Training Data for Machine Learning to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure — A Case Study in Tuberculosis Drug Resistance

Final Report

National Library of Medicine

National Institute of Allergy and Infectious Diseases

Training Data for Machine Learning to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure — A Case Study in Tuberculosis Drug Resistance

Final Report

Manohar Karki[1], Karthik Kantipudi[2], Babak Haghighi[1], Vy Bui[1], Feng Yang[1], Hang Yu[1], Michael Harris[2], Yasmin M. Kassim[1], Darrell E. Hurt[2], Alex Rosenthal[2], Ziv Yaniv[2], and Stefan Jaeger [1]

[1]Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
8600 Rockville Pike, Bethesda, MD 20894

[2]Office of Cyber Infrastructure and Computational Biology
National Institute of Allergy and Infectious Diseases
National Institutes of Health
8600 Rockville Pike, Bethesda, MD 20894

March 7, 2023

Preface

This report is based on research conducted by the National Library of Medicine (NLM) in collaboration with the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH). The research was supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OSPCORTF), which is coordinated by the Assistant Secretary for Planning and Evaluation (ASPE), under Interagency Agreement #750119PE080057, during the period from August 1, 2019, to September 30, 2022. This work was also supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, and had been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases under BCBB Support Services Contract HHSN316201300006W/HHSN27200002. Furthermore, this research was supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the National Library of Medicine. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-SC0014664.

All opinions, findings, and conclusions expressed in this paper are the authors' and do not necessarily reflect the policies and views of NIH, NLM, DOE, or ORAU/ORISE. Therefore, no statement in this report should be construed as an official position of NLM/NIH or of the U.S. Department of Health and Human Services.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

NLM/NIH or U.S. Department of Health and Human Services endorsement of any derivative products that may be developed from this report, such as clinical practice guidelines, other quality enhancement tools, or reimbursement or coverage policies, may not be stated or implied.

Suggested citation: Karki et al. Training Data for Machine Learning to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure — A Case Study in Tuberculosis Drug Resistance. Final Report. National Library of Medicine, November 2022.

Contents

Executive Summary

This report describes the research plan and results of a PCOR project that followed a statement of work defined in an intra-agency agreement (IAA) between the Office of the Assistant Secretary for Planning and Evaluation (ASPE) and the National Library of Medicine (NLM). ASPE coordinates efforts to build data capacity for patient-centered outcomes research (PCOR). As part of these efforts, ASPE and NLM collaborated on a project about detecting tuberculosis drug resistance using artificial intelligence and machine learning. The project goal was to create a foundation to advance the use of artificial intelligence (AI) for Patient-Centered Outcomes Research (PCOR) and clinical practice, using existing and to be acquired TB Portals data from the National Institute of Allergy and Infectious Diseases (NIAID). The data of the TB Portals program provided by the Office of Cyber Infrastructure and Computational Biology (OCICB), NIAID, offers valuable training data for machine classifiers. The data allows researchers to train classifiers that discriminate between drug-resistant and drug-sensitive tuberculosis based on socioeconomic, geographic, clinical, laboratory, radiographic, and genomic data.

Machine learning is a type of AI where a computer uses training data sets composed of large and varied amounts of data to "learn" how to identify patterns with little human intervention. Industry experts have acknowledged that large amounts of high-quality training data are a critical part of the foundation that will support researchers' use of machine learning to accelerate the discovery of novel disease-outcome correlations and inform the design of prevention and treatment studies. High-quality training data sets that are well-labeled and structured, use standard data models and common data elements annotated by domain experts, and combine previously unconnected data resources that can be used to train algorithms to elucidate knowledge and extract relevant data points for research. AI and associated innovative technologies like machine learning have the power to consume large amounts of data in varied, complex formats to more quickly identify effective treatments, potentially accelerating clinical innovation by speeding up the research lifecycle and the application of evidence in clinical settings. This project established a foundation for researchers to use AI to develop scientific approaches so healthcare providers can match patients to the best treatments based on their specific health conditions, life experiences, and genetic/phenotypic profiles.

The project was executed as part of a larger FY2019 OS-PCORTF project, in which NLM worked on the drug-resistant tuberculosis use case. The general objective was to enhance the capacity of PCOR researchers to use machine learning by developing and disseminating several resources that will present not only training data and methods

but also lessons learned from the processes and implementation. The project curated high-quality training sets of quality clinical research data collected with NIAID. This included clinical images such as frontal chest X-rays (CXR) and computed tomography scans (CT), clinical and socioeconomic data, and genomic pathogen information of thousands of patients with drug-sensitive and drug-resistant tuberculosis (TB). These training data sets were used to develop, train, and improve machine learning models for detecting TB drug resistance. Building the capacity of researchers to compare the health outcomes of innovative approaches in delivering and managing care for TB supports the tenets outlined in the OS-PCORTF funding priority, value-based care and health outcomes. A consistent mission goal is to predict drug resistance in a patient early on and administer the appropriate patient-specific drugs for more efficient treatment. Successful implementation of this idea would be a significant breakthrough in the fight against drug-resistant TB and could save many lives.

Deliverables were made available via TB Portals and a GitHub software repository, including training data, machine learning algorithms, and trained models. The tools and knowledge generated from this project will help PCOR researchers to produce findings that could impact clinical practice. The results will encourage those building upon this project's deliverables to develop similar use cases in other areas. Furthermore, as regulatory agencies develop national policies that increasingly consider patient-generated information in the approval of drugs and devices, evidence generated from the application of machine learning to patient-centered outcomes research will be beneficial.

1 Introduction

This project aimed to advance the use of artificial intelligence (AI) for Patient-Centered Outcomes Research (PCOR) and clinical practice. According to the Office of the Assistant Secretary for Planning and Evaluation (ASPE), *"Patient-centered outcomes research is designed to produce new scientific evidence that informs and supports the health care decisions of patients, families, and their health care providers. PCOR studies focus on the effectiveness of prevention and treatment options with consideration of the preferences, values, and questions patients face when making healthcare choices. The validity of PCOR findings is strengthened by a robust data infrastructure within HHS agencies that supports rigorous analyses and generates relevant findings that help inform decisions."* Please see the following page for a current portfolio of PCOR projects and their contributions to building data capacity that meets current HHS priorities, https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf. Specifically, this project was concerned with providing training data for machine learning to enhance PCOR data infrastructure.

The project was part of a larger FY 2019 project supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF), which featured the National Library of Medicine (NLM) and the National Coordinator for Health Information Technology (ONC) as project leads. While NLM developed a use case for drug-resistant tuberculosis (TB), as outlined in this report, ONC worked in parallel with ASPE on a second kidney-related use case. Both projects shared the same objective and common goal of enhancing the capacity of PCOR researchers to use machine learning by developing and disseminating several resources, including training data, training methods, and lessons learned. Further details regarding the kidney use case can be found in the final report by ONC.

In collaboration with the Office of Cyber Infrastructure and Computational Biology (OCICB) at the National Institute of Allergy and Infectious Diseases (NIAID), NLM collected high-quality training sets for TB drug resistance. These sets comprise clinical images in the form of X-rays and computed tomography scans (CTs), clinical data, and genome information about the pathogen for drug-sensitive and drug-resistant TB. During the project, the collected sets have been used to train various machine learning models, such as TB vs. not TB or drug-sensitive vs. drug-resistant classifiers. Furthermore, the data was used to evaluate and improve the performance of these classifiers for TB drug resistance detection.

NLM has a long tradition of carrying out research in machine learning and computational image analysis for biomedical image processing. Many staff scientists at NLM

are subject matter experts with strong publication records in these areas, including TB screening. Therefore, NLM was well set up to develop machine learning algorithms for detecting drug resistance in radiographs, using the latest techniques available, and to accomplish the project goal and the specific tasks described in this report. All activities of NLM were coordinated with NIAID so that all data and algorithms developed are consistent with TB Portals. NIAID provided images and clinical data, and NLM curated the data for machine learning and developed and tested machine learning models with this data.

Building the capacity of researchers to compare the health outcomes of innovative approaches in delivering and managing care for TB supports the tenets outlined in the OS-PCORTF funding priority, value-based care and health outcomes. Therefore, a consistent mission goal is to predict drug resistance in a patient early on and administer the appropriate patient-specific drugs for more efficient treatment. Successful implementation of this idea would be a significant breakthrough in the fight against drug-resistant tuberculosis and could save many lives.

2 Background and Rationale

NLM is part of the National Institutes of Health (NIH), which supports the generation and analysis of substantial quantities of biomedical research data, as outlined in the agency's strategic plan for data science. This includes the integration of clinical and observational data into biomedical data science. One of the ongoing large-scale efforts to build high-value resources that include patient clinical and observational data is the TB Portals Program (TBPP) initiated by NIAID, also part of NIH [22]. TBPP is a multi-national collaboration for data sharing and analysis to advance TB research. A consortium of clinicians and scientists from countries with a heavy burden of TB, especially drug-resistant TB, work with data scientists and information technology professionals to collect multi-domain TB data and make it available to clinical and research communities.

TB is a serious worldwide public health threat [9]. It is an airborne disease caused by Mycobacterium tuberculosis (MTB) bacteria, which were first discovered in 1882. Tuberculosis is a curable infection, and the worldwide TB mortality rate is decreasing due to a global effort to improve TB control and treatment. However, even today, after the development of advanced medical treatment and diagnostic technology, TB is the leading cause of death from infectious diseases worldwide. Approximately ten million people are estimated to have TB, which claims around 1.5 million lives each year.

Curing TB requires prolonged treatment with several drugs. There are currently more than 20 drugs in use against TB. The five most commonly used drugs, first-line drugs, are typically used for TB patients without prior TB drug treatment. This includes the drugs isoniazid, rifampicin, and three others. It is essential that several TB drugs are taken together to avoid becoming resistant to an individual drug. To avoid drug resistance, it is also very important that a patient adheres strictly to the treatment regimen over several months without interruptions. The drugs for the treatment of drug-resistant TB are more expensive and have more side effects. They are grouped according to their effectiveness and use experience and belong to the so-called second-line drugs, reserve drugs for treating drug resistance.

A patient with TB is drug-sensitive if the bacteria causing the infection respond to all drugs. Suppose a patient has contracted drug-resistant TB, either by direct transmission from another infected person or due to improper treatment. In that case, the TB bacteria will not respond to at least one of the primary drugs. Two main types of drug resistance are MDR-TB and XDR-TB. MDR-TB, or multidrug-resistant TB, is defined as resistance to at least isoniazid or rifampicin, two of the most effective first-line TB drugs. XDR-TB, or extensively drug-resistant TB, is caused by bacteria that, in addition to resistance against isoniazid or rifampicin, are resistant to several other drugs, including at least one of the second-line drugs. While these are the two main types of TB drug resistance, additional categories are sometimes used depending on the number of drugs to which the bacteria do not respond, including resistance against individual drugs and resistance against most existing drugs. Treatment of the latter is extremely difficult. In 2019, according to the World Health Organization (WHO), nearly half a million people developed rifampicin-resistant TB (RR-TB). Globally, 3.3% of new TB cases and 17.7% of previously treated cases had MDR/RR-TB [9].

MDR-TB is concerning because it is difficult to diagnose, and it takes more time, often more than two years, and costs to treat patients. One of the significant challenges for controlling MDR-TB lies in diagnosing drug resistance of TB-suspected patients during their first visit. Conventionally, drug susceptibility testing is performed on a sputum sample to identify the resistant status to several drugs, which requires a well-equipped laboratory facility and takes four to six weeks to obtain the laboratory results. The development of Xpert MTB/RIF, a real-time test based on polymerase chain reaction (PCR) for genetic mutations in the MTB genome associated with resistance, specifically Rifampicin (RIF) resistance, has reduced the laboratory time needed for the detection of MDR-TB. However, the test still produces many inconclusive results, and its deployment in resource-constrained settings is expensive. In addition, the test requires a sputum sample, which can be difficult to obtain, especially from children.

Therefore, detecting MDR-TB is still challenging, and the conventional chest X-ray (CXR) remains a valuable tool in the detection, screening, and surveillance of MDR-TB, thanks to its widespread availability.

TBPP is a web-based, open-access repository of socioeconomic/geographic, clinical, laboratory, radiographic, and genomic data from drug-resistant TB patient cases with linked physical samples [22]. Country sites in Eastern Europe, Central Asia, and sub-Saharan Africa share data through TBPP. Representatives from participating countries comprise the TBPP Steering Committee, which provides feedback and guidance on developing TBPP resources, tools, and new research studies. The program aims to add advanced analytical tools to perform domain-specific and meta-domain analyses of TBPP data, for example, using modern machine learning and computational image analysis. The data and tools brought together by TBP will enable researchers to answer important questions about diagnosing drug-resistant TB and treating TB patients. With the recent advances in computational methods for quantitative imaging, in particular, the application of powerful deep learning methods on sizeable medical image sets, automatic screening for infectious diseases has shown promising results.

Quantitative imaging allows computer-assisted detection methods to assess infectious disease severity and response to therapy. It is an interdisciplinary field involving clinical research of infectious diseases under various imaging modalities, including CXR, computed tomography (CT), and others. It can leverage radiology to detect and treat emerging pathogens, epidemics, and pandemics. Therefore, this project aims to develop new ways of detecting drug resistance in chest radiographs by using deep learning in combination with large image repositories and clinical data. The long-term goal is to create an efficient and accurate system that can discriminate between drug-resistant and drug-sensitive TB by detecting abnormal patterns predictive of treatment outcomes [17].

3 Objectives

This section describes the project's purpose and lists the key objectives and deliverables based on the original statement of work.

3.1 Project Purpose

The project aimed to acquire and curate radiographic images (CXR and CT) for machine learning and make this data publicly available to the research community via the NIAID TB Portals Program (TBPP). The TBPP platform engages researchers in con-

ducting patient-centered outcomes research related to drug-resistant tuberculosis and precision treatment. While the total number of tuberculosis cases has decreased over the last few years, the rise of drug-resistant tuberculosis has reduced the chance of controlling the disease. TBPP supports research in the decision-making and timely diagnosis of drug-resistant tuberculosis, which is essential to administering adequate treatment regimens and stopping the further transmission of drug-resistant tuberculosis.

In particular, the project addressed the following objectives:

1. Developing high-quality training data sets for machine learning to detect drug-resistant TB in radiographs in collaboration with NIAID.

2. Developing and evaluating machine learning methods for radiographic, clinical, and genomic data to discriminate between drug-resistant and drug-sensitive TB.

3. Disseminating resources, including training data, tools, and lessons learned to stimulate the application of these methods to a broader array of use cases by PCOR researchers.

4. Developing an implementation guide detailing each method used and the generic aspects of the data that each method leverages to facilitate applications to a broader array of use cases.

3.2 Relevance for Patient-Centered Outcomes Research

The project supported PCOR and comparative effectiveness by providing diagnostic tools based on machine learning and computational image analysis. Computational models are trained by evaluating the effectiveness of current treatments to learn better ways of delivering care to potentially drug-resistant TB patients. The idea was to learn early signs of drug-resistant TB infections by exploiting pictorial information from radiographs and clinical, patient-specific socioeconomic, and genomic data. Identifying drug resistance at an early stage of a TB infection allows using more appropriate drugs for treatment, leading to more effective treatments and faster recovery of patients. This is a critical PCOR problem as drug resistance in patients infected with tuberculosis has been identified as a significant global public health concern projected to expand in coming years; see also this WHO link for more information on TB drug resistance, https://www.who.int/activities/tackling-the-drug-resistant-tb-crisis (last accessed in December 2022).

The project supported federal and Department of Health and Human Services (HHS) investments in the Precision Medicine Initiative (PMI) and TBPP by using AI to identify optimal patient-specific treatments. Applying robust machine learning to PCOR

and the general healthcare space depends on large, high-quality training data sets for modeling data representations.

Using the deliverables of this project, PCOR researchers can execute and test the provided models on new data or retrain new models with their data. This may lead to improved models providing more insights. The project enhanced the capacity of PCOR researchers to use machine learning by developing and disseminating several resources that present not only training data and methods but also lessons learned from the execution of the tasks. Specifically, this project provided radiographs and clinical data of thousands of patients with drug-sensitive and drug-resistant TB, which have been made available via TBPP at NIAID. The results of this project, including the lessons learned, have been shared at conferences and public meetings to facilitate the application of the methods to a broader range of use cases.

3.3 Deliverables

The deliverables for the objectives fell into three categories:

1. **Developing high-quality training data (Objective 1)** This deliverable involved collecting training data for machine learning, including X-rays, CTs, and clinical data, in standard formats, such as DICOM or CSV files, suitable for training classifiers to detect TB drug resistance.

2. **Developing machine learning classifiers (Objective 2)** This class of deliverables defined several classification tasks to discriminate between drug-resistant and drug-sensitive TB and detect resistance types. In particular, the following classification tasks were addressed: normal vs. abnormal (TB vs. not TB for radiographs), drug-sensitive vs. drug-resistant TB, and MDR vs. XDR.

3. **Dissemination of resources, results, and lessons learned (Objective 3 and 4)** The third category of deliverables required disseminating the developed training data and software tools. It also asked for the presentation of the results in webinars and publications in conference proceedings and journal papers.

Figure 1 shows a summary of the deliverables defined in the inter-agency agreement with ASPE.

	Objective	**Deliverable**
1	Develop high-quality training data sets for machine learning in detecting drug-resistance tuberculosis in radiographs, in collaboration with NIAID.	Training data with X-rays, CTs, clinical data, socioeconomic information, genomic data, and statistical meta information in common formats used by the imaging and machine learning community
2	Develop machine learning algorithms that will be trained and tested on this data for discriminating between drug-resistant and drug-sensitive tuberculosis. Evaluate algorithms using conventional metrics.	Trained machine learning models for five binary TB classification problems, according to Task 2, including performance evaluation: i normal vs. abnormal ii drug-sensitive vs. drug-resistant iii MDR vs. XDR (and other resistance type pairs) iv drug-sensitive vs. drug-resistant, with clinical data v drug-sensitive vs. drug-resistant, with genomic data
3	Disseminate resources (tools and training data) and lessons learned to stimulate the application of these methods to a wider array of use cases by PCOR researchers.	i Publication of data, training models, and training scripts on TBPP and/or other websites as agreed with ASPE ii 2 webinars iii 2 presentations at conferences/-workshops
4	Develop implementation guides detailing each method used and the generic aspects of the data each method leverages, with detail sufficient to facilitate its application to a wider array of use cases.	Three publications will conclude the project: i Implementation guide available for researchers in a conventional format, with information on training and testing scripts, to be available via TBPP ii Scientific manuscript for submission to a peer-reviewed journal on detecting drug resistance with machine learning iii 508 compliant final report for posting on ONC and/or ASPE website.

Table 1: Objectives and Deliverables

4 Background - Problems Addressed

This section describes the problems addressed by the project and how this will advance the PCOR field, following the original statement of work. Specifically, it tells the research questions that can now be answered due to the project.

4.1 Data Acquisition

At the start of the project, there was little data available about drug-resistant TB. Although NIAID had already acquired X-rays and CTs for TB Portals in several countries [22], and the ImageCLEF 2017 and 2018 evaluation challenges had contributed some data [13], this was insufficient to train machine classifiers for detecting and predicting TB drug resistance. Since TB drug resistance is a complex medical condition requiring individual patient-centered treatment, many patient records are needed to arrive at statistically meaningful results. In tight collaboration, NIAID (OCICB) and NLM addressed this problem by significantly increasing the amount of data during the project. The data contains radiographs and clinical, genomic, and socioeconomic records. This makes this data unique and allows the training of deep learning networks.

To make the data public, the data needed to be de-identified, cleaned and saved in standard formats, such as DICOM. Furthermore, the data needed to be organized in a way suitable for machine learning. For example, the type of drug resistance, patient age and sex, and temporal information in the case of longitudinal data needed to be included. As important were data links connecting patient records with the corresponding genome data of the pathogen, which will enable future PCOR research across domains. Specifically, the data will help investigate what radiographic, clinical, and genomic factors contribute to TB drug resistance, how they are related, and how they affect recovery chances. More information on the mechanism and the general data sharing and analysis, including radiographs and genomic data, can be found here: https://jcm.asm.org/content/55/11/3267.

4.2 Machine Learning Classifiers

No reliable machine classifiers for TB drug resistance detection were available at the start of the project. The project addressed this problem by a) carrying out machine classification experiments on the newly acquired data and making the resulting models publicly available and b) providing a software interface allowing users to train their models on TBPP data or their data. Several classification tasks and PCOR questions for detecting drug resistance were addressed in this project:

- The project trained machine classifiers to detect manifestations of TB in radiographs (TB vs. not-TB).
- Using the newly acquired TB Portals data, the project trained classifiers to discriminate between drug-sensitive and drug-resistant TB in radiographs. This addresses the PCOR question of whether specific TB manifestations in a patient's lung indicate drug resistance, which is an open research problem.
- By training classifiers for MDR vs. XDR drug resistance, the project investigated the possibility of predicting different types of drug resistance based on the TBPP data, which is also an open research problem.
- The project used clinical and socioeconomic features, alone and in combination with TBPP's pictorial radiographs, to investigate if machine classifiers can predict drug resistance for a patient based on this information.
- Another problem addressed was the potential existence of links between phenotype (radiographs) and genotype (pathogen genome), which is largely unknown. The project tried to find connections and trained classifiers for drug resistance prediction using radiographic and genomic features.

All statistical models produced for these classification tasks were systematically tested based on standard performance metrics. The implemented machine classifiers provided to the research community will allow PCOR researchers to measure the relative usefulness of socioeconomic, clinical, and genomic data for TB drug resistance prediction.

4.3 Dissemination of Results

To enable PCOR researchers to continue the research of this project toward a definite answer to the questions above, disseminating training data, trained models, and lessons learned were deliverables of this project. While the training data has been made available via TB Portals (https://tbportals.niaid.nih.gov), the software, including programming interfaces, libraries, training algorithms, and trained models, have been made available via GitHub (https://github.com/NLMLHC/NLM-PCOR-TB). In addition, an implementation guide provides more information about the implemented software methods used to develop the deliverables of this project, including the software environment, which will encourage the application of these methods to TBPP data or other PCOR data. PCOR researchers can use this as a starting point to modify the software deliverables and apply them to other PCOR use cases, asking similar questions and following the best practices identified through this work.

Conference presentations and webinars were given to disseminate the results and lessons learned from the work; see Section 6 for more information. The scientific results of this project were summarized in several peer-reviewed publications for journals and conference proceedings. They provide details about the algorithms used and their performance evaluation; see Section 6 for a complete list of papers published during this project.

