set in the development and testing of image similarity algorithms.

This work is motivated by the current conditions in combating TB, the need for early diagnostic tests, and to enable physicians from areas of the world that do not have direct experience with drug resistant TB to easily obtain de-identified and physician curated patient case data from countries that experience a significant drug resistant TB burden. It is noted that although advanced genomic tests exist for early drug resistance testing, they remain a significant cost burden. In addition, there are clinical images in hospitals that are not annotated, but that are related to patient clinical data and outcome.

With the information known to the physician, at an initial visit, can a CXR be utilized either separately or in conjunction with the available and evident clinical features to identify, not only the important clinical manifestations of the disease, but also whether the patient might represent a case of drug resistance disease? In order to analyze whether TB Portals collection of CXR could be used as a reference and predict drug resistance category for user submitted images we capitalized on detailed patient case annotation including outcome, drug resistance, together with the CXR to create test and validation cohorts.

TB Portals Program database contains both X-Ray (n=496) and CT images (n=675). Roughly half of these images are annotated by Radiologist; however, clinical data is available for all patient cases. The Radiologist annotates lung specific features such as cavities, calcination, etc. These features and
combinations of features can be used to define cohorts of clinical cases having common clinical manifestation. However, for images that have not been annotated by the Radiologists, the only way to leverage these images in the creation of cohorts is to use image similarity. If most similar images indeed share the same clinical features then it makes sense to develop an automated procedure where new, unannotated images can be compared to the set of reference images. As a result of such a procedure the end user will get a list of similar images from TB Portals that can be studied for efficient patient treatment.

II. Existing Systems

Many papers have been written about content based image retrieval systems (CBIR). For example, in [1] the authors describe an algorithm that uses two phases: first, a set of images are retrieved using similarity measures; and second, an algorithm to retrieve clusters of images that are in the neighborhood of the query image.

A survey that reviewed current implementations for CBIR systems discusses a feature extraction module for every image, a distance (similarity) computation module, and a set of algorithms to retrieve the similar images [2]. In another survey, the authors analyze several CBIR systems utilizing pixel information of images, and assessing the similarity between the query image and the images database with various algorithms. [5]

The design of efficient indexing systems is also a focus for CBIR research [7]. New tools are required to index, store and retrieve medical images efficiently [6].

The latest literature sources use machine learning tools to improve medical images processing, clustering, sorting, and classification. The tasks such as segmentation and classification use deep learning which does not need a predefined feature set. Those techniques are gradually and increasingly incorporated into medical image retrieval systems [9,10,11,12].

III. Method

For the purpose of this study, we randomly selected 50 patient cases with drug sensitive TB and 50 patient cases with drug resistant TB. To accurately model the patient-physician encounter only the first visit CXR was utilized.

“Drug resistant” TB included both multiple drug resistant tuberculosis (MDR-TB) and extensively drug resistant tuberculosis (XDR-TB) cases. The WHO defines MDR-TB as resistance to at least both isoniazid and rifampicin; and extensive drug resistance (XDR) as resistance to any fluoroquinolone, and at least one of three second-line injectable drugs (capreomycin, kanamycin and amikacin), in addition to multidrug resistance. The “sensitive” class is limited to those patient cases that respond to all TB drugs. [3]

A. The Hypothesis

Our working hypothesis: there exists similarity between images that belong to the same resistance category. The most similar images to a “drug resistant” CXR will come from the reference set of “drug resistant” CXRs. Likewise, the most similar images for a “sensitive” CXR should come from the reference set of “sensitive” images. If our hypothesis is true, image similarity can help in primary diagnosis aimed at revealing drug resistant TB at the time of first patient visit.

![Figure 1: Data preparation process](image)

B. Image Preparation

Figure 1 described the workflow of activities performed. First, we selected two data sets of 50 images for both resistance categories. The images were segmented to analyze only the lungs area. The next procedure was done in two steps: we generated histograms for intensities in every image; and then we computed histogram of gradients (HOG) and local binary patterns(LBP). Processing every pixel of the images, we generated the intensity histograms. We used histograms with 32 bins, and with 256 bins.

We also used a combination of LBP and HOG because improves the detection performance considerably on some datasets [19,20,21]. While LBP is a description of local image texture, HOG helps to model appearance and shape in an image.