5 Methodology

Machine learning was the primary method used in this project for detecting and predicting drug resistance. It can be considered a subfield of artificial intelligence (AI), where a computer uses training data to develop statistical models that can be used to classify unknown patterns. The training data comprises many samples showing the patterns in all their variations so that training algorithms can learn the essential features and make inferences on unseen patterns, classifying them. Machine learning has seen considerable progress in the last decade, especially in developing convolutional neural networks (CNNs) for image analysis and recognition. These networks specialize in processing data with grid-like topologies, such as images [18]. This is also called deep learning and includes analyzing biomedical images for computer-aided diagnosis. For radiographs, for example, neural networks can detect manifestations of TB and perhaps drug resistance, as investigated in this project. Neural networks could also translate clinical patient data into drug resistance predictions. For computer-aided diagnosis, performances approaching or even outperforming the performance of human experts have been reported. However, large volumes of high-quality training data are critical to training classifiers and achieving high performance. Reliable machine classifiers can accelerate the discovery of novel disease-outcome correlations and inform the design of prevention and treatment studies, which has the potential to speed up the research life cycle and the application of evidence in clinical settings. High-quality training data sets are extensive and varied, covering all the different real-world occurrences of patterns. They are well-labeled and structured, use standard data models and common data elements annotated by domain experts, and combine previously unconnected data resources that can be used to train algorithms to elucidate knowledge and extract relevant data points for research.

The software to train deep learning networks for the TB drug resistance use case has been made available as a deliverable of this project. This software allows PCOR researchers to train statistical models to detect tuberculosis drug resistance, either on the data made available for this project or on their data following the same format.

Furthermore, using the software, researchers can easily import data, read the annotation labels, train a neural network or another machine classifier, let the classifier infer a class label, and evaluate the performance of the output. PCOR researchers can add their non-public training data to the existing data in TBPP and train new models based on larger training sets. Alternatively, they can directly apply the pre-trained models developed in this PCOR project to classify unknown input patterns without retraining the models. With only slight modifications, it would be possible to apply the software to other use cases for other lung diseases or entirely different image domains, for example, histopathology or retinal images. In summary, the project offers a machine learning software platform for the PCOR community to research radiographs and clinical data, with an emphasis on predicting TB drug resistance to determine the optimal treatment of a TB patient.

6 Accomplishments by Final Deliverables

This section lists the project accomplishments according to the objectives described in Section 1 and 3. It describes the most important research results and summarizes the software functionality for each deliverable. Most of the software modules offer two main functions, one for training and one for inference. The training function allows users to use their data to train a model. The inference function can execute a pre-trained model on TBPP data or a model that the user has trained on new data. For a more detailed description of the research results and the lessons learned, readers are referred to the research papers published during the project. A more extensive description of each software module can be found in the implementation guide, a deliverable of this project.

6.1 Training Data - Objective 1

A Data Use Agreement and a Memorandum of Understanding were signed between NLM and NIAID. This gave access to the existing NIAID TBPP data to train machine classifiers [22]. Under the leadership of NIAID, TBPP is designed to unite radiological, genomic, clinical, laboratory, and socioeconomic/geographic data from prospective and retrospective TB cases and their associated clinical samples and to freely share these curated data, including powerful and user-friendly analytical tools, with researchers and health care specialists throughout the world. This mission is accomplished through (i) international collaboration with clinical research sites and academic research organizations in countries with a heavy burden of drug-resistant TB and (ii) the creation,

support, expansion, and promotion of the multifactor repository of anonymized clinical data from patients with drug-resistant forms of TB.

TBPP is still in the data acquisition phase and includes both data acquired before and during this project. It should be noted that TBPP data differs in two main aspects from other data sources. First, it only contains data for patients with TB. Second, it is a much richer data source. For each patient, it includes clinical and socioeconomic information (e.g., treatment regimens, outcome, education level, etc.), imaging studies with manual and AI-generated labels (CXR and possibly CT), and finally, the pathogen's genomic information and the type of TB in terms of drug susceptibility (drug-sensitive or specific type of drug-resistant TB (DR-TB) such as Mono DR, Poly DR, and XDR).

The PCOR project collected 4,947 frontal chest X-rays (CXR) and 178 CT images with TB drug resistance information. Table 2 shows the number of patients from whom CXRs or CTs have been collected each project year. This data has been added to TB Portals, increasing the total data size to more than 6,000 patient records as of November 2022. The data includes clinical data, radiographs, patient information, and links to genomic information. It was the primary source of training data used in the experiments. Before using it for the experiments, the data needed to be cleaned up; for example, non-frontal chest X-rays and images of poor quality were removed. Lung fields were extracted using image segmentation.

Access to TBPP data is managed. To access the data, users must sign a data use agreement, describe their research project and agree not to distribute the data or attempt to identify patients. Once approved, the data is available in two forms: (a) download from a server using a dedicated protocol for the transfer of big data, which is more appropriate for images (b) use of a RESTful API to retrieve the tabular data (all data except images and genomic information). The former is updated quarterly and the latter weekly.

For a more detailed overview of TBPP data sharing and usage examples, see: https://datasharing.tbportals.niaid.nih.gov/.

	CXR	CT
2022	204(DS), 262(DR)	1(DS), 3(DR)
2021	955(DS), 1891(DR)	2(DS), 24(DR)
2020	527(DS), 1108(DR)	56(DS), 92(DR)
Total	1686(DS), 3261(DR)	59(DS), 119(DR)

Table 2: TBPP imaging data acquired during the project. The table shows the number of patients with drug-sensitive (DS) and drug-resistant (DR) TB manifesting in either a CXR or CT.

6.2 Machine Learning Algorithms - Objective 2

For this project, several machine classifiers were trained for different classification problems related to drug resistance using TBPP data; see also Objective 2 in Table 1. Specifically, classifiers were developed to discriminate between not-TB and TB cases and between drug-resistant and drug-sensitive cases using CXRs from about a dozen countries (Objective 2, Deliverable i, ii). The project also classified different types of drug resistance and used clinical and genomic data as additional information (Objective 2, Deliverable iii, iv, v). The following subsections provide more details about these classifiers.

6.2.1 TB vs. not TB

For CXR images, the project employed deep learning models to classify TB vs. not-TB (Objective 2, Deliverable i). Using an Xception network and pre-trained weights from ImageNet, the implementation of batch normalization and dropout methods avoided overfitting. The training was based on a cross-entropy loss function in combination with an Adam optimizer, using the publicly available Shenzhen and Montgomery County CXR sets as training data.

Because the classification of TB vs. not-TB required CXRs with TB and CXRs with no TB or with manifestations of other diseases for training, this classification problem was not the project's primary focus. TBPP is concerned exclusively with TB drug resistance and, thus, with radiographs featuring manifestations of TB. Nevertheless, the software provided as one of the deliverables allows users to apply the trained model or to train new models for this classification problem. For TB detection in CT images, please see the following publication [19].

6.2.2 Drug-sensitive vs drug-resistant

Several deep learning architectures were tested to discriminate between drug-sensitive and drug-resistant TB in CXRs, including AlexNet, InceptionV3, DenseNet169, and InceptionResNetV2 (Objective 2, Deliverable ii). Using 10-fold cross-validation and image augmentation to balance the data, AlexNet provided the highest accuracy of 75%, followed by InceptionV3 [18]. To visualize the results, GRAD-CAM heatmaps localized the areas a trained network deemed significant to distinguish between drug-resistant and drug-sensitive TB. The project also compared the abnormality distribution among sextants, extracted from the clinical data and radiological features, with the information extracted from the GRAD-CAM visualization to find correlations [17, 18].

For CT images, nnUNet (a self-configuring algorithm for deep learning-based biomedical image segmentation) detected lung regions on publicly available datasets, including lung regions in CTs from TBPP [14]. A lung segmentation performance of 0.98 was achieved, as measured by the Dice coefficient. Furthermore, the bounding boxes of these lung regions served as input to the training stage for drug-resistant vs. drug-sensitive TB classification in CTs, which resulted in roughly 60% AUC (area under the receiver operating characteristic curve - ROC).

6.2.3 MDR vs XDR

The project addressed MDR vs. XDR drug resistance classification to investigate the possibility of discriminating between both types with machine learning (Objective 2, Deliverable iii). Using seven demographic and 25 radiological features, a Random Forest Model (RF) was trained based on 592 MDR and 592 XDR cases from TBPP. The model achieved an average accuracy of 63% and an AUC of 69%.

The extent to which this classification problem is solvable with the given data in TB Portals remained unclear. Suppose MDR and XDR can be discriminated by computational means on radiographs and clinical data; in that case, more data would likely be needed to achieve a better classification performance.

6.2.4 Clinical data and radiological features

The relationship between lung nodules and drug resistance was a subject of investigation (Objective 2, Deliverable iv). Several nodule features in the clinical data were used to train a classifier that can discriminate between drug-resistant TB and drug-sensitive TB. The following nodule features were used in particular: number of nodules <3mm, number of nodules within 3-8mm, number of calcified or partially calcified nodules, number of non-calcified nodules, and number of clustered nodules. Altogether, 26 radiological features consisting of 17 nodule/cavity-related features and nine other lung abnormality features were used as the feature set in the experiments. After training a machine classifier, a support vector machine (SVM), with these features, including age and gender, the classifier achieved an average accuracy of 76% and an average AUC of 78% [26]. These results suggested that nodule features are possible indicators of drug resistance.

The project also investigated the role of demographic features in distinguishing drug-resistant TB from drug-sensitive TB [27]. In this study, seven demographic features were used: gender, patient type (New, Relapse, or Failure), BMI, education, employment, number of daily contacts, and number of children. For 2622 patients (1311 DS-TB +

1311 DR-TB), a combination of demographic and radiological features performed better than using only one feature type. The best performance was achieved by training a Random Forest classifier using the demographic and 26 radiological features, which provided an average accuracy of 75% and an average AUC of 83%. These results suggested that demographic features are meaningful indicators of drug resistance.

Another study investigated the relationship between the number of sextants affected by lesions and TB drug resistance. Pearson's chi-squared test showed that two clinical features and 22 radiological features are significantly associated with drug-resistant TB. Ten-fold cross-validation using an SVM classifier achieved an average accuracy of 68.42% and an average AUC of 72.16% when combing three clinical features, twelve nodule-related sextant features, and six cavity sextant features. Therefore, the study suggested that the number of sextants affected by lesions could be a predictor of drug-resistant TB. Features were more frequent in upper sextants than in lower sextants. For some features, such as non-calcified nodules, clustered nodules, and multiple nodules, the regions seemed to differ between drug-sensitive and drug-resistant patients, which could be a subject of further research.

In another experiment, a decision tree model (Random Forest) was trained on 840 drug-sensitive and drug-resistant patients to compute a ranking for 27 features. According to this model, pleural effusion, huge nodules, and lung collapse are the most critical features for detecting drug resistance.

For a clinical data analysis on 2237 patients (782 DS-TB + 1455 DR-TB), the project confirmed that classifiers using a combination of clinical and radiological features perform better than those using only one type of feature [26]. Data augmentation achieved the best performance for a combination of features, with an average accuracy of 72.34% and an average AUC of 78.42%.

Furthermore, the project analyzed risk factors for the recovery time of drug-sensitive and drug-resistant patients based on a Cox proportional hazard model. The result showed eight statistically significant factors, such as the number of very large lung nodules and the type of drug resistance, among others.

6.2.5 Genomic data

The project used machine learning on radiological and genomic TBPP data to discriminate between drug-sensitive and drug-resistant TB (Objective 2, Deliverable v). While radiological data consisted of radiographs annotated by radiologists, genomic data comprised a detailed breakdown of the genomic composition of the pathogen. Several experiments were performed for 4048 patient records, among which 2023 included genomic

features, and 1163 had both radiological findings from CXRs and genomic features.

The first experiment investigated the performance of radiological and genomic features for drug resistance prediction. Only one radiological feature (percent of pleural effusion of involved hemithorax) was mildly correlated with one of the gene mutation variants. Other radiological and genomic features were weakly correlated. Furthermore, eight genomic features were strongly correlated with DR-TB. Seven genomic features were mildly correlated with DR-TB. In contrast, radiological features were all weakly correlated with DR-TB. Pearson's chi-squared test showed that 14 out of 18 radiological features and 29 out of 43 genomic features were statistically significant regarding DR-TB vs. DS-TB ($p < 0.05$).

A Random Forest algorithm was used to classify DR-TB and DS-TB. The trained classification model achieved an average accuracy of 97% when combining radiological and genomic features or using genomic features alone. At the same time, the model's average accuracy based on radiological features alone was around 70%. This result showed that radiological and genomic features could be used to predict DR, with the genomic features alone or a combination of radiological and genomic features providing the best performance.

In another experiment, pleural effusion was analyzed in more detail, as it is the only radiological feature negatively correlated with one of the gene variants in the previous experiment. Unfortunately, pleural effusion was not a good predictor of drug resistance.

Furthermore, machine learning was used to predict the period of a successful treatment, again using both radiological and genomic features. Similar to prior experiments, the connection between radiological/genomic features and treatment period (TP) was investigated. TP correlated well with the genomic but poorly with the radiological features. Pearson's chi-squared test showed that TP depends on nine of 18 radiological features and most gene variants. A Gradient Boosting regression model predicted the first successful drug regimen length. This regression model had a mean absolute error of 110 days using radiological and 93 days using both radiological and genomic features. Because the treatment length could span up to three years, an error prediction of up to three months encouraged future work.

For TBPP, the project found 29 drug combinations to treat DS-TB and 668 drug combinations to treat DR-TB. Machine learning was used to predict the duration of the first treatment period of a specific drug regimen. Specifically, the Gradient Boosting regression model was trained on the top-3 most popular drug combinations. The most popular drug cocktail in the TBPP data was Ethambutol, Isoniazid, Pyrazinamide, and Rifampicin. For this drug combination, the regression model had a mean absolute error of 34 days using radiological features, 36 days using genomic features, and 33 days using

both features. The relative error was 17.7% using the radiological features, 19% using the genomic features, and 17.6% using both feature sets. For the second most popular drug combination, Ethambutol, Isoniazid, and Rifampicin, the genomic features did not help improve the model prediction. The relative error was 8.1% using the radiological features, 11.7% using the genomic features, and 11.4% using both feature sets. This was attributed to the fact that most cases were DS-TB. The third most popular drug regimen was Bedaquiline, Clofazimine, Cycloserine, Levofloxacin, and Linezolid, for which the relative error was about 36% using the radiological features, 33% using the genomic features, and 38% using both feature sets.

Finally, the first culture change period (FCCP) was predicted using both radiological and genomic features. There was a weak correlation between FCCP and genomic and radiological features. The radiological features and most of the mutations were independent of FCCP ($p > 0.05$), indicating that this is a challenging problem that needs further investigation.

6.3 Presentations and Publications

As part of the outreach activities, according to Objective 3 of the deliverables, an "International Workshop on Using Artificial Intelligence to Identify Multi-Drug Resistant TB and AIDS-TB Co-infections in Radiographs" was organized in Shenzhen, China, in October 2019. The project was presented in a talk titled "Developing AI for drug resistance detection in radiographs. Is it possible?"

The results of this project and the lessons learned were presented at conferences and in journal papers. For example, at the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, a talk was given on "Identifying Drug-Resistant Tuberculosis in Chest Radiographs: Evaluation of CNN Architectures and Training Strategies," in which an AUC of 84% was reported for classifying drug-resistant vs. drug-sensitive TB based on CXR images, using different augmentation techniques, synthetic data, and other publicly available sources. The following lists all papers published during the project; see the references at the end of this report.

- Karki M, Kantipudi K, Yang F, Yu H, Wang Y.X.J, Yaniv Z, Jaeger S. Generalization challenges in drug-resistant tuberculosis detection from chest X-rays. MDPI Diagnostics, vol. 12, no. 1, pp. 188-210, 2022.

- Yang F, Yu H, Kantipudi K, Karki M, Kassim YM, Rosenthal A, Hurt DE, Yaniv Z, Jaeger S. Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features. Quantitative Imaging

in Medicine and Surgery, vol. 12, no. 1, pp. 675-687, 2022.

- Karki M, Kantipudi K, Yu H, Yang F, Kassim YM, Yaniv Z, Jaeger S. Identifying drug-resistant tuberculosis in chest radiographs: Evaluation of CNN architectures and training strategies. 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2964-2967, 2021.

- Yang F, Yu H, Kantipudi K, Rosenthal A, Hurt DE, Yaniv Z, Jaeger S. Automated drug-resistant TB screening: Importance of demographic features and radiological findings in chest X-ray. 50th IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1-4, 2021.

- Ma L, Wang Y, Guo L, Zhang Y, Wang P, Pei X, Qian L, Jaeger S, Ke X, Yin X, Lure F. Developing and verifying automatic detection of active pulmonary tuberculosis from multi-slice spiral CT images based on deep learning. Journal of X-Ray Science and Technology, vol. 28, no. 5, pp. 939-951, 2020.

- Zheng Q, Lu Y, Lure F, Jaeger S, Lu P. Clinical and radiological features of novel coronavirus pneumonia. Journal of X-Ray Science and Technology, vol. 28, no. 3, pp. 391-404, 2020.

The project was part of a webinar on "Balancing Access to Federal Data Sets with Enhancing Privacy and Security," which was organized by ASPE. Together with the project collaborators at NIAID, the presentation described the anonymization process of the TBPP data so that it can be used for research, addressing potential privacy problems and their solutions.

7 Lessons Learned and Future Directions

This section discusses the lessons learned as a consequence of this project and proposes directions for future research, following a more detailed presentation of the challenges in [17].

7.1 Encountered Challenges and Solutions

During this project, extensive experiments have been conducted to discriminate between DR-TB and DS-TB in radiographs, clinical patient records, and genomic pathogen data. Many of these experiments have either established or pushed the current state-of-the-art. The results suggest that detecting DR-TB and predicting treatment outcomes and

times could be possible. Nevertheless, the early detection of DR-TB in chest radiographs and clinical data remains a research subject.

For CXRs from TBPP, with image augmentation and the addition of synthetic and publicly available images, the project achieved a high classification accuracy when evaluating models on data from one source using cross-validation. However, a significant performance degradation could be observed when evaluating trained models on unseen data from a different source. The models did not generalize well on new data. This so-called domain shift problem is a problem for many machine learning systems and is thus a general AI problem. In [17], the project authors suggested a multi-task approach to overcome this problem, but more research is needed to alleviate generalization problems. Domain shift can be caused by differences in image acquisition and non-pathological and non-anatomical aspects interfering with the information relevant to DR-TB. The capability of an AI system to provide consistent performance across different hospitals and countries is vital for DR-TB detection and computer-aided diagnosis as a whole.

Multiple publications have described attempts to utilize images or clinical data to distinguish between DR-TB and DS-TB. Several studies have shown that radiological findings in a CT or CXR could help differentiate between the two classes. A literature review from 2018 [25] concluded that thick-walled multiple cavities are a valuable predictor for DR-TB, with reasonable specificity but low sensitivity. Another study [12] compared 183 DR-TB cases and 183 DS-TB cases, concluding that there were substantial differences in findings between the two classes regarding lesion size and morphology. A slightly larger study [11], which compared 468 DR-TB and 223 DS-TB cases, concluded that a combination of the number and size of consolidated nodules is a good predictor for DR-TB. Furthermore, a small study in [7] used data from 144 patients and found that the presence of multiple cavities is a good predictor for DR-TB. Then, a much larger study [5] compared 516 DR-TB and 1030 DS-TB cases, obtaining an AUC of 0.83 using a regression model. This study observed that the co-existence of multiple findings was indicative of DR-TB. Finally, our more recent work compared 1455 DR-TB and 782 DS-TB cases, using two clinical features and 23 radiological findings [26]. An SVM classifier was used to distinguish between DR-TB and DS-TB with an AUC of 0.78. One limitation of these approaches is the reliance on a radiologist's reading.

Several fully automated solutions were presented in the ImageCLEF 2017 and 2018 evaluation challenges [13]. These challenges included differentiating between DR-TB and DS-TB using thoracic CT images. Among the proposed approaches, Gentili et al. [8] reformatted the CT images to the coronal plane and used a pre-trained ResNet50 CNN for classification. For the same challenge, Ishay et al. [15] used an ensemble of 3D CNNs, Cid et al. [6] used a 3D texture-based graph model with SVMs, and Allaouzi et

al. [2] replaced the softmax function of a 3D CNN architecture with an SVM. All entries had limited success, resulting in AUCs of about 0.6. After two editions, the organizers discontinued the TB drug resistance challenge, concluding that it was impossible to solve based only on the image.

While the observations made by the radiologists seemed encouraging, the challenges did not yield the desired results, suggesting that the sub-optimal performance may be due to the small number of images made available for training. However, increasing the number of CTs for this task is not trivial considering that the use of CT imaging for DS-TB cases is uncommon, with the standard imaging modality being CXR. For CXR images, the authors in [24] used a customized CNN architecture to classify DR-TB and DS-TB in 2973 images from TBPP. They achieved a classification performance of 66%, which improved to 67% when follow-up images were also included. Several authors of this report have previously proposed fully automated methods for CXRs, as described in [16, 18]. In [16], using 135 CXRs from a single source and a shallow neural network resulted in an AUC of 0.66. In [18], a more extensive set was used, containing 3642 images from multiple sources. Using an InceptionV3 deep neural network pre-trained on ImageNet, an AUC of 0.85 was obtained for this set, which is the current state-of-the-art performance achieved with TBPP data. It is a significant improvement compared to other approaches. However, even though 10-fold cross-validation was performed, the ability of the trained network to classify CXRs from unseen domains was not evaluated.

The common weakness of most current automated methods is that they have not been evaluated on a separate data source. As different imaging technologies and devices produce images with different features and qualities, models must be robust to these differences. An underlying assumption of most machine learning algorithms is that training and test data are independent and identically distributed. The learned parameters will not perform well if the two distributions are different. That is, the trained model will not generalize well on unseen data. While CXR imaging is a low-cost modality in widespread use, the image feature variations in the acquired images are significant [20, 4], bringing into question the utility of any proposed method that is not evaluated on its generalization capability. Sathitratanacheewin et al. [23] observed that a model for CXR-based TB diagnosis performed well with 0.85 AUC when tested on images within their intramural dataset but observed a significant performance deterioration of the model when tested on extramural images, yielding an AUC of only 0.7. Harris et al. [10] reported that 80% of published works using CXR for TB diagnosis either used the same databases for training and testing or did not comment on the databases they used for testing. For domain shift, when the distributions between training and test sets differ, it has been shown that formulating training as a multi-task

approach can reduce performance degradation [21]. In addition to domain shift, the generalization capability of deep learning methods can also deteriorate when they learn irrelevant features. Deep learning algorithms can pick up features in the training set that are arbitrarily correlated with the disease and, thus, are entirely unrelated. For example, these features can stem from the imaging devices' different characteristics or clinical practices implemented, such as patient positioning [28, 3, 1]. If a model has incorporated such features, it will not generalize well for different imaging devices or clinical workflows, which are irrelevant to disease diagnosis. In [17], authors of this report explored various strategies to improve the generalization of models for the classification of CXRs as DR-TB or DS-TB using different normalization and attention mechanisms, both explicit (segmentation-based) and implicit (multi-task based).

7.2 Ongoing Governance

The training data collected and the machine learning tools developed during this project have been made publicly available. As stated above, the data is available via TBPP (https://tbportals.niaid.nih.gov) [22], and the software is downloadable via GitHub (https://github.com/NLMLHC/NLM-PCOR-TB). The TBPP data repository and the GitHub software repository will receive regular updates in the foreseeable future. TBPP will continue to be serviced and maintained by NIAID. Its patient records were collected with the primary focus on acquiring drug-resistant cases, especially cases that reflect the specific research interest in the country of origin. As a result, there is an imbalance between drug-resistant and drug-sensitive cases, which does not necessarily reflect the prevalence of TB from either class in a contributing country. More data will be added to TBPP over time to a) increase the overall volume of TBPP and b) reduce these imbalances. This will increase the importance of TBPP, which is already the most significant source of data regarding TB drug resistance. Adding more data will make TBPP more useful for machine learning and, thus, more interesting for PCOR researchers. Depending on projects and research results, NLM will update the software in the GitHub repository accordingly. This could mean updating the models to incorporate new data added to TBPP or rewriting current methods to reflect recent developments in machine learning. If desired, the GitHub repository could also be expanded with code from the PCOR community or other open-source developers.

7.3 Additional Clinical Areas

The methods developed in this PCOR project are generic because they are not restricted to TB or drug resistance detection. Several underlying inference engines are

based on CNNs, which have established themselves as successful tools for image analysis. Theoretically, they can detect various abnormalities or diseases in radiographs or biomedical images in general. For example, in light of the COVID-19 crisis, the project investigated how machine learning may help detect this disease in radiographs (CXRs and CTs). Techniques similar to methods in this PCOR project can potentially detect COVID-19 and increase the specificity of COVID-19 diagnosis. Specifically, lung features were identified that could be seen by an AI tool, as published in a survey paper on clinical and radiological manifestations of coronavirus pneumonia [29]. Features of other lung abnormalities and diseases could be targeted similarly. In addition, computer-aided prediction of the treatment outcome and treatment length for other diseases can be accomplished with the software tools developed in this project. In this sense, the project has enhanced PCOR data infrastructure, providing a basis for future PCOR activities beyond TB drug resistance, and offering a platform for future research in AI generalizability and explainability.

8 Conclusion

Deliverables and resources of this project have been made available via existing infrastructure that researchers can access, such as GitHub and NIAID's openly-accessible TBPP for TB data sharing and analysis. Evidence generated from the application of the developed AI methods supports the TBPP platform by setting the foundation for researchers to develop scientific approaches so healthcare providers can match patients to the best treatments based on their specific health conditions and profiles.

The hope is that the results of this project will encourage those building upon this project's deliverables to advance these methods for the TB drug resistance use case, develop similar use cases in other areas, and make their research resources publicly available. Subsequent users of this project's tools, resources, and generated knowledge could include a larger array of PCOR researchers and healthcare stakeholders. For example, as regulatory agencies develop national policies that increasingly consider patient-generated information in the approval of drugs and devices, evidence generated from the application of machine learning to patient-centered outcomes research will be beneficial. Healthcare systems innovators and clinicians can use the tools developed and lessons learned to produce findings that could impact clinical practice.

References

[1] Kaoutar Ben Ahmed, Gregory M Goldgof, Rahul Paul, Dmitry B Goldgof, and Lawrence O Hall. "Discovery of a Generalization Gap of Convolutional Neural Networks on COVID-19 X-Rays Classification". In: *IEEE Access* 9 (2021).

[2] Imane Allaouzi and Mohamed Ben Ahmed. "A 3D-CNN and SVM for Multi-Drug Resistance Detection." In: *CLEF (Working Notes)*. 2018.

[3] Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. "Deep learning predicts hip fracture using confounding patient and healthcare variables". In: *NPJ Digit Med.* 2 (2019).

[4] Daniel C Castro, Ian Walker, and Ben Glocker. "Causality matters in medical imaging". In: *Nature Communications* 11.1 (2020), pp. 1–10.

[5] Nianlan Cheng, Shuo Wu, Xianli Luo, Chunyan Xu, Qin Lou, Jin Zhu, Lu You, and Bangguo Li. "A Comparative Study of Chest Computed Tomography Findings: 1030 Cases of Drug-Sensitive Tuberculosis versus 516 Cases of Drug-Resistant Tuberculosis". In: *Infection and Drug Resistance* 14 (2021), pp. 1115–1128.

[6] Yashin Dicente Cid and Henning Müller. "Texture-based Graph Model of the Lungs for Drug Resistance Detection, Tuberculosis Type Classification, and Severity Scoring: Participation in ImageCLEF 2018 Tuberculosis Task." In: *CLEF (Working Notes)*. 2018.

[7] Samantha Flores-Trevino, Eduardo Rodriguez-Noriega, Elvira Garza-Gonzalez, Esteban Gonzalez-Diaz, Sergio Esparza-Ahumada, Rodrigo Escobedo-Sanchez, Hector R. Perez-Gomez, Gerardo Leon-Garnica, and Rayo Morfin-Otero. "Clinical predictors of drug-resistant tuberculosis in Mexico". In: *PLoS One* 14.8 (2019).

[8] Amilcare Gentili. "ImageCLEF2018: Transfer Learning for Deep Learning with CNN for Tuberculosis Classification." In: *CLEF (Working Notes)*. 2018.

[9] *Global tuberculosis report*. Geneva: World Health Organization, 2020.

[10] Miriam Harris, Amy Qi, Luke Jeagal, Nazi Torabi, Dick Menzies, Alexei Korobitsyn, Madhukar Pai, Ruvandhi R Nathavitharana, and Faiz Ahmad Khan. "A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis". In: *PloS one* 14.9 (2019), e0221339.

[11] Xi-Ling Huang, Aliaksandr Skrahin, Pu-Xuan Lu, Sofia Alexandru, Valeriu Crudu, Andrei Astrovko, Alena Skrahina, Jessica Taaffe, Michael Harris, Alyssa Long, Kurt Wollenberg, Eric Engle, Darrell E. Hurt, Irada Akhundova, Sharafat Ismayilov, Elcan Mammadbayov, Hagigat Gadirova, Rafik Abuzarov, Mehriban Seyfaddinova, Zaza Avaliani, Sergo Vashakidze, Natalia Shubladze, Ucha Nanava, Irina Strambu, Dragos Zaharia, Alexandru Muntean, Eugenia Ghita, Miron Bogdan, Roxana Mindru, Victor Spinu, Alexandra Sora, Catalina Ene, Eugene Sergueev, Valery Kirichenko, Vladzimir Lapitski, Eduard Snezhko, Vassili Kovalev, Alexander Tuzikov, Andrei Gabrielian, Alex Rosenthal, Michael Tartakovsky, and Yi Xiang J Wang. "Prediction of multiple drug resistant pulmonary tuberculosis against drug sensitive pulmonary tuberculosis by CT nodular consolidation sign". In: *bioRxiv* (2019).

[12] Aziza Ghanie Icksan, Martin Raja Sonang Napitupulu, Mohamad Arifin Nawas, and Fariz Nurwidya. "Chest X-ray findings comparison between multi-drug-resistant tuberculosis and drug-sensitive tuberculosis". In: *Journal of Natural Science, Biology, and Medicine* 9.1 (2018), p. 42.

[13] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba Garcia Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, et al. "Overview of ImageCLEF 2018: Challenges, datasets and evaluation". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2018, pp. 309–334.

[14] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[15] Adam Ishay and Oge Marques. "ImageCLEF 2018 Tuberculosis Task: Ensemble of 3D CNNs with Multiple Inputs for Tuberculosis Type Classification." In: *CLEF (Working Notes)*. 2018.

[16] Stefan Jaeger, Octavio H Juarez-Espinosa, Sema Candemir, Mahdieh Poostchi, Feng Yang, Lewis Kim, Meng Ding, Les R Folio, Sameer Antani, Andrei Gabrielian, et al. "Detecting drug-resistant tuberculosis in chest radiographs". In: *International journal of computer assisted radiology and surgery* 13.12 (2018), pp. 1915–1925.

[17] Manohar Karki, Karthik Kantipudi, Feng Yang, Hang Yu, Yi Xiang J Wang, Ziv Yaniv, and Stefan Jaeger. "Generalization Challenges in Drug-Resistant Tuberculosis Detection from Chest X-rays". In: *Diagnostics* 12.1 (2022), p. 188.

[18] Manohar Karki, Karthik Kantipudi, Hang Yu, Feng Yang, Yasmin M Kassim, Ziv Yaniv, and Stefan Jaeger. "Identifying Drug-Resistant Tuberculosis in Chest Radiographs: Evaluation of CNN Architectures and Training Strategies". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 2964–2967.

[19] Luyao Ma, Yun Wang, Lin Guo, Yu Zhang, Ping Wang, Xu Pei, Lingjun Qian, Stefan Jaeger, Xiaowen Ke, Xiaoping Yin, and Fleming Lure. "Developing and verifying automatic detection of active pulmonary tuberculosis from multi-slice spiral CT images based on deep learning". In: *Journal of X-ray Science and Technology* 28.5 (2020), pp. 939–951.

[20] Eduardo HP Pooch, Pedro L Ballester, and Rodrigo C Barros. "Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification". In: *arXiv preprint arXiv:1909.01940* (2019).

[21] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Phil Chen, Amirhossein Kiani, Jeremy Irvin, Andrew Y Ng, and Matthew P Lungren. "CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting". In: *arXiv preprint arXiv:2002.11379* (2020).

[22] Alex Rosenthal, Andrei Gabrielian, Eric Engle, Darrell E. Hurt, Sofia Alexandru, Valeriu Crudu, Eugene Sergueev, Valery Kirichenko, Vladzimir Lapitskii, Eduard Snezhko, Vassili Kovalev, Andrei Astrovko, Alena Skrahina, Jessica Taaffe, Michael Harris, Alyssa Long, Kurt Wollenberg, Irada Akhundova, Sharafat Ismayilova, Aliaksandr Skrahin, Elcan Mammadbayov, Hagigat Gadirova, Rafik Abuzarov, Mehriban Seyfaddinova, Zaza Avaliani, Irina Strambu, Dragos Zaharia, Alexandru Muntean, Eugenia Ghita, Miron Bogdan, Roxana Mindru, Victor Spinu, Alexandra Sora, Catalina Ene, Sergo Vashakidze, Natalia Shubladze, Ucha Nanava, Alexander Tuzikov, and Michael Tartakovsky. "The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis". In: *Journal of Clinical Microbiology* 55.11 (2017), pp. 3267–3282.

[23] Seelwan Sathitratanacheewin, Panasun Sunanta, and Krit Pongpirul. "Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability". In: *Heliyon* 6.8 (2020), e04614.

[24] J Ureta and A Shrestha. "Identifying drug-resistant tuberculosis from chest X-ray images using a simple convolutional neural network". In: *Journal of Physics: Conference Series*. Vol. 2071. 1. IOP Publishing. 2021, p. 012001.

[25] Yì Xiàng J Wàng, Myung Jin Chung, Aliaksandr Skrahin, Alex Rosenthal, Andrei Gabrielian, and Michael Tartakovsky. "Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences". In: *Quant Imaging Med Surg* 8.2 (2018), pp. 161–173.

[26] Feng Yang, Hang Yu, Karthik Kantipudi, Manohar Karki, Yasmin M. Kassim, Alex Rosenthal, Darrell E. Hurt, Ziv Yaniv, and Stefan Jaeger. "Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features". In: *Quantitative Imaging in Medicine and Surgery* 12.1 (2022), pp. 675–687.

[27] Feng Yang, Hang Yu, Karthik Kantipudi, Alex Rosenthal, Darrell E Hurt, Ziv Yaniv, and Stefan Jaeger. "Automated Drug-Resistant TB Screening: Importance of Demographic Features and Radiological Findings in Chest X-Ray". In: *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. 2021, pp. 1–4.

[28] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". In: *PLoS Med.* 15.11 (2018).

[29] Qiuting Zheng, Yibo Lu, Fleming Lure, Stefan Jaeger, and Puxuan Lu. "Clinical and radiological features of novel coronavirus pneumonia". In: *Journal of X-ray Science and Technology* 28.3 (2020), pp. 391–404.

Appendix

2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Identifying Drug-Resistant Tuberculosis in Chest Radiographs: Evaluation of CNN Architectures and Training Strategies

Manohar Karki[1], Karthik Kantipudi[2], Hang Yu[1], Feng Yang[1],
Yasmin M. Kassim[1], Ziv Yaniv[2] and Stefan Jaeger[1]

***Abstract*— Tuberculosis (TB) is a serious infectious disease that mainly affects the lungs. Drug resistance to the disease makes it more challenging to control. Early diagnosis of drug resistance can help with decision making resulting in appropriate and successful treatment. Chest X-rays (CXRs) have been pivotal to identifying tuberculosis and are widely available. In this work, we utilize CXRs to distinguish between drug-resistant and drug-sensitive tuberculosis. We incorporate Convolutional Neural Network (CNN) based models to discriminate the two types of TB, and employ standard and deep learning based data augmentation methods to improve the classification. Using labeled data from NIAID TB Portals and additional non-labeled sources, we were able to achieve an Area Under the ROC Curve (AUC) of up to 85% using a pretrained InceptionV3 network.**

I. INTRODUCTION

Tuberculosis (TB) is a global disease caused by the bacterium *Mycobacterium tuberculosis*, which is spread through the air. According to the World Health Organization, in 2019 an estimated 10 million people were infected with TB and about 1.4 million died from the disease [1]. Efforts to control TB have been hindered by the rise of drug-resistant strains, where in 2019 about half a million people developed rifampicin-resistant TB out of which 78% were multidrug-resistant [1]. Early detection of drug resistance enables more specific drug treatment, reduces the period of infectiousness and disease spread in addition to improving outcomes [2].

Current diagnostic methods for identifying drug-resistant TB (DR-TB) infections include conventional culture growth over several weeks and rapid molecular testing [3]. These procedures are not feasible globally, especially for countries unable to scale up their testing capacities. An automated computational approach that utilizes widely available technology is thus desirable. Chest X-rays (CXRs) are extensively used in detection of tuberculosis, and thus offer a potentially natural avenue for discriminating between DR-TB and drug-sensitive TB (DS-TB).

In this work, we evaluate multiple CNN architectures and training strategies with the aim of differentiating between DR-TB and DS-TB. We evaluate both pre-trained CNNs as simple N-layer custom CNNs. In terms of training strategies, we evaluate the use of different data augmentation approaches, augmenting the data statically beforehand or dynamically during training. Along with that, we generate synthesized images for DR-TB and DS-TB from the original images using Generative Adversarial Networks (GANs). We utilize a unique TB dataset provided by the US National Institute of Allergy and Infectious Diseases [4]. This patient based dataset includes clinical, genomic, and radiological data (CXRs and CT), but most importantly, it includes the results of drug susceptibility testing. Finally, we utilize several publicly available TB image datasets with unknown drug susceptibility to further enhance our classifier training.

II. PREVIOUS WORK

Computational identification and classification of lung diseases in medical images has been greatly facilitated by advancements in deep learning [5]. In the context of TB, usage of CXRs to classify an image as TB/not-TB has been described in multiple publications. Even simpler architectures such as AlexNet and GoogleNet, used with around 1000 training images, have shown good performance, exceeding 95% accuracy on some datasets [6]. The specific task of detecting TB in CXRs has seen great success, with multiple commercial products available, and a recent study reporting an area under the receiver operating characteristic curve of 0.92 or greater, when evaluated on unseen data [7].

Very few works have dealt with identifying the type of TB, DR-TB or DS-TB, from images. As part of the ImageCLEF 2017 and 2018 challenges, this question was posed using CT images. In 2017/2018 participants of the challenge were provided with a training set comprised of 230/259 training CTs and 223/236 testing CTs. After running the challenge for two years, the organizers said that "After two editions we concluded that the MDR (Multi-Drug Resistant) subtask was not possible based only on the image." [1] It should be noted that the size of the training dataset was very small, and likely adversely affected deep learning based approaches.

In a different study [8], our group had moderate success in differentiating between DR-TB and DS-TB using CXRs, achieving an AUC of 0.66 utilizing hand-crafted shape and texture features. In clinical research, several publications describe using imaging (CXR or CT) to identify clinical findings that potentially differentiate between DR-TB and DS-TB. In [9], the authors found the DR-TB class to have more large lesions whereas the DS-TB class had more medium and small lesions. In [10], the authors found that the DR-TB class was characterized by having thick-walled

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

[1] https://www.imageclef.org/2019/medical/tuberculosis/

cavities. Finally, in [11], the authors found that presence of multiple cavities was a predictor of DR-TB.

Based on our initial results, and the more recent clinical observations, we believe CXRs can potentially be used for differentiating DR-TB from DS-TB using a deep learning approach, which is described in the following sections.

III. Methods

To discriminate between DR-TB and DS-TB, this work collects and processes CXR images from different sources, selects models trained with deep learning based approaches, and uses training strategies to improve classification performance. The CXRs used in this work are from the following sources: TB Portals [4], Montogomery County and Shenzhen chest X-ray sets [12], and the TBX11K large scale tuberculosis dataset [13]. Table I lists the number of samples for each set. The TB Portals dataset is the only one which contains results of drug susceptibility testing, indicating if the image is DR-TB or DS-TB. For all other datasets, we assume the images are DS-TB as that is significantly more common. To ensure that our evaluation is valid, we only use images from the TB Portals dataset in our testing.

A. Data preprocessing

1) Data Selection: The TB portals dataset contains images from hospitals in 16 countries. Because of this, there are variations in the quality of images. We discarded images that are non-pulmonary, from lateral views, or non-grayscale.

Because an early-stage distinction of drug sensitivity or resistance is desirable, only images from a patient's first visit were selected for this analysis. Further, to give equal weight to all patients, a single image was used per patient even when multiple images were acquired on that visit.

2) Cropping of lung regions from CXRs: CXR images often include significantly sized regions that are outside the lungs, such as shoulders and neck. These regions are not relevant for the classification and in fact can be a hindrance in developing accurate models. Cropping a tight region around the lungs and removing unnecessary regions also allows for a more consistent size of the lungs across multiple images once they are rescaled. We therefore use a deep learning approach to crop the original CXRs to the lung region. During the cropping process, the CXRs are blurred by Gaussian smoothing with a standard deviation of 0.5 to reduce the high frequency signal components. Each smoothed image is resampled to a fixed dimension (256x256), before normalizing the intensities to zero mean and unit standard deviation. The CXRs are passed through a U-Net based segmentation model [14]. The resulting lung masks are used to compute a bounding box to crop the lung region from the original CXRs. The segmentation model was trained on two datasets [12], [15] yielding an IoU of 0.971 and 0.956, respectively. Subsequently, the original images are cropped using the bounding box coordinates, downsampled, and renormalized. Figure 1 illustrates this process.

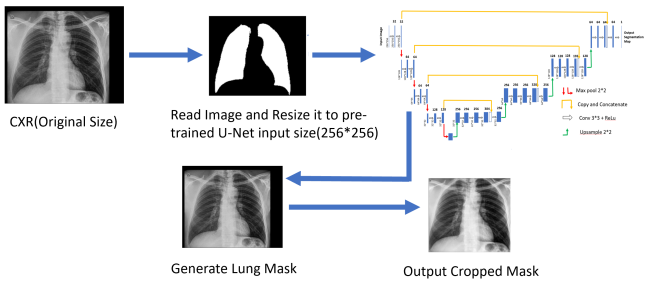


Fig. 1: Preprocessing pipeline for CXRs

B. Network Architectures

For the classification task, we evaluate several standard CNN architectures as well as three custom CNNs. The standard networks include: AlexNet [16], DenseNet [17], InceptionV3 [18], ResNet [19], and Xception [20]. For each of the standard networks, we removed the dense layers after the final convolutional layer and added new dense layers. Table II shows the number of parameters for each of the networks.

C. Data Augmentation

Most deep CNNs require a large amount of good quality data for the models to generalize well. As the number of available samples for each class is relatively small, we use two augmentation approaches:

1) Image transformations: The following transformations are applied to the original images: rotation ($\pm 10^o$), translation (± 5 pixels), blurring ($\mathcal{N}(0.0, 1.0)$), and histogram equalization. We intentionally apply small parameter values for these methods as they replicate the relatively small variations in X-ray images compared to images from other domains. We evaluate the usage of **static augmentation**, one time application of the transformations to the entire dataset before training starts, and **dynamic augmentation**, where original images are modified on the fly during batch training.

2) Synthetic Image Generation: Aside from image transformations, we synthesize images from both categories to increase the number of samples. We use the progressive growing of generative adversarial networks (PG-GANs) [21]. PG-GANs were chosen as they have been shown to generate relatively stable, quality, and variant images. For generating synthetic images for each category during each growth phase of $[4{\times}4, 8{\times}8, \ldots, 128{\times}128]$, batch sizes and epochs

Sources	DR-TB	DS-TB
TB Portals	1821	878
Montgomery County [12]*	0	58
Shenzhen [12]*	0	336
TBX11K [13]*	0	549
Synthetic (using GAN)	1000	1000
Total	2821	2821

TABLE I: Number of images from each source. * Patients from [12] and [13] are assumed to be drug sensitive.

of $[128, 64, 64, 32, 32, 16]$ and $[100, 250, 250, 250, 250]$ were used respectively. The final outputs are up-sampled to the input size of the classifying network.

IV. Experiments

We evaluate model capabilities to distinguish between DR-TB and DS-TB on a patient-level basis. In all experiments, we use 10-fold cross validation.

We start by evaluating multiple models on the TB portals dataset using non-augmented training. We then evaluate the effects of various augmentation strategies on the best models. Finally, we add TB images from external sources, labeling all of them as DS-TB, to the best model from the last step.

A. Model Selection

The pretrained network architectures were designed to address multi-class classification tasks. While we only deal with two classes (DR-TB and DS-TB), the size of the available dataset is much smaller in comparison. We therefore initially evaluate multiple standard architectures and several custom CNNs using the TB Portals dataset.

B. Effects of Augmentation

Once we identify the more promising architectures, we explore the effects of dynamic and static augmentation strategies as well as utilizing synthetically generated images in the training stage. For this experiment, we only select the best performing pretrained-networks (InceptionV3 and Xception) and the best performing custom network. We also evaluate the effect of increasing the amount of statically augmented data on the balanced dataset created in the previous experiment.

C. Including Additional Data

As shown in Table I, the number of DR-TB images in the TB portals dataset is significantly higher than the number of DS-TB images. In an effort to utilize images from all available patients in the imbalanced TB portals dataset, additional TB images from other sources were also included. We label these images as DS-TB, as this is the prevalent type of TB. The previous augmentation strategies were combined with the additional data to see if they influence the overall AUC performance. Note that these images are only used for training purposes as there is no drug susceptibility testing associated with them.