The extracted features allow us to represent an image as a vector of numerical values. The feature vectors were used as input to the KNN algorithm.

We used an implementation of KNN available as described in [8]. The KNN algorithm was run three times: one for histogram of 32 bins, one for the histogram of 256 bins; and one for the file with information about LBP and HOG. To establish the optimum number of neighbors required for correct classification, we tested the top 1, 3, 5, 7 and 10 similar images using the simple majority rule. Similarity was calculated as the Euclidian distance between feature vectors. The closer the vectors the more similar were considered the corresponding images.

Finally, the accuracies for the three indices were compared. If the hypothesis (above) is correct then nearest neighbors to “resistant” vectors will be from the same resistance category.
IV. Results

We concluded that information provided by calculating image similarity based solely on intensity histograms predicted tuberculosis drug resistant category.

Image similarity based on texture information (LBP and HOG) did not predict better resistance category as can be seen in Figure 4.

As evident from Figures 2 to 4 the best performance was achieved using 256 bins intensity histograms reaching 73% of correct predictions when using 10 nearest neighbors.

V. Discussion

As we have shown, image similarity alone predicted drug resistance category with a reasonable accuracy of 73%. The concept of similarity promises to enable valuable contributions by providing a unique means to query databases and retrieve patient case data for patients with similar CXR.

The TB Portals Program determined to utilize the similarity function between CXR developed web application and service to work with both the existing content of TB Portals and with user submitted novel CXRs. Within the framework of the TB Portals Program research we have implemented a version of images similarity search for a database analysis tool called the Data Exploration Portal, or DEPOT. [18]
Within DEPOT the similarity function exists within an ecosystem of patient-centric case data relating clinical, socioeconomic, genomic and image annotation data. The similarity matrix for all versus all images in the database is pre-computed and stored.

Figure 5 shows a screen capture for one patient with summary case information. The DEPOT allows the user to examine CXR image with the standard viewer and a sampling of clinical data as shown.

If the user wants to further explore the content of TB Portals he can use image similarity as shown in Figure 6. From similar images, one can examine the treatment drug regiments, drug resistance test results, and other relevant clinical and genomic data. The achieved performance of the similarity algorithm does not allow one to recommend it as a diagnostic tool; however, it may provide the user with helpful reference patient cases.

The DEPOT user may upload an X-ray image and search for many similar images. The corresponding case data for each similar image record can be reviewed in detail, suggesting the type of drug resistance, treatment regiment, and listing socioeconomic, clinical and genomic data.

For both above mentioned scenarios, the similarity function enables the creation of cohorts of images. The user may exclude non-interesting patient cases from the cohort one image at a time. The user can create and save multiple cohorts in this manner. Statistical analysis within DEPOT enables the comparison of two cohorts across hundreds of clinical, genomic and image annotation data in minutes. Easily readable graphical results ordered by the statistical significance of all features are provided. Output is available in print and Excel formats as well.

Querying TB Portals for similar images thus enables physicians and medical students with limited exposure with drug resistant TB patients to easily obtain and study de-identified and physician curated patient cases from countries with significant drug resistant TB burden. The TB Portals Program team continues to search for improvements to the similarity algorithms. TB Portals is an ongoing and growing multi-country collaboration and new patient cases are added to the database which may result in further improvements in similarity based classification schemes.

The available repositories of clinical imaging data are quickly growing making evident the need for the development of computer-assisted tools taking advantage of information that is otherwise difficult to systematically explore. Connecting visual exploration with algorithmic analysis of big clinical data will assist physicians in advancing personalized medicine in the increasingly connected digital world.

Appendix

The data set used in this paper is available [15,16]. All CXRs were represented as histograms of intensity after segmentation removed non-lung areas of the CXR [17].

Acknowledgment

This work is supported under BCBB Support Service. Acknowledge the guidance and leadership of Michael Tartakovsky, CIO, and Alexander Rosenthal, CTO, OCICB NIAID NIH.

References


[18] NIAID TB Portals Data Exploration Portal, or DEPOT: https://depot.tbportals.niaid.nih.gov/


[21] Chandrasekhar et al. CHoG: Compressed Histogram of Gradients - A low bit rate feature descriptor, CVPR.