V. Results

In our network architecture comparison, without any augmentation, pretrained InceptionV3 and Xception networks had the best performance, as can be seen in Table II. These two networks, and several custom CNNs (6-layer, 10-layer, 12-layer), were also trained with random initialization. Among the custom networks, the 6-layer CNN had the best area under the ROC curve (AUC) with $0.74 \pm .04$ compared to the rest of the custom networks. When randomly initialized, the performance of InceptionV3 and Xception deteriorated.

Different augmentation methods and addition of synthetic images did not yield better performance for these networks, as shown in Table III. Performance did not scale with the increase in augmented data. When number of samples was increased to 3X and 4X original samples size by static augmentation, performance decreased. Interestingly, the performance of the custom 6-layer network improved with the same training strategy. We chose to continue our evaluation using the InceptionV3 network as its performance remained the most consistent with these augmentation strategies.

Architecture	**Parameters (in millions)**	**AUC**
Pretrained Networks		
AlexNet [16]	5.7	0.79 (± .02)
DenseNet121 [17]	7.2	0.79 (± .02)
DenseNet201 [17]	18.6	0.80 (± .02)
InceptionV3 [18]	22.3	**0.81 (± .03)**
InceptionResNetV2 [18]	54.7	0.77 (± .05)
ResNet50 [19]	24.1	0.80 (± .03)
ResNet152 [19]	58.7	0.77 (± .03)
Xception [20]	21.3	**0.81 (± .02)**
Random initialization		
6-layer CNN	3.0	**0.74 (± .04)**
10-layer CNN	8.6	0.70 (± .03)
12-layer CNN	9.1	0.65 (± .04)
InceptionV3 [18]	22.3	0.76 (± .03)
Xception [20]	21.3	0.76 (± .03)

TABLE II: Mean AUC (Area Under ROC Curve) of 10-fold cross validation results when various pretrained and custom networks are tested on TB Portals dataset

Figure 2 summarizes the performance evaluation of InceptionV3, using various datasets and augmentation strategies. We see that static augmentation has an overall positive effect on model performance compared to the dynamic augmentation strategy. We also see that the addition of images from other sources to the training set combined with static augmentation lead to the best AUC overall performance of 85%.

Although the inclusion of *both* the *synthetic* data and data from *additional sources* improves performance when using dynamic augmentation, it did not have an effect when using static augmentation. The variance in performance is slightly better when *both* synthetic and additional data are included.

Finally, to inspire confidence in the predictions of our network, we utilize GradCAM heatmaps [22] to visualize its focus. Figure 3 shows two heatmaps for correctly predicted .

Network Architecture	**Dynamic**	**Static**			**Synthetic**
		2X	*3X*	*4X*	
InceptionV3 (pretrained)	0.80 (±.03)	**0.81** (±.03)	0.80 (±.02)	0.79 (±.02)	**0.81** (±.02)
Xception (pretrained)	0.80 (±.03)	0.80 (±.03)	0.77 (±.04)	0.79 (±.03)	**0.81** (±.03)
6-layer CNN	0.76 (±.03)	0.76 (±.02)	0.74 (±.02)	0.76 (±.03)	0.75 (±.03)

TABLE III: AUC with *dynamic* and *static* augmentation and with augmentation using GAN generated images.

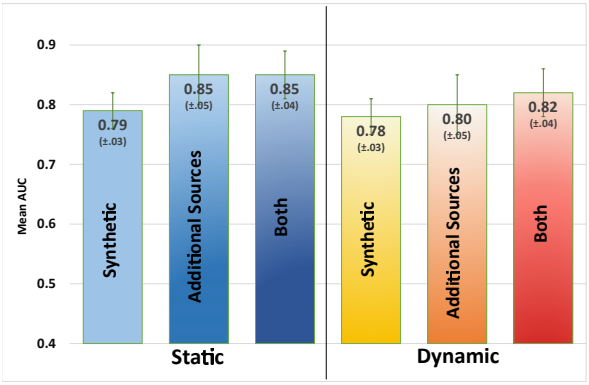


Fig. 2: Mean AUC performances of the InceptionV3 network with *static* or *dynamic* augmentation and including a) *synthetic* images, b) images from [12] and [13] (referred in figure as *additional sources*) c) *both*. These additional images with static augmentation provided the best performance.

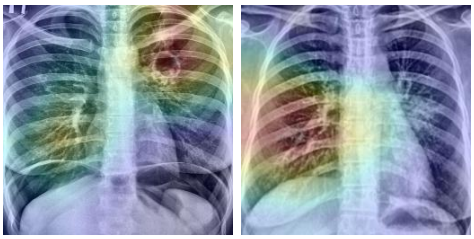


Fig. 3: GradCAM heatmaps superimposed on the original images, Classified as DR-TB (left image) and DS-TB (right image) due to likelihood values of .99 and .05 respectively.

VI. CONCLUSIONS

This paper presents an evaluation of models for discriminating between drug-resistant and drug-sensitive TB in the TB portals dataset, using augmentation strategies and other publicly available data. With a 10-fold cross validation, we achieve the best AUC performance of 85%. Even without augmentation and additional data, but with pretrained weights, we achieve a 81% AUC performance with InceptionV3 and Xception networks. GradCAM heatmaps affirm that the models learn from relevant areas from the CXRs during the training process. Despite discouraging earlier work in the literature, our work has shown that discriminating between DR-TB and DS-TB can be possible in CXRs for a sufficiently large training set.

ACKNOWLEDGMENT

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

REFERENCES

[1] World Health Organization, *Global tuberculosis report*, 2020.
[2] P. O'Riordan, U. Schwab *et al.*, "Rapid molecular detection of rifampicin resistance facilitates early diagnosis and treatment of multidrug resistant tuberculosis: case control study," *PLoS One*, vol. 3, no. 9, p. e3173, 2008.
[3] C. Lange, K. Dheda *et al.*, "Management of drug-resistant tuberculosis," *The Lancet*, vol. 394, no. 10202, pp. 953–966, 2019.
[4] A. Rosenthal, A. Gabrielian *et al.*, "The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis," *Journal of Clinical Microbiology*, vol. 55, no. 11, pp. 3267–3282, 2017.
[5] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A survey of deep learning for lung disease detection on medical images: state of the art, taxonomy, issues and future directions," *Journal of Imaging*, vol. 6, no. 12, 2020.
[6] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.
[7] Z. Z. Qin, M. S. Sander *et al.*, "Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems," *Sci Rep.*, vol. 9, no. 1, 2019.
[8] S. Jaeger, O. H. Juarez-Espinosa *et al.*, "Detecting drug-resistant tuberculosis in chest radiographs," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1915–1925, 2018.
[9] 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," *Journal of Natural Science, Biology, and Medicine*, vol. 9, no. 1, p. 42, 2018.
[10] Y. X. J. Wang, M. J. Chung *et al.*, "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.
[11] S. Flores-Trevino, E. Rodriguez-Noriega *et al.*, "Clinical predictors of drug-resistant tuberculosis in Mexico," *PLoS One*, vol. 14, no. 8, 2019.
[12] S. Jaeger, S. Candemir *et al.*, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, p. 475, 2014.
[13] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
[14] P. Ivan, P. Vitali, and B. Uladzislau, "Lung segmentation (2D)," https://github.com/imlab-uiip/lung-segmentation-2d, 2020.
[15] J. Shiraishi, S. Katsuragawa *et al.*, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *AJR. American Journal of Roentgenology*, vol. 174, pp. 71–4, 02 2000.
[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 2017.
[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 2016.
[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
[22] R. R. Selvaraju, M. Cogswell *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

Original Article

Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features

Feng Yang[1]^, Hang Yu[1], Karthik Kantipudi[2], Manohar Karki[1], Yasmin M. Kassim[1], Alex Rosenthal[2], Darrell E. Hurt[2], Ziv Yaniv[2], Stefan Jaeger[1]

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

Contributions: (I) Conception and design: F Yang, Z Yaniv, S Jaeger; (II) Administrative support: A Rosenthal, DE Hurt, S Jaeger; (III) Provision of study materials or patients: K Kantipudi, Z Yaniv, A Rosenthal, DE Hurt; (IV) Collection and assembly of data: K Kantipudi, Z Yaniv, A Rosenthal, DE Hurt; (V) Data analysis and interpretation: F Yang, H Yu, K Kantipudi, M Karki, YM Kassim; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Dr. Feng Yang; Dr. Stefan Jaeger. Lister Hill National Center of Biomedical Communications, National Library of Medicine, NIH, Bethesda, MD 20894, USA. Email: feng.yang2@nih.gov; stefan.jaeger@nih.gov.

Background: Tuberculosis (TB) drug resistance is a worldwide public health problem that threatens progress made in TB care and control. Early detection of drug resistance is important for disease control, with discrimination between drug-resistant TB (DR-TB) and drug-sensitive TB (DS-TB) still being an open problem. The objective of this work is to investigate the relevance of readily available clinical data and data derived from chest X-rays (CXRs) in DR-TB prediction and to investigate the possibility of applying machine learning techniques to selected clinical and radiological features for discrimination between DR-TB and DS-TB. We hypothesize that the number of sextants affected by abnormalities such as nodule, cavity, collapse and infiltrate may serve as a radiological feature for DR-TB identification, and that both clinical and radiological features are important factors for machine classification of DR-TB and DS-TB.

Methods: We use data from the NIAID TB Portals program (https://tbportals.niaid.nih.gov), 1,455 DR-TB cases and 782 DS-TB cases from 11 countries. We first select three clinical features and 26 radiological features from the dataset. Then, we perform Pearson's chi-squared test to analyze the significance of the selected clinical and radiological features. Finally, we train machine classifiers based on different features and evaluate their ability to differentiate between DR-TB and DS-TB.

Results: Pearson's chi-squared test shows that two clinical features and 23 radiological features are statistically significant regarding DR-TB *vs.* DS-TB. A ten-fold cross-validation using a support vector machine shows that automatic discrimination between DR-TB and DS-TB achieves an average accuracy of 72.34% and an average AUC value of 78.42%, when combing all 25 statistically significant features.

Conclusions: Our study suggests that the number of affected lung sextants can be used for predicting DR-TB, and that automatic discrimination between DR-TB and DS-TB is possible, with a combination of clinical features and radiological features providing the best performance.

Keywords: Differential diagnosis; tuberculosis (TB); drug-resistance (DR); clinical features; radiological features; machine learning

^ ORCID: 0000-0002-8334-7450.

Submitted Mar 16, 2021. Accepted for publication Jul 23, 2021.
doi: 10.21037/qims-21-290
View this article at: https://dx.doi.org/10.21037/qims-21-290

Introduction

Tuberculosis (TB) drug resistance is a global public health concern since it threatens the progress made in TB care and control (1). In 2019, there were an estimated 10 million new TB cases; approximately half a million cases are resistant to rifampicin, of which 78% are multidrug-resistant TB (MDR-TB) (2). MDR-TB is a type of TB that is resistant to at least two first-line anti-TB drugs: isoniazid and rifampicin. Drug-resistant TB is a growing public health concern since it requires more complex treatment than drug-sensitive TB and incurs more costs. In 2019, it was estimated that globally 3.3% of new TB cases and 17.7% of previously treated TB cases are MDR-TB (2).

Early identification of drug resistance enables patient-specific drug treatment, which reduces the period of infectiousness and disease spread in addition to improving outcomes. However, discrimination between drug-resistant TB (DR-TB) and drug-sensitive cases (DS-TB) using readily available clinical information and images, preferably during the first visit, is still an open problem. Currently, there are two types of TB drug susceptibility tests: conventional culture-based phenotypic testing and molecular testing. The former involves looking at the bacteria behavior, which requires a well-equipped laboratory facility and may take several weeks to obtain the results (3). The latter involves looking at genetic mutations, which is fast but expensive and may produce inconclusive results (4). Therefore, it is desirable to predict the suspicion of DR-TB automatically from radiological findings and clinical information in patient medical records.

Inspired by the work in (5), we hypothesize that the number of affected sextants may serve as important radiological features for DR-TB identification, and that both clinical and radiological features are important factors for machine classification between DR-TB and DS-TB.

Previous work

There is evidence that certain clinical data and radiological findings may enable differentiation between DR-TB and DS-TB. Faustini *et al.* (6) conducted a review of twenty-nine studies before 2006 and reported that prior treatment for TB is the strongest determinant of MDR-TB in Europe. Tembo and Malangu (7) collected 2,568 medical records in Botswana and found that previous treatment and positive sputum smear microscopy are associated with the prevalence of MDR-TB or rifampicin-resistant TB (RR-TB). Mdivani *et al.* (8) reported three risk factors for MDR-TB based on analysis of 1,422 patients from Georgia: retreatment case, history of injection drug use and female gender. O'Donnell *et al.* (9) found that women admitted to hospital in KwaZulu-Natal, South Africa with drug-resistant TB are 38% more likely than men to have XDR-TB based on an analysis of 4,514 patients. Shen *et al.* (10) conducted an analysis of 8419 patients from Shanghai, China, and stated that patients aged 30–59 years are more likely to be DR-TB in previously treated cases. Lv *et al.* (11) reported from 3,552 patients in Dalian, China, that previously treated patients and older age are more likely to have MDR-TB. Icksan *et al.* (12) compared chest X-ray (CXR) findings from 183 MDR-TB and 183 DS-TB cases in Indonesia and reported that the MDR-TB group has more large-size lesions while the DS-TB group has more small- and medium-size lesions. Wáng *et al.* (13) performed a review on available articles before 2018 for radiological signs of MDR-TB and found that thick-walled multiple cavities (particularly with count ≥3 and size ≥30 mm) present the most promising radiological sign for MDR-TB with good specificity but at the cost of low sensitivity. Huang *et al.* (5) reported from 468 DR-TB cases and 223 DS-TB cases that a combination of consolidated nodule number and size can be used to predict the probability of MDR-TB. Flores-Treviño *et al.* (14) found from 144 patients in Mexico that multiple cavities is a predictor for DR-TB.

Based on these past studies, we hypothesized that the number of affected sextants may serve as a useful feature when applying machine learning techniques to the discrimination of DR-TB and DS-TB. To date, very few works have been concerned with discriminating between DR-TB and DS-TB in an automated manner. Kovalev *et al.* (15) apply different machine learning methods to features extracted from CXR images or CT images or both. They achieve an accuracy of 73% with an AUC value of 72%, a sensitivity of 82% and a specificity of 58% when combining the features from both X-ray and CT images. The accuracy

Table 1 Origin of the TB cases analyzed in this work

Country	No. (%) of cases		
	DS-TB	DR-TB	Total
Azerbaijan	0 (0)	7 (0.31)	7 (0.31)
Belarus	113 (5.05)	461 (20.61)	574 (25.66)
Georgia	364 (16.27)	340 (15.20)	704 (31.47)
India	151 (6.75)	49 (2.19)	200 (8.94)
Kazakhstan	35 (1.56)	223 (9.97)	258 (11.53)
Kyrgyzstan	0 (0)	110 (4.92)	110 (4.92)
Moldova	1 (0.04)	26 (1.16)	27 (1.21)
Republic of the Congo	0 (0)	1 (0.04)	1 (0.04)
Romania	4 (0.18)	87 (3.89)	91 (4.07)
South Africa	94 (4.20)	3 (0.13)	97 (4.34)
Ukraine	20 (0.89)	148 (6.62)	168 (7.51)

for CXR images alone or CT images alone falls to 62% and 65%, respectively. Our previous work (16) applied both traditional machine-learning methods and deep learning networks to CXR images, achieving an AUC value around 66% and an accuracy around 60%.

Methods

Data collection

We use a dataset of 2,237 patients, which includes de-identified clinical data and CXR images publicly available from the NIAID TB Portals program (17). Each patient record is manually annotated with clinical information and radiological findings based on CXR images. Clinical information includes age of onset, gender, patient type (new, relapse or failure), type of sample (pulmonary or extrapulmonary), BMI, diagnosis, 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 (17). Radiological findings include chest radiography patterns such as the number and location of affected sextants, 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. Due to the large number of radiologists participating in this study, their annotations are not biased toward a single radiologist. The 2,237 patients include 782 DS-TB and 1,455 DR-TB patients, acquired from 11 countries. Distribution of the origin is listed in *Table 1*. The type of drug susceptibility was determined by sputum cultured drug resistance testing and/or molecular testing, specifically Bactec, Hain (FL-LPA), SL-LPA, GeneXpert, and Lowenstein -Jensen testing (17,18). Data usage is exempt from local institutional review board review as it is publicly available from the TB portals program. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The TB portals program participants are responsible for ensuring compliance with their countries' laws, regulations, and ethics considerations (17).

Statistical analysis

In this work, we hypothesize that the number of affected sextants may serve as an important radiological feature for DR-TB identification, and that both clinical and radiological features are relevant for machine classification of DR-TB and DS-TB. Lung sextants are defined by dividing each lung into three equal sections from apex to base, as shown in *Figure 1*. A sextant is said to be affected if either nodules, cavities, collapses, or infiltrates are present. We utilize three clinical features and 26 radiological features, as listed in *Table 2*. Age in this paper means age of onset. To gain insight into the statistical significance of extracted features with different types of resistance (DR-TB or DS-TB), Pearson's chi-squared test is applied. A feature with $P<0.05$ is considered statistically significant.

Machine classification of drug-sensitive and drug-resistant TB

Based on the clinical and radiological features selected by Chi-squared tests, we train a machine learning classifier,

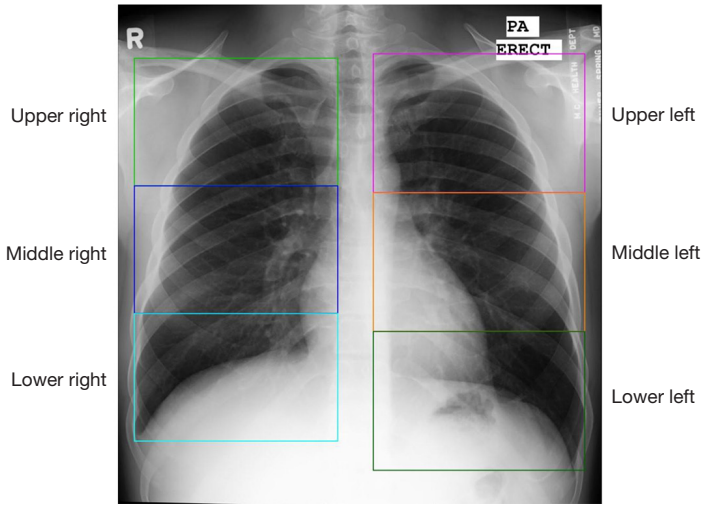


Figure 1 Definition of lung sextants in our study.

Table 2 Summary of clinical features and radiological features extracted from medical records

Type and number of features	Description
Three clinical features	Age; gender; patient type (new case, relapse, or failure)
Radiological features	
12 nodule features	Number of sextants affected by (I) either small nodules (<3 mm), (II) medium nodules (3–8 mm), (III) large nodules (8–30 mm), (IV) huge nodules (≥30 mm), (V) multiple nodules, (VI) calcified or partially calcified nodules, (VII) non-calcified nodules, (VIII) clustered nodules, (IX) low ground glass density active fresh nodules, (X) medium density stabilized fibrotic nodules, (XI) high density calcified typically sequellae nodules, or (XII) any kind of nodules
Six cavity features	Number of sextants affected by either (I) small cavities (<10 mm), (II) medium cavities (10–25 mm), (III) large cavities (>25 mm), (IV) multi-sextant cavities, (V) visible multiple cavities; or (VI) any kind of cavities
Eight other lung abnormality features	(I) Overall percentage of abnormal volume; (II) pleural effusion percentage of involved hemithorax; (III) number of sextants affected by collapse; (IV) number of sextants affected by low ground glass density infiltrates; (V) number of sextants affected by medium density infiltrates; (VI) number of sextants affected by high density infiltrates; (VII) presence of mediastinal lymphadenopathy; (VIII) presence of other non-TB abnormalities

Age in our work means age of onset.

a Support Vector Machine (SVM) (19), to discriminate between DS-TB and DR-TB. We illustrate the pipeline of our machine classification in *Figure 2*. To compare the contributions of different features for classifying DR-TB *vs.* DS-TB, we train the SVM classifier using different feature combinations.

Our dataset includes 782 DS-TB cases and 1,455 DR-TB cases, and is thus biased toward DR-TB. Machine learning classifiers are sensitive to the proportions of different classes in the training set. If ignored, data imbalance will bias predictions in favor of the majority class, leading to inaccurate results. To balance the dataset, several approaches can be applied: down-sampling of the majority class, over-sampling of the minority class, or synthetic minority over-sampling (SMOTE) type techniques (20-23). Data down-sampling involves randomly removing samples from the majority class, which might discard useful information. Data over-sampling is a process of randomly duplicating samples of the minority class, which will not lose any information from the original dataset, but is prone to over-fitting to the training data. The SMOTE technique is an improved over-sampling method that synthesizes new samples from

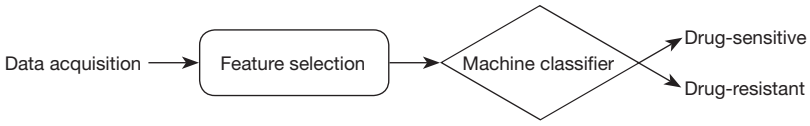


Figure 2 Workflow of the proposed machine classification between DR-TB and DS-TB. DS-TB, drug-sensitive tuberculosis; DR-TB, drug-resistant tuberculosis.

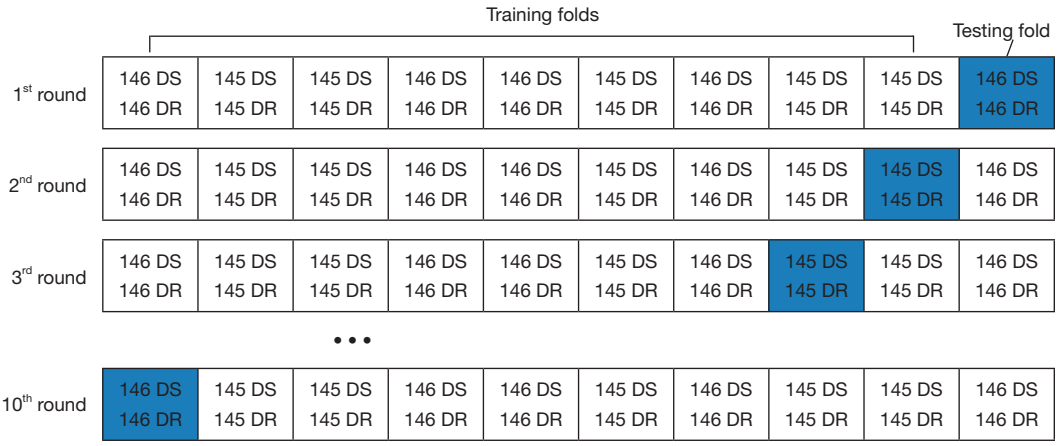


Figure 3 Ten-fold cross validation based on balanced data with oversampled DS-TB cases. DS-TB, drug-sensitive tuberculosis.

the minority class. It often outperforms simple/random over-sampling (21-23). In our study, we use the classic SMOTE method (20) for data over-sampling in imbalanced binary dataset classification. We evaluate the classification performance based on ten-fold cross validation: 782 DS-TB cases and 1,455 DR-TB cases are first divided into ten folds, each of which includes 78 (or 79) DS-TB cases and 145 (or 146) DS-TB cases. To balance the data between DS-TB and DR-TB cases, the 78 (or 79) DS-TB cases in each fold are oversampled to 145 (or 146) cases using the SMOTE method, or the 145 (or 146) DR-TB cases in each fold are down-sampled to 78 (or 79). Then, nine folds are used for training data and the remaining fold is used as testing data to calculate the accuracy. The process is repeated ten times, with each fold serving once as testing data, and the average performance of the ten rounds is reported as the final evaluation result. *Figure 3* visually illustrates the ten-fold cross validation scheme on balanced data obtained via the SMOTE oversampling method.

Results

Figure 4 illustrates the distribution of the three clinical features we use (patient type, age and gender) for the 1,455 DR-TB and 782 DS-TB cases analyzed in this work. We find visible differences for patient type distributions between DR-TB and DS-TB: For DR-TB cases, 62.96% are in the *New* category, 23.71% are in the *Relapse* category, and 13.33% are in the *Failure* category. However, for DS-TB cases, 90.41% are in the *New* category, 9.34% are in the *Relapse* category, and only 0.26% are in the *Failure* category. Age categories follow those defined in (24), with a slightly finer resolution. Clear differences are observed among age groups between DR-TB and DS-TB. Both DS-TB and DR-TB patients are more likely to be less than 65 years old, whereas the frequency of DR-TB is higher in the age groups of 35–44 and 45–54. No clear difference is observed for the gender distribution between DR-TB and DS-TB. Differences observed in feature distributions are consistent with the statistical significance analysis using Pearson's chi-squared test, which is performed based on the null hypothesis that the clinical feature categories are independent from being drug-sensitive or drug-resistant. Patient type ($P<0.001$) and age ($P<0.01$) show a statistically significant association with resistance type. More details about the distribution and chi-squared test results of the three clinical features can be found in Table S1.

Figure 5 shows the distribution of 12 nodule features.

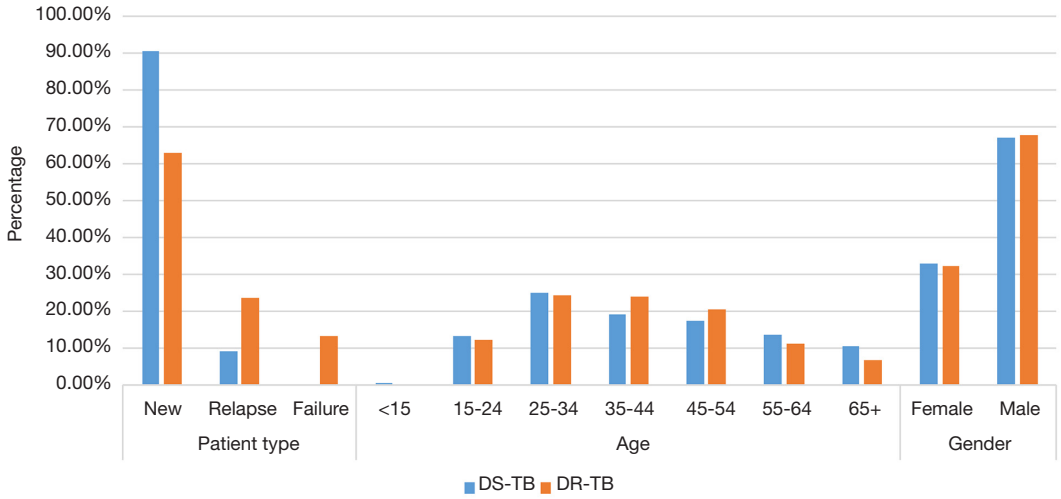


Figure 4 Distributions of three clinical features for 782 DS-TB patients (blue) and 1,455 DR-TB patients (orange). Shown are the percentage of cases for each feature present in a given category (e.g., the blue bars for patient type show that 90.41% of DS-TB patients are in the *New* category, 9.34% in the *Relapse* category, and 0.26% in the *Failure* category). DS-TB, drug-sensitive tuberculosis; DR-TB, drug-resistant tuberculosis.

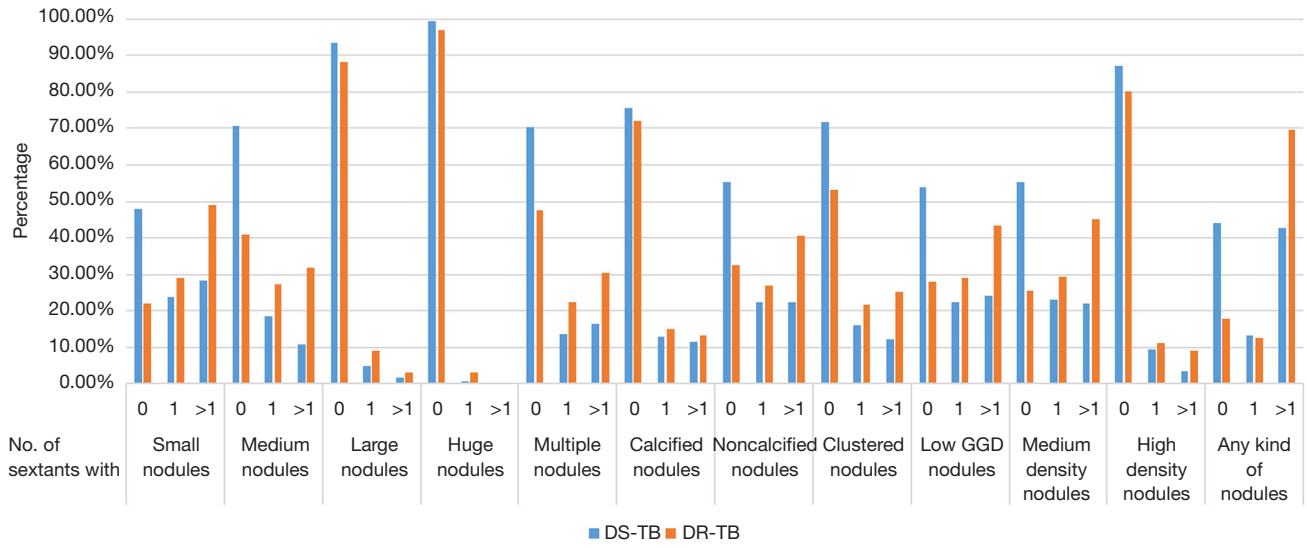


Figure 5 Distributions of 12 nodule related features for 782 DS-TB patients (blue) and 1,455 DR-TB patients (orange). Shown are the percentages of cases for each feature present in a given category (e.g., the blue bars in the category *Small nodules* show that 48.08% of DS-TB cases have no sextant affected by small nodules, 23.79% have only one sextant affected by small nodules, and 28.13% have more than one sextant affected by small nodules). DS-TB, drug-sensitive tuberculosis; DR-TB, drug-resistant tuberculosis.

Visible differences are observed between DS-TB and DR-TB. Nodules occur in approximately 82% of DR-TB patients and in approximately 56% of DS-TB patients. Multiple small or medium nodules, large nodules, huge nodules, multiple non-calcified nodules, multiple clustered nodules, multiple low ground-glass-density (GGD) nodules,

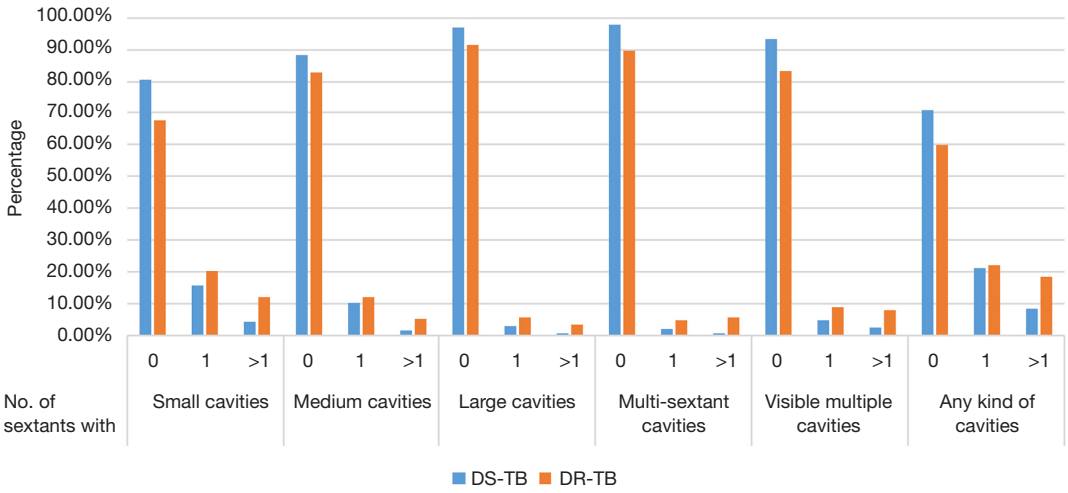


Figure 6 Distributions of six cavity-related features for 782 DS-TB patients (blue) and 1,455 DR-TB patients (orange). Shown are the percentages of cases for each feature present in a given category (e.g., the blue bars in the category *Small cavities* show that 80.31% of DS-TB cases have no sextant affected by small cavities, 15.60% have only one sextant affected by small cavities, and 4.09% have more than one sextant affected by small cavities). DS-TB, drug-sensitive tuberculosis; DR-TB, drug-resistant tuberculosis.

multiple medium-density nodules and multiple high-density nodules are more likely to occur in DR-TB. Pearson's chi-squared tests on the 12 nodule features show that all the nodule features are statistically significant regarding DR-TB *vs.* DS-TB, except for the number of sextants with calcified nodules (P>0.05). *Figure 6* depicts the distribution of six cavity features. We find that cavities occur in approximately 40% of DR-TB patients and in approximately 29% of DS-TB patients. Multiple small cavities, multiple medium cavities, and large cavities tend to occur in DR-TB patients. Pearson's chi-squared tests indicate that all six cavity features are statistically significant regarding DR-TB *vs.* DS-TB. *Figure* 7 shows the distribution of eight other lung abnormality features. We observed that multiple sextants with collapses, multiple sextants with low GGD infiltrates, high-density infiltrates, more than 50% abnormal volume, and mediastinal lymphadenopathy are more likely to occur in DR-TB. Chi-squared test results show that, the number of sextants affected by medium density infiltrates (P>0.05) and the presence of other non-TB abnormalities (P>0.05), are not statistically significant with respect to discriminating between DR-TB and DS-TB. Differences visible in the feature distributions in *Figures 4*-7 are consistent with Pearson's chi-squared tests. More details about distributions and chi-squared test results of the 26 radiological features can be found in Tables S2,S3.

To investigate the possibility of automatically differentiating between DR-TB and DS-TB and to evaluate the importance of specific features, we train SVM machine classifiers using eight different combinations of features listed in *Table 2*: (I) three clinical features; (II) 12 nodule features; (III) six cavity features; (IV) 12 nodule features plus six cavity features; (V) three clinical features plus 12 nodule features plus six cavity features; (VI) all 26 radiological features; and (VII) 25 significant clinical and radiological features. *Tables 3*,*4* show the results for machine classification of DR-TB cases *vs.* DS-TB cases using data down-sampling (*Table 3*) and data augmentation (*Table 4*) for data balancing, respectively. Comparing the results in *Tables 3*,*4*, we find that (I) classifiers with data augmentation achieve better results; (II) classifiers using only clinical features obtain an accuracy around 61%, with very low sensitivity; (III) compared to classifiers using only clinical features, classifiers using radiological features achieve a higher sensitivity at the cost of much lower specificity; (IV) classifiers using a combination of clinical and radiological features achieve a better performance than those using any of them alone; (V) the classifier using the 25 statistically significant features with SMOTE data augmentation achieves the best performance, with an average accuracy

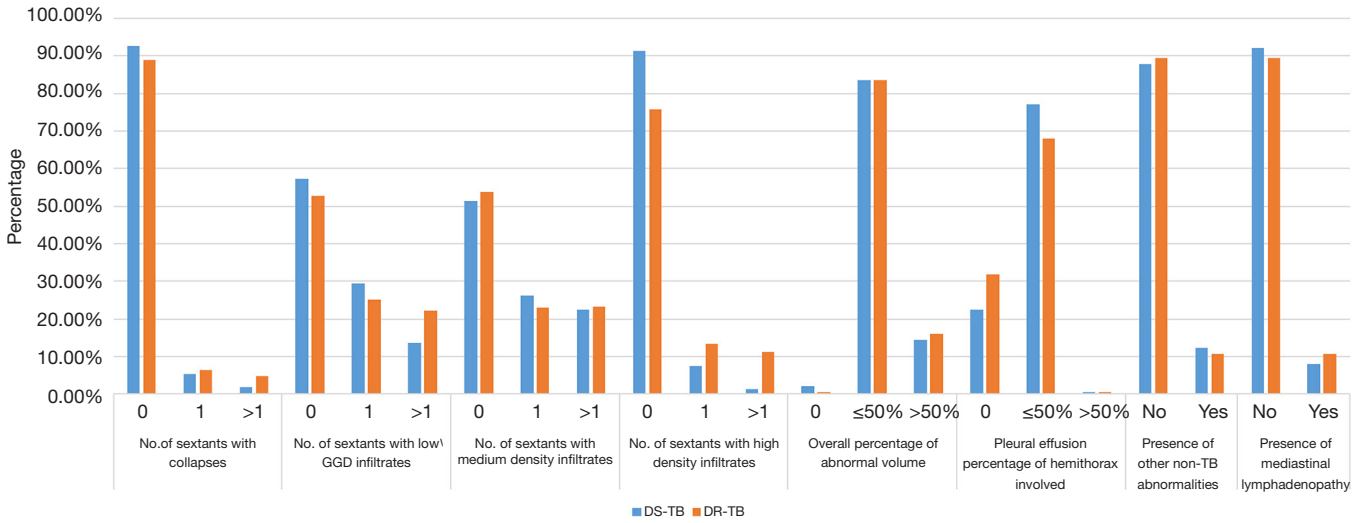


Figure 7 Distributions of eight other lung abnormality features for 782 DS-TB patients (blue) and 1,455 DR-TB patients (orange). Shown are the percentages of cases for each feature present in a given category (e.g., the blue bars in the category *No. of sextants with collapses* show that 92.71% of DS-TB cases have no sextant affected by collapse, 5.37% have only one sextant affected by collapse, and 1.92% have more than one sextant affected by collapse). DS-TB, drug-sensitive tuberculosis; DR-TB, drug-resistant tuberculosis.

Table 3 Ten-fold cross validation on balanced dataset with data down-sampling between DS-TB (782 cases) and DR-TB (782 cases)

Training features	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC (%)
3 clinical features	60.22±3.05	40.04±7.54	84.40±4.84	72.33±5.57	60.77±4.95
12 nodule features	64.38±4.26	71.99±6.81	56.77±4.83	62.46±3.65	67.18±4.58
6 cavity features	55.62±3.08	30.55±5.70	80.69±5.60	61.69±7.51	55.03±3.29
12 nodule + 6 cavity features	64.32±4.82	71.99±6.66	56.64±5.98	62.43±4.25	67.38±4.96
3 clinical + 12 nodule + 6 cavity features	66.63±4.53	66.12±6.02	67.14±5.35	66.85±4.45	70.94±4.68
26 radiological features	66.44±4.87	*76.47±6.12*	56.40±5.50	63.70±3.97	71.13±4.68
3 clinical + 26 radiological features	68.22±4.77	73.01±5.06	*63.43±6.85*	*66.78±4.61*	*74.05±5.84*
25 significant features	*68.29±3.40*	75.70±5.49	60.87±5.91	66.04±3.38	73.60±4.79

Twenty-six radiological features include 12 nodule features, six cavity features and eight other lung abnormality features. Twenty-five significant features are obtained by excluding the four non-significant features from the 29 features. The best performance in each column is marked in italic.

of 72.34% and an average AUC value of 78.42%. The 25 significant features are obtained by excluding the four non-significant features (i.e., gender, number of sextans with calcified nodules, number of sextants with medium density infiltrates, and presence of other non-TB abnormalities) from the 29 features (three clinical features and 26 radiological features). The ROC curve for SVM-based machine classification using the 25 significant features is shown in *Figure 8*. More details of our trained SVM model can be found in Table S4.

Discussion

Limitation of the study

The current study has some limitations. First, we did not

Table 4 Ten-fold cross validation on balanced dataset with SMOTE data augmentation between DS-TB (1,455 cases) and DR-TB (1,455 cases)

Training features	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC (%)
3 clinical features	61.89±3.31	38.43±5.27	85.35±7.11	73.63±9.26	64.44±3.07
12 nodule features	64.94±3.56	72.71±3.20	57.17±4.96	63.00±3.39	68.79±3.64
6 cavity features	56.77±2.43	34.02±3.34	79.53±5.64	63.02±6.29	56.60±3.76
12 nodule + 6 cavity features	65.33±4.03	70.38±3.64	60.28±6.42	64.08±4.16	68.76±4.42
3 clinical + 12 nodule + 6 cavity features	67.28±3.30	70.44±3.12	64.13±6.77	66.47±3.83	73.14±2.98
26 radiological features	67.28±5.35	84.13±2.81	56.95±13.34	67.40±7.46	72.25±6.24
3 clinical + 26 radiological features	70.99±3.18	74.36±2.84	67.63±7.36	69.98±4.42	77.56±3.72
25 significant features	*72.34±2.65*	*75.33±3.36*	*69.35±6.29*	*71.30±3.63*	*78.42±2.63*

Twenty-six radiological features include 12 nodule features, six cavity features and eight other lung abnormality features. Twenty-five significant features are obtained by excluding the four non-significant features from the 29 features. The best performance in each column is marked in italic.

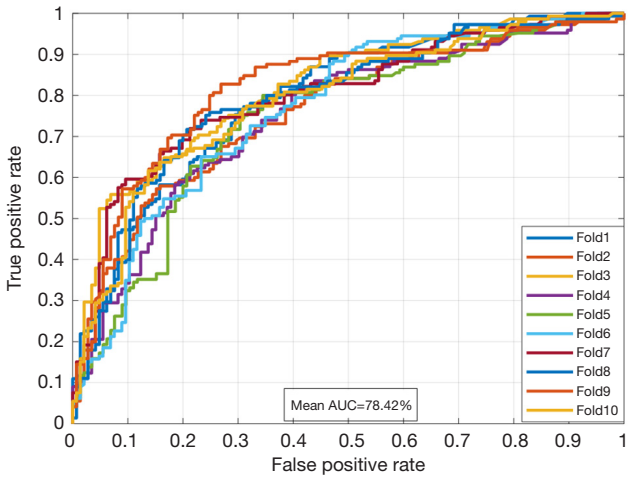


Figure 8 ROC curve of SVM classifier based on 25 significant features with data over-sampling and 10-fold cross validation. ROC, receiver operating characteristic; SVM, Support Vector Machine; AUC, area under the curve.

use radiological features from chest CT images, since many of the existing records do not include annotated CT radiological findings. Second, distribution bias is present in the source of both DR-TB and DS-TB. This work includes 2,237 patients from 11 countries. However, about 97% of all drug-sensitive cases are from five countries (Georgia, Belarus, India, South Africa, and Kazakhstan) and about 81% of all drug-resistant cases are from four countries (Belarus, Georgia, Kazakhstan and Ukraine). That is, our machine classifier learns drug-sensitive and drug-resistant features primarily from six countries, and the classification performance will decrease when we use it to identify DR-TB from other countries or when we perform a country-level evaluation. Third, DR-TB and DS-TB cases are imbalanced. We used SMOTE data augmentation techniques to balance the dataset; it would be better to have a truly balanced dataset for both the statistical significance analysis and machine classifier training. We are now working to increase the sample size of DR-TB and DS-TB cases to obtain a balanced data set with more uniformly distributed countries of origin. Fourth, our best machine learning model based on 25 significant features requires radiological findings reported by a radiologist, which limits the full automation of our machine learning model. Our next step will be to automatically detect such radiological features from radiographs based on deep learning methods.

Roles of patient type, age, and gender

In our study, the strongest predictor of DR-TB is patient type (New, Relapse or Failure). This is consistent with previous studies in Europe, Botswana, and Georgia (6-8). The treatment history is a well-known risk factor for the development of DR-TB. The WHO Global Report on Surveillance on MDR-TB and XDR-TB (25) stated that TB cases with a history of previous TB treatment are

significantly associated with DR-TB. Such a predictor can be used in early screening of DR-TB cases, especially in resource-limited clinical settings. For example, a relapsed patient would indicate that drug susceptibility testing is recommended at treatment start.

The association between DR-TB and age is not well established in the previous reports since different studies use different cut-off points for age groups. The European MDR-TB analysis (6) indicated that MDR-TB patients are more likely to be younger than 65 years, while the report on MDR-TB in Shanghai, China (10) stated that the age group of 30–59 years is associated with MDR-TB. In our data, age also showed a significant association with DR-TB. However, we found both that DR-TB and DS-TB are more likely to happen in age groups less than 65 years, and that the frequency of DR-TB is higher than DS-TB in the age groups of 35–44 and 45–54.

Tuberculosis is more common in males (26-28). In (6) it was reported that male gender is a risk factor for MDR-TB cases in Western Europe, but not in Eastern Europe. Contrary to that observation, it was found in Georgia, that female gender is significantly associated with MDR-TB (8,9). The authors in (8) assumed that this association is related to the fact that the majority of health care workers are females in Georgia. In our study, we did not find a significant association between gender and DR-TB. We hypothesize that gender is likely a regional risk factor for DR-TB, but does not present a general association in the TB data from 11 countries used in this work.

Roles of radiological features

Previous works (29-31) have revealed that active TB is likely to affect the upper lung regions exhibiting cavities, consolidations, and nodules, and to affect unilateral lung regions exhibiting pleural effusions. However, based on a review paper from 2018 (13) and our literature search in PubMed on Feb 4, 2021, only a small number of reports have mentioned the importance of lesion types and their locations in the development of DR-TB. Both the systematic review of radiological signs for MDR-TB before 2018 (13) and the report on MDR-TB in Mexico (14) found that multiple cavities is a promising sign for identifying MDR-TB. This radiological feature may offer good specificity at the cost of low sensitivity (13). Our work confirms that there is a significant association between DR-TB and multiple cavities ($P<0.001$), and also quantitatively demonstrates that a SVM classifier using cavity lesions can predict DR-TB with an average specificity of 80%, at the cost of an average sensitivity of 34%. It was reported that MDR-TB patients are more likely to have large-size lesions, and DS-TB patients are more likely to have small- or medium-size lesions (12). Our study confirms that large nodules and large cavities are more common in DR-TB. In addition, we found that multiple nodules and multiple cavities are more common in DR-TB, which confirms the analysis in (5,13).

In our study, we also found that DR-TB patients are more likely to have more abnormalities in all the six lung sextants. *Figure 9* shows the abnormality distribution in the six lung sextants for DR-TB and DS-TB patients, using an abnormality occurrence index (AOI) that is calculated by dividing the sum of abnormalities of all patients in a given group for a given sextant by the number of patients in this group and sextant. A higher index indicates a higher possibility of abnormality occurrence in that sextant. In our future work, we will investigate the possibility of incorporating abnormality location into features for automated identification of DR-TB.

Conclusions

In this paper, we investigated the possibility of using the number of affected sextants for drug resistance prediction and the possibility of applying machine learning to discriminate between drug-resistant TB and drug-sensitive TB by incorporating both clinical and radiological features. We found that, clinical features can predict DR-TB cases with an accuracy of around 61%, with a relatively low sensitivity, while radiological features based on the number of affected sextants can predict DR-TB cases with an accuracy of around 67%, with low specificity. The combination of clinical and radiological features improves these results. For the combined features, our machine classifier achieves an average accuracy of 72.34% and an average AUC value of 78.42%. Our study suggests that the number of affected sextants can be used for identifying drug-resistant TB, and that automatic discrimination between drug-resistant TB and drug-sensitive TB is possible by utilizing both clinical features and radiological features.

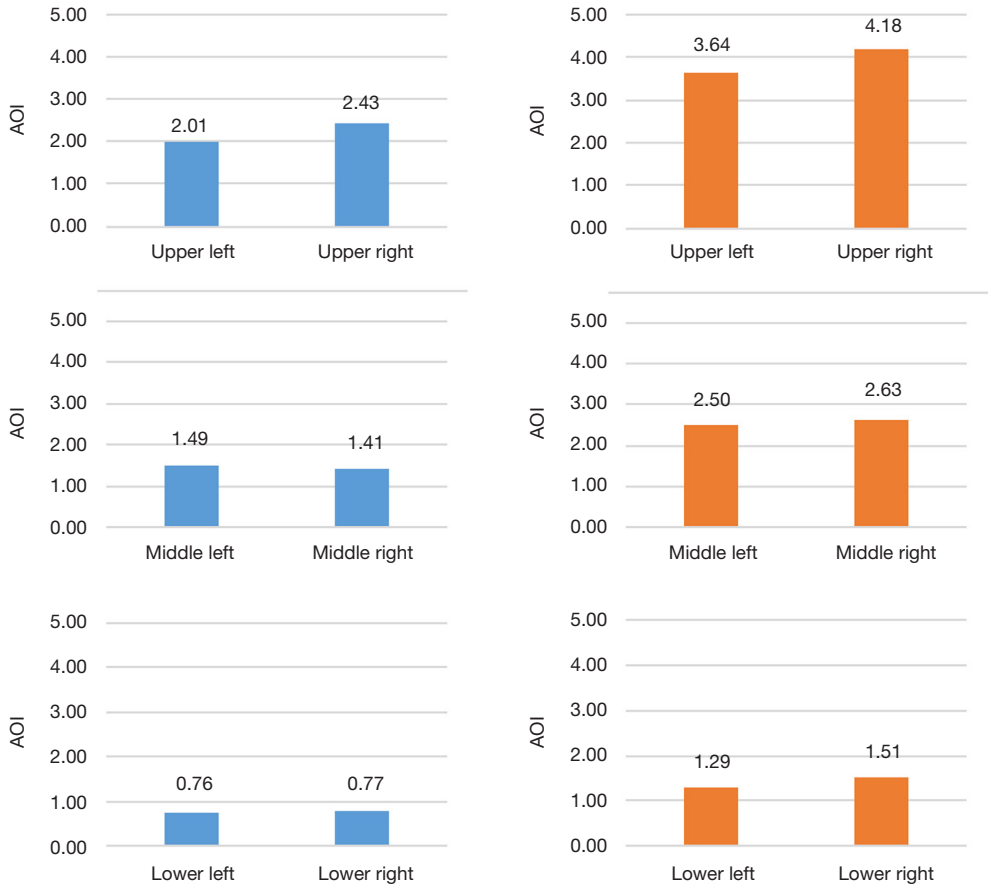


Figure 9 Abnormality distribution for 782 DS-TB patients (blue) and 1,455 DR-TB patients (orange). Shown are the abnormality occurrence indices (AOI) in each lung sextant (upper left, upper right, middle left, middle right, lower left and lower right). Taking DS-TB patients for example, the abnormality occurrence index in a given sextant is calculated by dividing the number of abnormalities in this sextant for all DS-TB patients by the number of DS-TB patients. A higher index indicates a higher possibility of abnormality occurrence in this sextant. DS-TB, drug-sensitive tuberculosis; DR-TB, drug-resistant tuberculosis.

Acknowledgments

Funding: 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, National Institutes of Health. It had been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases under BCBB Support Services Contract HHSN316201300006W/HHSN27200002. This research was also supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the National Library of Medicine. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of NIH, NLM, DOE, or ORAU/ORISE.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/qims-21-290). Dr. SJ serves as an unpaid editorial board member of *Quantitative Imaging in Medicine and Surgery*. The other authors have no conflicts of interest

to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Data usage is exempt from local institutional review board review as it is publicly available from the TB portals program. The TB portals program participants are responsible for ensuring compliance with their countries' laws, regulations, and ethics considerations.

References

1. WHO. Module 1: Tuberculosis preventive treatment. Consolidated guidelines on tuberculosis 2020.
2. WHO. Global Tuberculosis Report 2020: Executive summary 2020.
3. World Health Organization (WHO). Technical report on critical concentrations for TB drug susceptibility testing of medicines used in the treatment of drug-resistant TB. WHO 2018.
4. Heyckendorf J, Andres S, Köser CU, Olaru ID, Schön T, Sturegård E, Beckert P, Schleusener V, Kohl TA, Hillemann D, Moradigaravand D, Parkhill J, Peacock SJ, Niemann S, Lange C, Merker M. What Is Resistance? Impact of Phenotypic versus Molecular Drug Resistance Testing on Therapy for Multi- and Extensively Drug-Resistant Tuberculosis. Antimicrob Agents Chemother 2018;62:e01550-17.
5. Huang XL, Skrahin A, Lu PX, Alexandru S, Crudu V, Astrovko A, et al. Prediction of multiple drug resistant pulmonary tuberculosis against drug sensitive pulmonary tuberculosis by CT nodular consolidation sign. bioRxiv [Internet] 2019; Available online: https://www.biorxiv.org/content/early/2019/11/07/833954
6. Faustini A, Hall AJ, Perucci CA. Risk factors for multidrug resistant tuberculosis in Europe: a systematic review. Thorax 2006;61:158-63.
7. Tembo BP, Malangu NG. Prevalence and factors associated with multidrug/rifampicin resistant tuberculosis among suspected drug resistant tuberculosis patients in Botswana. BMC Infect Dis 2019;19:779.
8. Mdivani N, Zangaladze E, Volkova N, Kourbatova E, Jibuti T, Shubladze N, Kutateladze T, Khechinashvili G, del Rio C, Salakaia A, Blumberg HM. High prevalence of multidrug-resistant tuberculosis in Georgia. Int J Infect Dis 2008;12:635-44.
9. O'Donnell MR, Zelnick J, Werner L, Master I, Loveday M, Horsburgh CR, Padayatchi N. Extensively drug-resistant tuberculosis in women, KwaZulu-Natal, South Africa. Emerg Infect Dis 2011;17:1942-5.
10. Shen X, DeRiemer K, Yuan ZA, Shen M, Xia Z, Gui X, Wang L, Gao Q, Mei J. Drug-resistant tuberculosis in Shanghai, China, 2000-2006: prevalence, trends and risk factors. Int J Tuberc Lung Dis 2009;13:253-9.
11. Lv XT, Lu XW, Shi XY, Zhou L. Prevalence and risk factors of multi-drug resistant tuberculosis in Dalian, China. In: Journal of International Medical Research 2017:1779-86.
12. Icksan AG, Napitupulu MRS, Nawas MA, Nurwidya F. Chest X-Ray Findings Comparison between Multi-drug-resistant Tuberculosis and Drug-sensitive Tuberculosis. J Nat Sci Biol Med 2018;9:42-6.
13. Wáng YXJ, Chung MJ, Skrahin A, Rosenthal A, Gabrielian A, Tartakovsky M. Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences. Quant Imaging Med Surg 2018;8:161-73.
14. Flores-Treviño S, Rodríguez-Noriega E, Garza-González E, González-Díaz E, Esparza-Ahumada S, Escobedo-Sánchez R, Pérez-Gómez HR, León-Garnica G, Morfín-Otero R. Clinical predictors of drug-resistant tuberculosis in Mexico. PLoS One 2019;14:e0220946.
15. Kovalev V, Liauchuk V, Kalinovsky A, Rosenthal A, Gabrielian A, Skrahina A, Astrauko A, Tarasau A, Kalinouski A, Rosenthal A, Gabrielian A, Skrahina A, Astrauko A, Tarasau A. Utilizing radiological images for predicting drug resistance of lung tuberculosis. In: Computer Assisted Radiology and Surgery 2015:S129-30.
16. Jaeger S, Juarez-Espinosa OH, Candemir S, Poostchi M, Yang F, Kim L, Ding M, Folio LR, Antani S, Gabrielian A, Hurt D, Rosenthal A, Thoma G. Detecting drug-resistant tuberculosis in chest radiographs. Int J Comput Assist Radiol Surg 2018;13:1915-25.

17. Rosenthal A, Gabrielian A, Engle E, Hurt DE, Alexandru S, Crudu V, et al. The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. J Clin Microbiol 2017;55:3267-82.
18. Afsar I, Gunes M, Er H, Gamze Sener A. Comparison of culture, microscopic smear and molecular methods in diagnosis of tuberculosis. Rev Esp Quimioter 2018;31:435-8.
19. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn 1995;20:273-97.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
21. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013;14:106.
22. Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: Lecture Notes in Computer Science 2005:878-87.
23. Nekooeimehr I, Lai-Yuen SK. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. Expert Syst Appl 2016;46:405-16.
24. Zhang X, Andersen AB, Lillebaek T, Kamper-Jørgensen Z, Thomsen VØ, Ladefoged K, Marrs CF, Zhang L, Yang Z. Effect of sex, age, and race on the clinical presentation of tuberculosis: a 15-year population-based study. Am J Trop Med Hyg 2011;85:285-90.
25. WHO. Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response [Internet]. Geneva: World Health Organization 2010. Available online: http://whqlibdoc.who.int/publications/2010/9789241599191_eng.pdf
26. Hertz D, Dibbern J, Eggers L, von Borstel L, Schneider BE. Increased male susceptibility to Mycobacterium tuberculosis infection is associated with smaller B cell follicles in the lungs. Sci Rep 2020;10:5142.
27. Neyrolles O, Quintana-Murci L. Sexual inequality in tuberculosis. PLoS Med 2009;6:e1000199.
28. Murphy ME, Wills GH, Murthy S, Louw C, Bateson ALC, Hunt RD, McHugh TD, Nunn AJ, Meredith SK, Mendel CM, Spigelman M, Crook AM, Gillespie SH; REMoxTB consortium. Gender differences in tuberculosis treatment outcomes: a post hoc analysis of the REMoxTB study. BMC Med 2018;16:189.
29. El-Solh AA, Hsiao CB, Goodnough S, Serghani J, Grant BJ. Predicting active pulmonary tuberculosis using an artificial neural network. Chest 1999;116:968-73.
30. Yeh JJ, Chen SC, Teng WB, Chou CH, Hsieh SP, Lee TL, Wu MT. Identifying the most infectious lesions in pulmonary tuberculosis by high-resolution multi-detector computed tomography. Eur Radiol 2010;20:2135-45.
31. Bhalla AS, Goyal A, Guleria R, Gupta AK. Chest tuberculosis: Radiological review and imaging recommendations. Indian J Radiol Imaging 2015;25:213-25.

Cite this article as: Yang F, Yu H, Kantipudi K, Karki M, Kassim YM, Rosenthal A, Hurt DE, Yaniv Z, Jaeger S. Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features. Quant Imaging Med Surg 2022;12(1):675-687. doi: 10.21037/qims-21-290

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 $\geqslant$3 and size $\geqslant$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 $\mathbf{Y}$:

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

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.

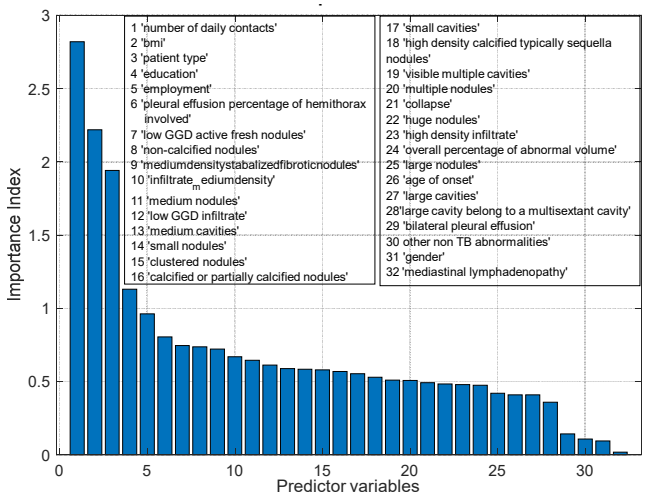


Fig. 2. Importance index of random forest model for the seven demographic features and 25 radiological features. Predictors: 1 - 'number of daily contacts', 2 - 'bmi', 3 - 'patient type', 4 - 'education', 5 - 'employment, 6 - 'pleural effusion percentage of hemithorax involved', 7 - 'low GGD active fresh 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.

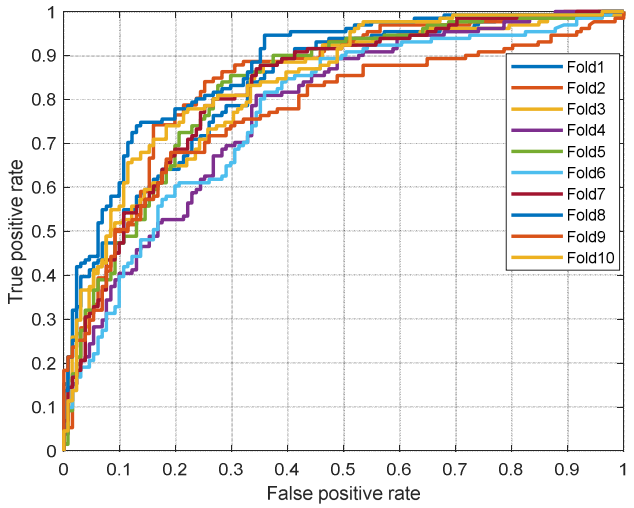


Fig. 3. ROC curves for ten-fold cross validation using random forest classifier based on the top 10 features.

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.

diagnostics

MDPI

Article

Generalization Challenges in Drug-Resistant Tuberculosis Detection from Chest X-rays

Manohar Karki [1,*], Karthik Kantipudi [2,*], Feng Yang [1], Hang Yu [1], Yi Xiang J. Wang [1,3], Ziv Yaniv [2] and Stefan Jaeger [1,*]

[1] Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD 20894, USA; feng.yang2@nih.gov (F.Y.); hang.yu@nih.gov (H.Y.); yixiang_wang@cuhk.edu.hk (Y.X.J.W.)
[2] Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20894, USA; zivyaniv@nih.gov
[3] Department of Imaging and Interventional Radiology, Faculty of Medicine, The Chinese University of Hong Kong, Prince of Wales Hospital, New Territories, Hong Kong
* Correspondence: mkarki2@gmail.com (M.K.); karthik.kantipudi@nih.gov (K.K.); stefan.jaeger@nih.gov (S.J.)

Abstract: Classification of drug-resistant tuberculosis (DR-TB) and drug-sensitive tuberculosis (DS-TB) from chest radiographs remains an open problem. Our previous cross validation performance on publicly available chest X-ray (CXR) data combined with image augmentation, the addition of synthetically generated and publicly available images achieved a performance of 85% AUC with a deep convolutional neural network (CNN). However, when we evaluated the CNN model trained to classify DR-TB and DS-TB on unseen data, significant performance degradation was observed (65% AUC). Hence, in this paper, we investigate the generalizability of our models on images from a held out country's dataset. We explore the extent of the problem and the possible reasons behind the lack of good generalization. A comparison of radiologist-annotated lesion locations in the lung and the trained model's localization of areas of interest, using GradCAM, did not show much overlap. Using the same network architecture, a multi-country classifier was able to identify the country of origin of the X-ray with high accuracy (86%), suggesting that image acquisition differences and the distribution of non-pathological and non-anatomical aspects of the images are affecting the generalization and localization of the drug resistance classification model as well. When CXR images were severely corrupted, the performance on the validation set was still better than 60% AUC. The model overfitted to the data from countries in the cross validation set but did not generalize to the held out country. Finally, we applied a multi-task based approach that uses prior TB lesions location information to guide the classifier network to focus its attention on improving the generalization performance on the held out set from another country to 68% AUC.

Keywords: Tuberculosis (TB); drug resistance; deep learning; chest X-rays; generalization; localization

Citation: Karki, M.; Kantipudi, K.; Yang, F.; Yu, H.; Wang, Y.X.J.; Yaniv, Z.; Jaeger, S. Generalization Challenges in Drug-Resistant Tuberculosis Detection from Chest X-rays. *Diagnostics* **2022**, *12*, 188. https://doi.org/10.3390/diagnostics12010188

Academic Editors: Philippe A. Grenier, Henk A. Marquering, Sameer Antani and Sivaramakrishnan Rajaraman

Received: 1 December 2021
Accepted: 5 January 2022
Published: 13 January 2022

1. Introduction

According to the 2020 World Health Organization (WHO) report [1], it is estimated that in 2019 about 10 million people fell ill with Tuberculosis (TB) and about 1.4 million died from the disease. Based on the same report, it is estimated that in 2019 about 0.5 million individuals were infected with rifampicin-resistant TB out of which about 400,000 were multidrug-resistant.

Drug-resistant TB (DR-TB) is a growing public health concern requiring longer and more complex treatment than drug-sensitive TB (DS-TB), in addition to incurring higher financial costs. Treatment for DR-TB requires a course of second-line drugs for at least 9 months and up to 20 months, supported by counselling and monitoring for adverse events. In comparison, treatment of DS-TB only lasts between 6–9 months. Early diagnosis of DR-TB is crucial for selecting appropriate, patient-specific, treatment regimens. Thus,

improving early decision making has the potential to increase favorable patient outcomes, combat the spread of infection and reduce the overall financial costs associated with the disease.

Currently, the diagnostic methods for identifying DR-TB infections require culture and drug susceptibility testing. These procedures are not feasible globally, especially for countries unable to scale up their testing capacities. An automated, low cost, computational approach that utilizes readily available resources such as medical images and other clinical information is thus desirable.

In the context of TB diagnosis, automated deep learning based systems which only utilize Chest X-rays (CXRs) have seen significant success, with multiple commercial offerings available [2,3]. In one evaluation study, these systems classified CXRs as TB/not-TB with an Area Under the Curve (AUC) of above 0.9 [2]. In another study, they outperformed radiologists, with two of the systems meeting the WHO's target product profile for triage tests [3].

Currently, discrimination between DR-TB and DS-TB using readily available clinical images and possibly additional clinical side information is still an open problem. In this work, we used both deep and classical machine learning algorithms to classify drug-resistant (DR) and drug-sensitive (DS) tuberculosis in chest X-ray images, radiological features, and clinical patient information. Specifically, we have:

- analyzed a state-of-the-art classifier in terms of its capability to generalize on unseen data from another country. This has been an issue largely neglected by the research community in the past. However, we show that a high classification performance is not sufficient for practical usefulness. The capability to provide consistent performance across different datasets, hospitals, and countries, is essential;
- investigated the problem of poor generalization to unseen data by comparing the performance of the deep learning based classifier with other classifiers trained on texture features extracted from X-ray images. We also explore if these generalization problems exist in other clinical data, and if non-disease related attributes such as the origin of chest X-rays can influence the drug resistance detection performance;
- studied explicit and implicit ways to steer the attention of the classifier. We use segmented lungs as a means to guide the network to learn explicitly from the lungs. We also propose a novel multi-task approach that uses prior information (TB lesions' locations) to implicitly focus on the important regions and improve DR/DS classification performance.

2. Previous Work

Attempts to utilize images and clinical data to distinguish between DR-TB and DS-TB have been previously described in multiple publications. These works are either based on the utilization of radiological findings identified in the image by a clinician or via fully automated methods which receive as input the image and potentially other available clinical data and output the likelihood for each of the two classes.

Several studies have shown that radiological findings based on a radiologist reading of a CT or CXR have the potential to differentiate between the two classes. A literature review from 2018 [4] concluded that the presence of thick-walled multiple cavities in the images is a useful predictor for DR-TB, with good specificity but low sensitivity. Another study [5] compared 183 DR-TB cases and 183 DS-TB cases from a single hospital. This study concluded that there were substantial differences in findings between the two classes in terms of lesion size and morphology. A slightly larger study [6], which compared 468 DR-TB cases and 223 DS-TB cases concluded that a combination of the number and size of consolidated nodules is a good predictor for DR-TB. Another, small study [7], utilized data from 144 patients and found that the presence of multiple cavities is a good predictor for DR-TB. A much larger study [8], compared 516 DR-TB and 1030 DS-TB cases, obtaining an AUC of 0.83 using a regression model. This study observed that the co-existence of multiple findings (multiple cavities, thick-walled cavities, disseminated lesions along the

bronchi, whole-lung involvement) was indicative of DR-TB. Finally, more recent work [9] compared 1455 DR-TB and 782 DS-TB cases, using two clinical features and 23 types of radiological findings. A support vector machine was used to distinguish between DR-TB and DS-TB with an AUC of 0.78. It should be noted that reliance on a radiologist reading is a significant limitation. The lack of consensus on radiological findings for drug resistance further hinders the clinical usefulness of these approaches. Because of these reasons, fully automated solutions, described next, are more desirable.

Several fully automated solutions were presented as part of the ImageCLEF 2017 and 2018 evaluation challenge forums [10]. These challenges included a subtask, differentiating between DR-TB and DS-TB using thoracic Computed Tomography (CT) images. This classification task included 259 training images and 236 test images with about half of the cases DR-TB and half DS-TB. Proposed solutions included Gentili et al. [11] who reformatted the CT images to the coronal plane and used a pre-trained ResNet50 Convolutional Neural Network (CNN). For the same challenge, Ishay et al. [12] used an ensemble of 3D CNNs and Cid et al. [13] used a 3D texture-based graph model and support vector machines (SVM). Allaouzi et al. [14] replaced the softmax function of a 3D CNN architecture with an SVM to tackle this classification task. All entries had limited success, resulting in AUCs of about 0.6. After two editions, the organizers removed the subtask from the competition with the conclusion that "the MDR subtask was not possible to solve based only on the image". While these challenges did not yield the desired results, the results obtained using radiologist readings are more favorable, suggesting that the sub-optimal performance may be due to the small number of images available for training. It should be noted that increasing the number of CTs for this task is not trivial as the use of CT imaging in DS-TB cases is uncommon, with the standard imaging modality being CXR. The rare use of CT imaging in standard practice, and the consequential lack of data to analyze, limits the usage of CT images to train a model to distinguish between DR and DS-TB.

On CXR images, [15] utilized a customized CNN architecture to classify DR-TB and DS-TB from 2973 images from the TB portals dataset. They achieved a classification performance of 66%, which improved to 67% when follow up images were also included. Our group has previously proposed fully automated methods utilizing CXRs as described in [16,17]. In [16], we utilized 135 CXRs from a single source. Using a shallow neural network we obtained an AUC of 0.66. In [17], we utilized a much larger dataset, 3642 images from multiple sources. Using a deep neural network, InceptionV3 pre-trained on ImageNet, we obtained an AUC of 0.85. This result is the current state-of-the-art performance achieved on the TB portals data. This is a significant improvement of results from other approaches. However, even though a 10-fold cross validation was performed, the capability of the trained network to classify chest X-rays from unseen domains was not evaluated. In fact, the common weakness of all of these automated methods is that they have not been evaluated for generalization by separating the source of the data. As different medical imaging technologies and devices produce different standards and quality of images, it is important for our models to be robust to these changes.

An underlying assumption of most machine learning algorithms is that the population, test, and training data are independent and identically distributed. If the two distributions are different, then the learned parameters will not yield a good performance. That is, the model will not generalize well to unseen data. While CXR imaging is a low cost modality that is in widespread use, the variations in the standards of the acquired images is significant [18,19], bringing into question the utility of any proposed method which is not evaluated on its generalization capability. More specifically, Harris et al. [20] found that 80% of published works on using CXR for TB diagnosis either used the same databases to train and test their software, or did not comment on databases they used for testing their models. Sathitratanacheewin et al. [21] also observed that a model for CXR-based TB diagnosis performed well with 0.85 AUC when tested on images within their intramural dataset with significant performance deterioration when tested on extramural images, yielding an AUC of 0.7. For domain shift, when the change in image distribution between the training and

testing sets is inevitable, it has been shown that these effects can be ameliorated if training is formulated using a multi-task approach [22].

In addition to the generalization issues due to domain shift, the generalization of deep learning algorithms can also deteriorate if they learn irrelevant features. This is a specific shortcoming of deep learning algorithms as they do not preclude the algorithm from learning features present in the training set that are arbitrarily correlated with the disease, yet are completely irrelevant. These can stem from characteristics of the imaging devices or clinical practices such as patient positioning [23–25] used at the specific locations. If a model implicitly learns such features it will not generalize well when presented with data obtained on different imaging devices or using different clinical workflows, both of which are irrelevant to disease diagnosis.

In this work, we explore various strategies to improve the generalization of models for classification of CXRs as DR-TB or DS-TB using various normalization and attention mechanisms, both explicit (segmentation based) and implicit (multi-task based).

3. Data

3.1. TB Portals Data

The primary data source used in this work is from the NIAID TB Portals program (https://tbportals.niaid.nih.gov (accessed on 10 January 2022)), with a public data release date of October 2020. The dataset contains clinical data and CXR images that are anonymized and made available for public use [26]. Each patient record is manually annotated with clinical information and radiological findings based on the associated CXR image. For this work, data from 1756 patients from ten countries were used. Table 1 shows the data distribution based on country of origin and gender. It should be noted that the TB portals data were collected with a primary focus on acquisition of drug-resistant cases and cases that reflect the specific research interest at the country of origin. As a result, the data are imbalanced in terms of the ratio between drug-resistant and drug-sensitive cases, which does not necessarily reflect the prevalence of TB from either class in the contributing country. Interestingly, we also see that the data are not balanced in terms of gender with about double the number of males to females. This does reflect known differences in TB prevalence in females versus males and has been linked to both societal and biological differences between the sexes [27–29].

Table 1. Patient distribution from different countries and genders for the chest X-ray data used in this work.

Country	Number of Patients			
	Drug-Sensitive	*Drug-Resistant*	*Male*	*Female*
Belarus	118	344	294	168
Georgia	399	236	472	163
Romania	15	114	91	38
Azerbaijan	0	32	24	8
India	197	21	165	53
Moldova	12	32	37	7
Kyrgyzstan	0	18	11	7
Ukraine	8	25	25	8
Kazakhstan	15	53	36	32
South Africa	114	3	72	45
Total	878	878	1227	529

3.2. Clinical Data

The clinical data contain an extensive set of features associated with each patient. This includes demographic data, radiologists' findings for each CXR, different diagnostic tests and treatment information. Additionally, it includes demographic features such as age of onset, gender, patient type (New, Relapse or Failure), body mass index, country of

origin, education, employment, number of daily contacts, number of children, prescription drug usage, laboratory tests, treatment period, treatment status and outcome. The radiologists' findings include chest radiography patterns such as nodules, cavities, collapses and infiltrates and their location in the lungs. Due to financial constraints and the size of the TB portals CXR dataset, radiological findings 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 findings are not biased towards a single radiologist. Table 2 lists all finding types used by the radiologists to annotate the images. These are abnormalities commonly associated with TB. In addition to the type of abnormality, the findings are further differentiated based on their size (small, medium, large) and number of occurrences (single, multiple).

Table 2. Twenty features derived from the presence of abnormalities that are localized to different sextants.

Types of Abnormalities	
collapse	small nodules
small cavities	medium nodules
medium cavities	large nodules
large cavities	huge nodules
large cavity belonging to multiple sextants	non-calcified nodule
multiple cavities	clustered nodules
low ground glass density, active fresh nodules	multiple nodules
medium density stabilized fibrotic nodules	infiltrate: low ground glass density
high density calcification, typically sequella	infiltrate: medium density
calcified or partially calcified nodule	infiltrate: high density

3.3. Chest X-ray Images

All TB Portals CXRs used in this work are from a frontal, AP or PA, view and have varied resolutions (206 × 115 to 4453 × 3719). The intensity range found in the images also varies, with 1177 images having a low dynamic, intensities in the 0–255 range, and 579 images having a high dynamic, intensities in the 0–65,536 range.

It should be noted that the drug susceptibility label associated with each image is obtained via drug susceptibility testing and is not derived from the image. Additionally, the usage of radiological findings for predicting drug susceptibility has shown moderate success. Thus, the question of whether good performance for predicting drug susceptibility from CXRs from unseen sources is possible remains open.

In addition to the CXRs from the TB Portals program, we use a publicly available TB CXR dataset collected from a hospital in China [30] (Download from http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip (accessed on 10 January 2022)). This dataset contains 662 frontal chest X-rays, of which 326 are labeled as non-TB cases and 336 are labeled as TB. There are two sets of annotations where each abnormal TB image has been manually annotated by two radiologists. Figure 1 shows one such segmentation.

Sextant Division

To further differentiate between radiological findings, we associate them with their spatial location in the lungs. To do so, we define lung sextants by dividing each lung into three equal sections from apex to base, as shown in Figure 2. The division of the sextants can be subjective for findings close to sextant borders, when the division boundaries may not be strictly adhered to by the annotating radiologist. In this work, we say a sextant is affected by TB if at least one of the abnormalities listed in Table 2 is present in the sextant.

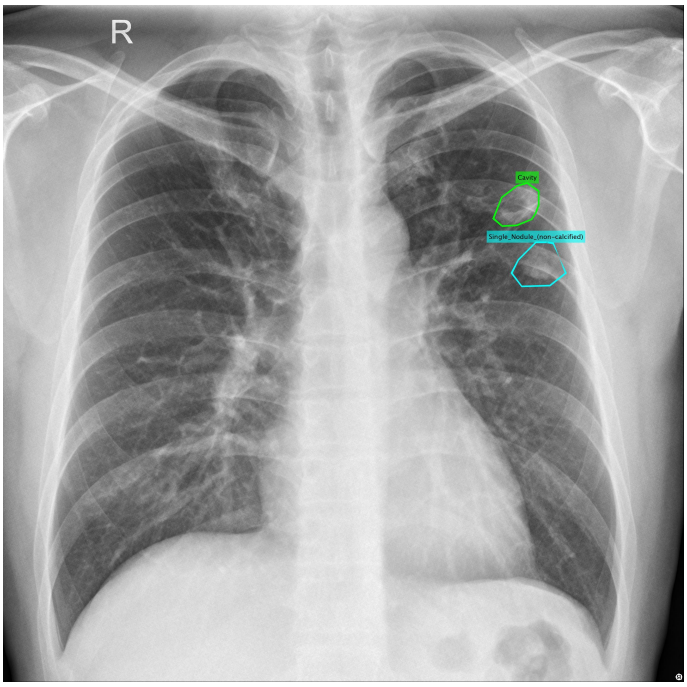


Figure 1. Example of a lung segmentation for a nodule and a cavity.

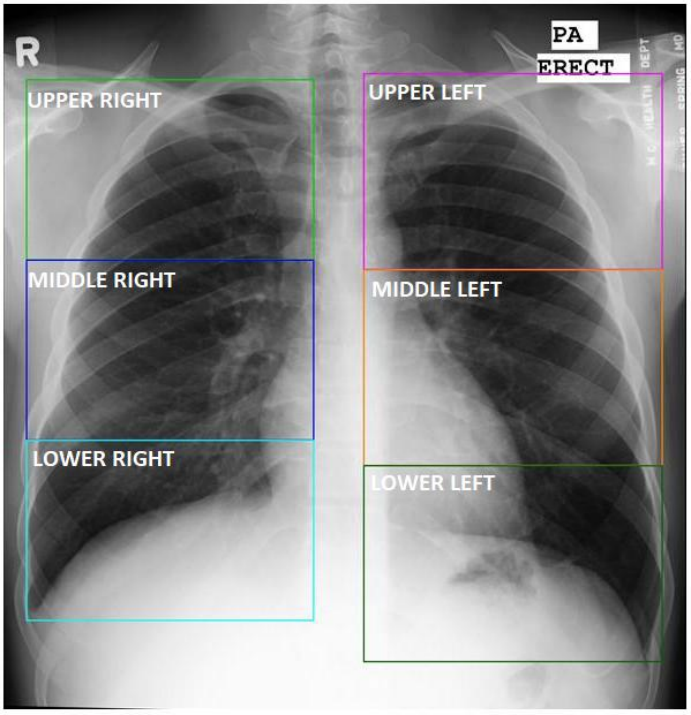


Figure 2. Definition of six lung sextants. Abnormal annotations are assigned to one or more of these sextant divisions.

3.4. Dataset Definitions

For our experiments, we only select the first image taken in the clinical process for a patient; hence, the number of images is equal to the number of patients. All of our drug resistance classification experiments feature an equal number of DR and DS patients in the

training set. For data balancing, we use a conservative approach, excluding images from the majority class. The subsets used for training and evaluation are listed below:

- **Generalization Dataset (Gen. Dataset)**: A total of 1520 samples are selected for this set, 760 samples for each class. All samples originating from the country Belarus are excluded.
- **Dataset with sextant annotations (Sext. Dataset)**: This set contains a total of 1118 samples, 559 samples for each class. This set also does not include any data from Belarus.
- **Validation Set**: This is a cross validation set, which varies for each fold. It contains a randomly selected set containing 20% of the dataset (5-fold CV) for each cross validation training. The numbers reported for the validation set are the average performance values of all cross validation folds. Because, this would be a subset of the above two datasets, no samples from Belarus will be present in this set either.
- **Belarus Dataset**: This dataset contains a maximum of 118 samples from each class with 236 samples in total. The Belarus dataset is used as the test set for most experiments. When sextant-based data should be required, five samples from each class are removed as they do not contain sextant information.

3.5. Data Standardization

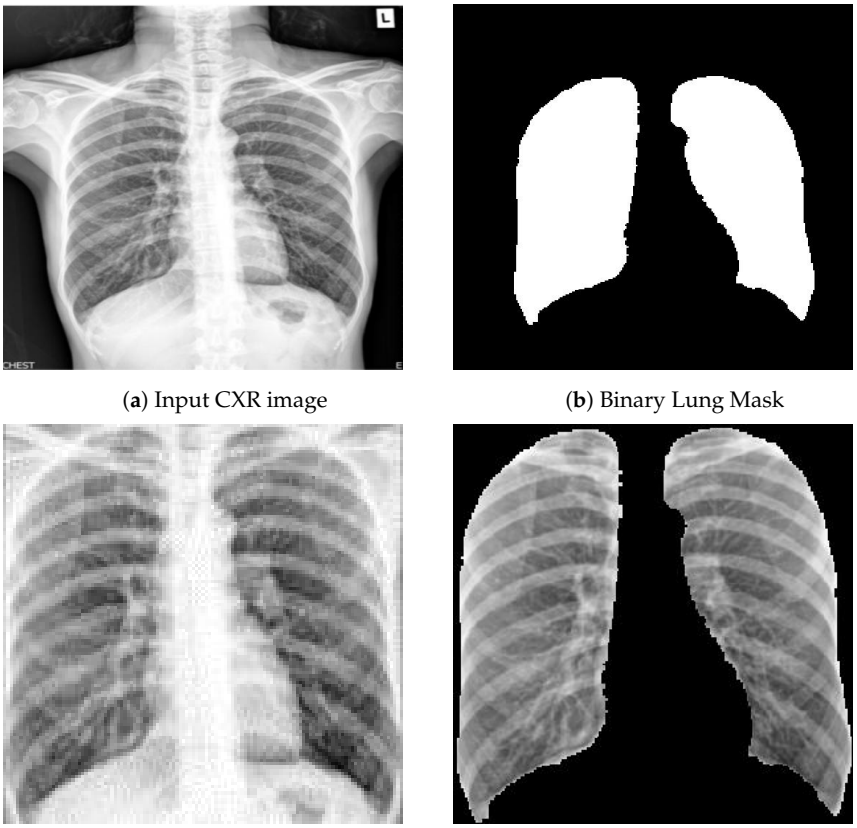
Lung segmentation is used to explicitly address the challenges associated with generalization due to domain shift and the possible existence of confounding factors due to class-correlated yet irrelevant features. Segmentation enables us to limit the input images for the binary DR/DS classifier so that they only contain regions relevant for classification of pulmonary tuberculosis, meaning the lungs. Additionally, the lungs are scaled to a uniform size and position within the image, removing potential confounding factors such as lung size and patient placement that are often correlated with the clinical sites and thus with the local prevalence of TB types. Once the lung regions are segmented, the image is cropped to the lung bounding box, Figure 3c, and all information outside the lung is removed, Figure 3d.

For lung segmentation we initially utilized a publicly available U-Net model which was trained on two datasets with a total of 385 images and corresponding manual lung segmentations [30,31] (https://github.com/imlab-uiip/lung-segmentation-2d (accessed on 10 January 2022)). Unfortunately, this model failed frequently when applied to the TB portals images. Often, one or both sides of the lung were not segmented.

Furthermore, segmentation using this model failed on pathological lung regions in a significant number of images, which is detrimental for disease analysis.

To address these performance limitations a U-Net based [32] segmentation model with a ResNet50 backbone [33] was trained using the publicly available v7 COVID-19 X-ray dataset, which contains 6500 images and corresponding manual lung segmentations (https://github.com/v7labs/covid-19-xray-dataset (accessed on 10 January 2022)).

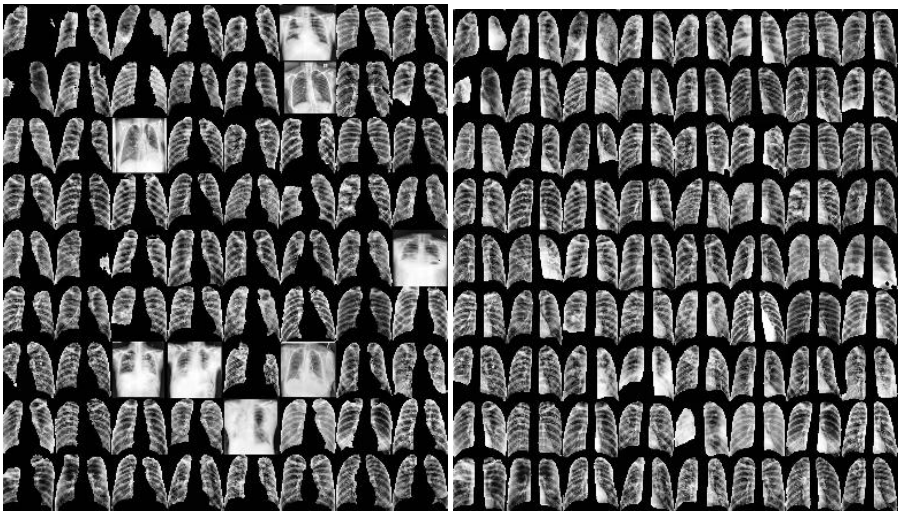
As the TB portals dataset does not provide ground truth lung segmentations, results were visually evaluated as either failure or success. The segmentation failure rates of this model and the previous model were 0.06% and 3% respectively. Aside from that, the old model segmented one of the lungs with less than 10% of the corresponding ground truth pixels in 0.8% of the cases. No such cases were observed in the new model. Figure 4 illustrates the difference between the two models applied to the same set of 72 images.



(**a**) Input CXR image (**b**) Binary Lung Mask

(**c**) Cropped Image (**d**) Lung Segmented

Figure 3. Original CXR (**a**) is fed to the U-Net, which outputs a binary lung mask (**b**) with which the original CXR is cropped(**c**) and the lungs are segmented (**d**) in the cropped bounding box.



(**a**) (**b**)

Figure 4. Cropped images based on the lung segmentation results obtained using a publicly available UNet model trained on the combined JSRT and Montgomery datasets (**a**), and using a customized UNet model trained on the v7 COVID-19 X-ray dataset (**b**). The performance of the customized model is clearly better. Note that if the segmentation fails, the entire image is used.

4. Drug Resistance Classification

For classifying between drug-resistant and drug-sensitive TB, we primarily use the chest X-rays but also utilize text data to assist with the classification and to compare the performance when using just the images. Classic machine learning algorithms and CNNs with pretrained weights were used on the clinical text data and chest X-ray images respectively. Figure 5 shows the setup of our classification network where the preprocessed chest X-ray image is the input and the prediction is either drug-resistant (DR) or drug-sensitive (DS).

As the focus of this work is to evaluate, understand and propose solutions to the issue of generalization to unseen data, we describe in the following subsections: (a) the need of domain adaptation for a network to generalize to unseen data, (b) the use of radiomics features derived from chest X-rays, (c) multi-task learning as a means to provide implicit attention to the main task of DR/DS classification, (d) classifying per-sextant abnormality, and (e) segmenting abnormal regions.

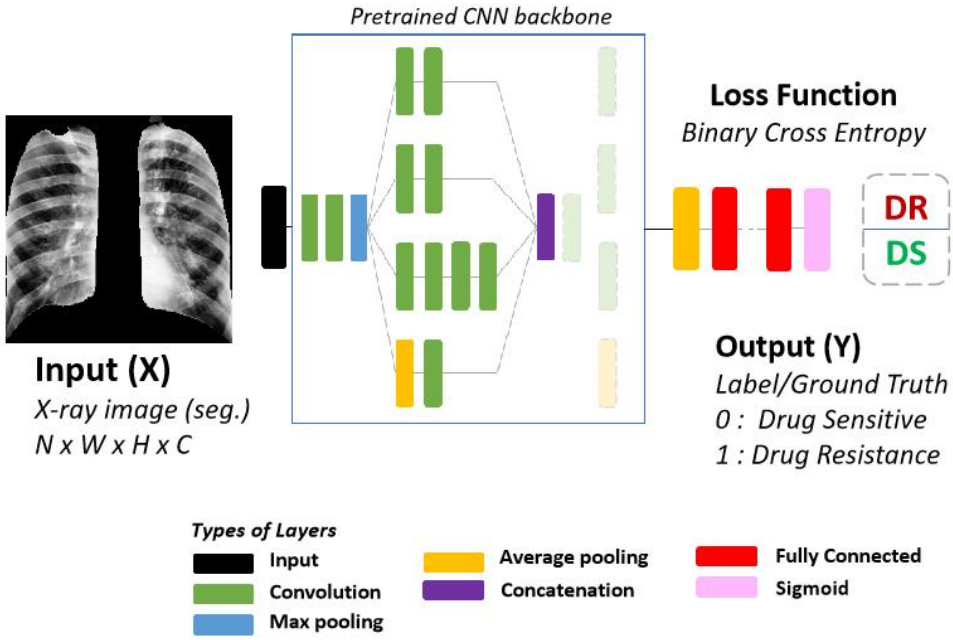


Figure 5. Standard CNN architecture for drug resistance classification. The Input (X) is a preprocessed X-ray image with segmented lungs. The Output (Y) is one of two classes, DR-TB or DS-TB. For this work, we use the ResNet18 [33] architecture as the backbone for all of our experiments.

4.1. Domain Adaptation

The distribution for which a trained model is tested can often be significantly different from the distribution that it was trained on. There is no guarantee that a trained model will be robust to data it has not seen before. Different acquisition standards [34], equipment, and even personnel can create vastly different looking images for medical images. Even after acquisition, other processing and storage differences can create differences in the images. For a human, these variations may be easier to overcome but a machine learning model needs to be trained to understand the differences. Either smart features and algorithms need to be employed or a large and diverse set of data is required to train such a model.

Evaluation of models on unseen data from different domains is the logical way to evaluate such models. Besides that, interpreting the model's decision can also be valuable to understand a prediction. For drug resistance TB classification, localizing the prediction decisions is worthwhile, as tuberculosis itself is frequently observed in certain regions of the lung.

Furthermore, it is worth exploring how easily images from different domains can be discriminated. Easily distinguishable domains in the input coupled with an imbalanced dataset can readily result in failure to generalize.

The usage of transfer learning enables a model to adapt to the new domain, but is less desirable when compared to a fixed model which does not require additional training per domain. Starting from pretrained weights allows for the high-level features to be consistent and not overly dependent on the domain of the training images. This explains why networks initialized with pretrained weights consistently outperformed networks randomly initialized [17].

4.2. Radiomics Features

Usage of explicit, engineered features can be used as a counter measure to prevent a network from learning correlated, yet irrelevant, features within a dataset. Radiomic features have been used to extract patterns that may have been missed by radiologists to identify abnormalities present in medical images [35].

The features used in the paper include 2D-shape based features (e.g. axis lengths), first order statistics (e.g. skewness), gray-level co-occurence matrix features (GLCM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM) and neighbouring gray-tone difference matrix (NGTDM) features.

4.3. Multi-Task Learning

To encourage a network to focus on desirable localities, such that the network is generalizing based on the actual abnormalities within the provided anatomical regions, adding a secondary task sharing some of the features is a promising approach. When networks have been trained to predict different but related tasks in tandem, performance on each of the tasks have benefited [36]. The auxiliary information from a secondary task can be beneficial to the main task and is useful to regularize the network as well.

A big motivation behind using multi-task learning for this work is the availability of the radiologists' annotations for different abnormalities in the lung, localized to the six divisions (sextants) of the lungs. While this information will not be available during testing, it can be utilized to regularize the main model and to focus the attention of the network to supposedly relevant areas of the image. For our drug resistance classification, the network consists of a pretrained CNN network that is trained with binary cross entropy loss. For multi-tasked networks, the combined loss [37] for the two tasks is as follows:

$$\mathcal{L} = \frac{1}{\sigma_1^2} \times \mathcal{L}_1(W) + \frac{1}{\sigma_2^2} \times \mathcal{L}_2(W) + \log \sigma_1 + \log \sigma_2, \tag{1}$$

where W represents the weights of the network, and σ_1 and σ_2 are the noise parameters for the respective tasks, which are used to determine the relative weights given to each of the losses.

4.4. Abnormal Sextant Classification

The abnormal sextant information provides the locations of TB-related abnormalities to the network. These are expert-annotated features that provide additional context and information during DR/DS classification. Figure 6 shows the architecture for this type of multi-task learning. The classification model is modified such that the output of the last convolution layer is diverged into two stacks of fully-connected layers. The first path is the same as the normal architecture where the network decides if the X-ray image shows manifestations of drug resistance or drug sensitivity. The second path outputs a vector of length 6. Each of the six values represents the presence or absence of any of the 20 abnormality features described in the Data section above. Hence, for each sextant, if one of the abnormalities is present, the sextant is considered 'abnormal,' whereas if none of the abnormalities is present, it is considered 'normal.' In Equation (1), $\mathcal{L}_1$ and $\mathcal{L}_2$ are both

binary cross entropy losses in this case. For the secondary task, the loss is the average loss among all six outputs.

4.5. Abnormality Segmentation

Segmenting abnormalities provides location information as the locations of each of the sextants are also available. For this task, the losses from Equation 1 are modified such that $\mathcal{L}_1$ is the binary cross entropy loss and $\mathcal{L}_2$ is the combination of Jaccard loss and binary cross entropy loss for the segmentation of abnormal regions. Figure 7 shows the modification of the base model into an encoder-decoder U-Net style architecture for the additional task of abnormality segmentation. Two approaches are taken into account to determine the abnormal ground truth regions.

4.5.1. Sextant Segmentation from Radiologist Annotation

The sextant annotations from the radiologist are converted into masks such that the location information of each sextant is also available to the network. Each pixel in the sextant with presence of any abnormality is set to 1, and 0 otherwise. This is similar to the sextant classification with added information of the location of each of the sextants.

4.5.2. TB Abnormality Segmentation

Instead of using the abnormalities from clinical text data, we alternatively use the TB abnormality segmentation network to derive the ground truth. The Shenzhen dataset with annotations has a finer segmentation of abnormalities. In this approach, the chest X-ray images are segmented for lesions using the network trained on the Shenzhen data [30]. The advantage of this approach is that even images without annotations can be used for training.

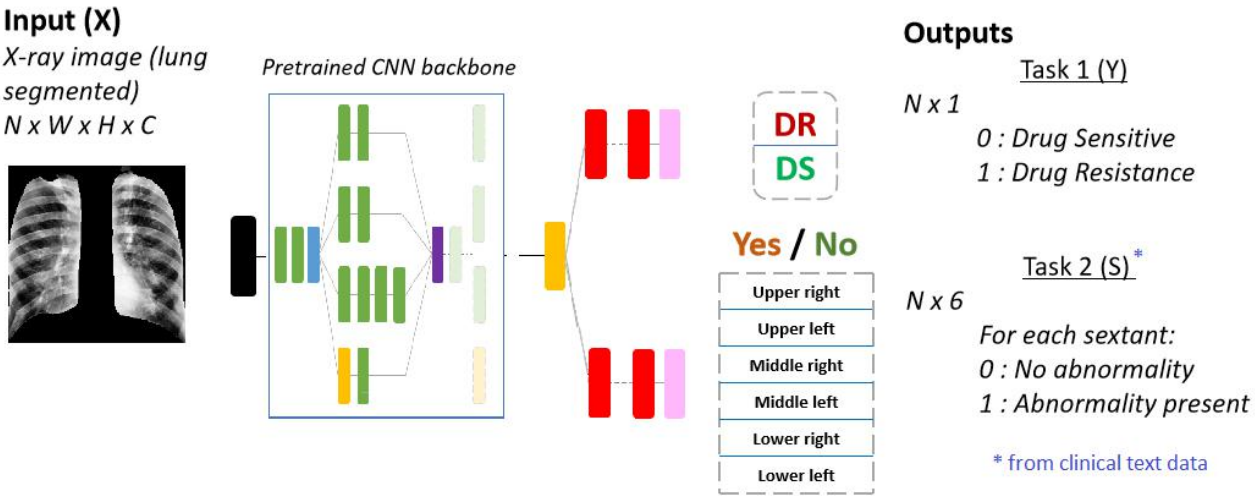


Figure 6. Multi-output network with the same pretrained backbone (ResNet18) for the additional task of abnormal sextant classification. The data used for this task is multi-modal. The inputs to the network are the chest X-ray images, whereas the labels for abnormal sextants are derived from the clinical text data described in Section 3.2.

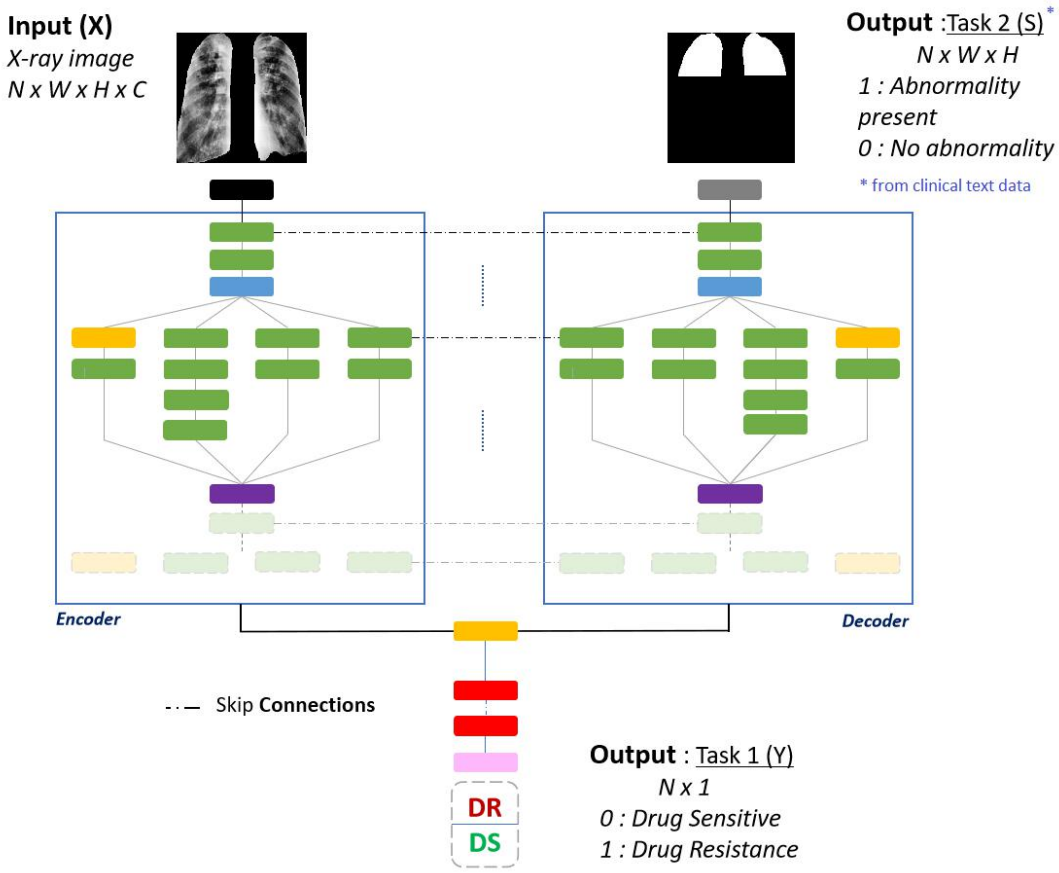


Figure 7. Multi-output network for the additional task of abnormality segmentation. The pretrained backbone (ResNet18) is modified to be a U-Net with encoder and decoder. The inputs for this network are chest X-ray images whereas the segmentation output masks are derived from the clinical text data.

5. Experimental Results

We perform our experiments with a 5-fold cross validation stratification. We also separate 7.5% of the training data, to check inter-epoch performance and stop the model early once the performance degrades for a long period on this set. The pretrained network backbone, as described in Figure 5, is the ResNet18 [33] architecture. The number of parameters for ResNet18 (11 million) are half of that of InceptionV3 (22.3 million), which we previously used [17]. Even with the smaller network and smaller dataset (since samples are held out), the performance on the validation set was 79% AUC. As we convert these networks to a U-Net style segmentation network for secondary tasks, the difference in parameters is increased even more. With the choice of ResNet18, we are able to transfer the pretrained weights from ImageNet [38] and keep the number of total trainable parameters small while having a consistent network to compare different approaches.

5.1. CNN-Based Drug Resistance Classification

To examine if our network is robust against domain changes and if it generalizes well to unseen data, we exclude the data from one country before cross validation stratification and use it as a held out set. As we see from Table 1, the only two countries that have more than 100 samples in each class are Belarus and Georgia. Aside from these two countries, every other country has less than 25 samples in the minority class. Choosing other countries would lead to a highly imbalanced testing set or a testing set with very few samples per class. When we excluded Georgia to use it as a held out set for evaluating generalization, the total number of samples decreased by almost half. There were only 479 samples per

class for the balanced training set. The AUC performance on the validation set was 78% but the performance on the held out Georgia data was at 52%. Also, less than 25% of the Georgia patients had sextant information available and hence the evaluation for the multi-task learning was not feasible. The data from Belarus has been used previously in [16], to both train and evaluate the classification of DR/DS TB from chest X-rays. Because of these reasons, we only used the data from Belarus as our held out set and used Georgia's data as part of the training sets.

For our classification training, we experimented with two different sets of initialization weights. The first set of weights is from the ImageNet classification task and the second set is from the network trained for TB-abnormality segmentation described in Section 5.6. When we trained the model with the cropped images (Figure 3c), similar to the previous approach [17], the performance on the validation set was 73% AUC and on the Belarus dataset it was 55% AUC.

In an effort to improve generalization, explicit attention on the lung regions was provided by setting the areas outside the segmented lung to 0 (Figure 3d). This approach improved the classification performance on each of the datasets. Table 3 shows that the best AUC performance was observed on the validation set, using both the ImageNet classification and TB abnormality segmentation weights, with 79% AUC. On the Belarus dataset, the best AUC of 65% was observed with the dataset with sextant information and with ImageNet weights. Achieving a much better performance with this approach, we use the segmented lungs as an input to the following experiments.

Table 3. DR-TB/DS-TB Classification Performance on the Validation Set and the Belarus Set.

		Validation Set		**Belarus Dataset**	
Trained on	***Initialization Weights***	**AUC**	**Accuracy**	**AUC**	**Accuracy**
Gen. Dataset	*ImageNet classification [38]*	0.79 ± 0.03	**0.72 ± 0.04**	0.60 ± 0.01	0.55 ± 0.03
	TB abnormality segmentation	**0.79 ± 0.02**	**0.72 ± 0.03**	0.60 ± 0.02	0.57 ± 0.02
Sext. Dataset	*ImageNet classification*	0.77 ± 0.03	**0.72 ± 0.03**	**0.65 ± 0.02**	**0.62 ± 0.02**

5.2. Classification with Radiomic Features

With the usage of non-learnable features, some acquisition-specific details can be hidden, which may be easily identified by a sufficiently large deep network. For this purpose, 104 radiomic features are extracted with the aid of the pyradiomics (https://pyradiomics.readthedocs.io/en/latest/features.html (accessed on 10 January 2022)) library [35]. The library calculates the features based on the X-ray image and the mask of the object of interest. We evaluate these features on both the lungs and the rest of the image by providing the lung masks and the complement of the lung masks, respectively. For our classifiers we use standard machine learning algorithms such as support vector classifiers (SVC), k-nearest neighbors (k-NN), Random Forest (RF) and multi-layer Perceptron (MLP).

The support vector classifier achieved the best performance on the validation set and the Belarus dataset, as seen in Table 4. Surprisingly, the best validation performance (74.5%) was computed when the lung region was excluded, that is, only non-lung parts of the image were used to derive these features. The performance on the Belarus dataset was 62.8%.

Table 4. AUC Performance with radiomic features.

Testing Data	**Input Data**	***SVC***	***k-NN***	***RF***	***MLP***
Validation Set (Gen. dataset)	Lung only	0.722	0.681	0.725	0.720
	Lungs excluded	**0.745**	0.688	0.732	0.725
Belarus dataset	Lung only	**0.628**	0.577	0.620	0.620
	Lung excluded	0.583	0.563	0.621	0.530

5.3. Classification with Sextant Divisions

The location of abnormalities are useful for classifying tuberculosis [39,40]. The sextant-based annotations are localized features that show different abnormalities within the lung. We also divide the chest X-rays into six divisions similar to how they were annotated by radiologists.

We classify DR-TB and DS-TB from the annotations acquired and our divided chest X-ray images. As described in Table 2, there are 20 such features for each of the sextants. Hence, there are 120 features in total. Figure 8 shows how abnormalities are more frequent in the apex of the lung.

Figure 9 shows the classification performance when individual sextants, the entire lung, and top, bottom, and middle regions were evaluated regarding DR-TB vs DS-TB classification. On the validation set, the CNN classifier trained on chest X-rays performed indiscriminately to the lung location used for training. With the annotated data and classical machine learning classifiers, the top sextants were more discriminatory than the bottom sextants. On the Belarus dataset, however, classical machine learning classifiers were not able to discriminate between the two classes with much success. An MLP (multi-layer Perceptron) classifier was able to achieve 60% AUC performance. When a single sextant was used, they all performed similarly at around 60%. Training on the entire image yielded the best results (65%) on the CXR images.

When we reduce the number of features to use just the location information or the type of abnormality, providing the location yielded better AUC performance (62.9%) on the Belarus dataset. However, on the validation set, providing the type of abnormality performed better (AUC of 70.0%) as shown in Table 5. 'Location' refers to the presence of any abnormality in sextants whereas 'Type' refers to the presence of one of the 20 abnormalities listed in Table 2 in any area of the lung.

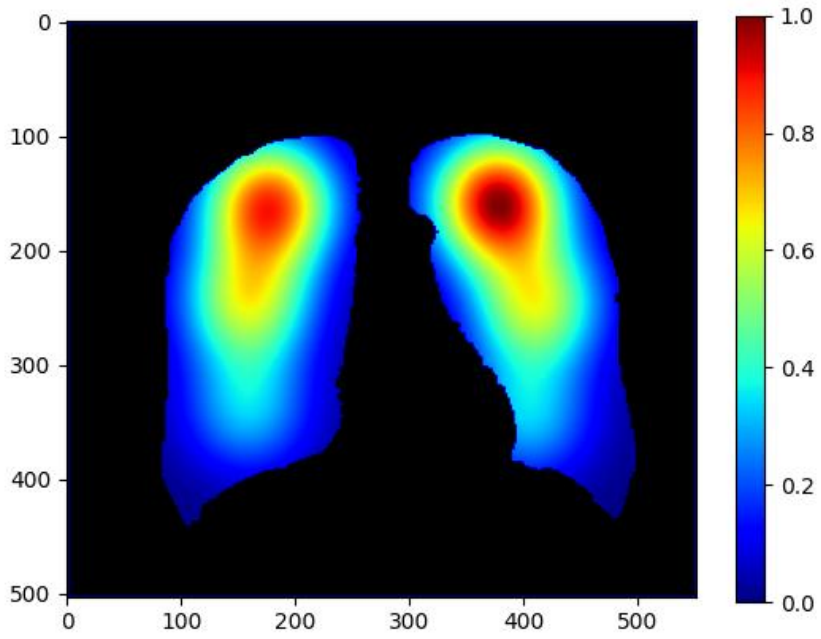
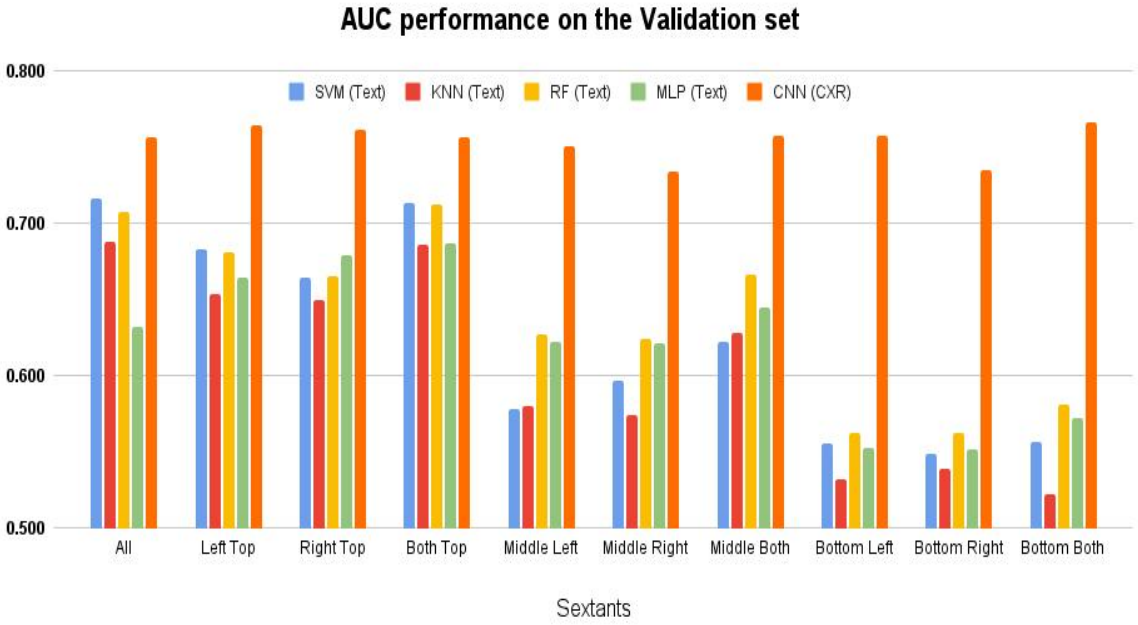


Figure 8. Abnormality occurrence heatmap in different regions of the lung derived from radiologists' annotated sextant data.



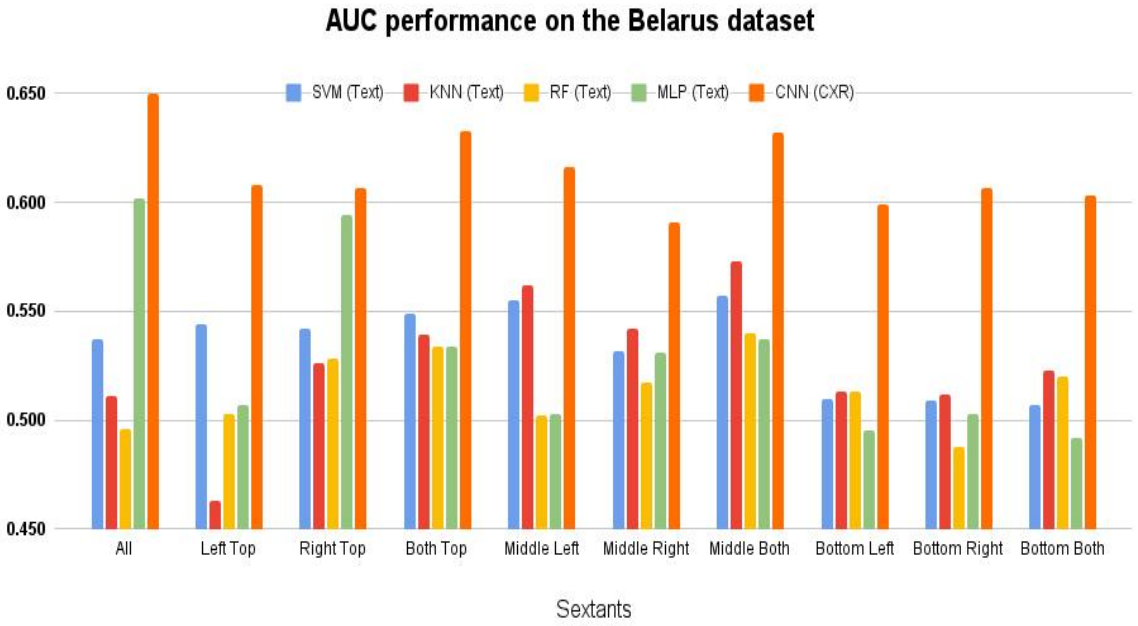


Figure 9. Drug Resistance Classification—AUC performance on sextants.

Table 5. Performance when location and type are separated for annotated radiologic features.

Classifier:	Validation Set		Belarus Dataset	
	Location	Type	Location	Type
SVC	0.573	0.700	**0.629**	0.539
KNN	0.501	0.670	0.484	0.506
RF	0.547	0.687	0.527	0.497
MLP	0.505	0.663	0.596	0.504

5.4. Classification with Data and Network Capacity Limitations

The classification performance with radiomic features derived from the non-lung region also prompted us to further examine the performance of our CNN classifier with limited information in Table 6. The limitations we added are regarding the input data and the training networks.

To further investigate the bias in the data that is supposedly not related to the underlying disease manifestations, we apply limitations to the information received by the network or limit the capacity of the network itself. This was achieved by modifying the data as well as the training network. Figure 10 shows examples of different ways the X-ray images were manipulated to reduce the information input to the classifier network. The information from the chest X-rays were limited or diminished by randomizing pixel locations. For example, in a particular experiment, entire image intensities were randomized. This would conceal the spatial relationships between pixels but still preserve the histograms and first order statistics of image intensity values. For further experiments, only certain regions of the image were randomized and the rest were set to 0. Lung masks were also used as input where the pixel intensity values are lost but the shape of the lungs are still intact. Another approach was subtracting the mean of the background (non-lung) and re-normalizing each image.

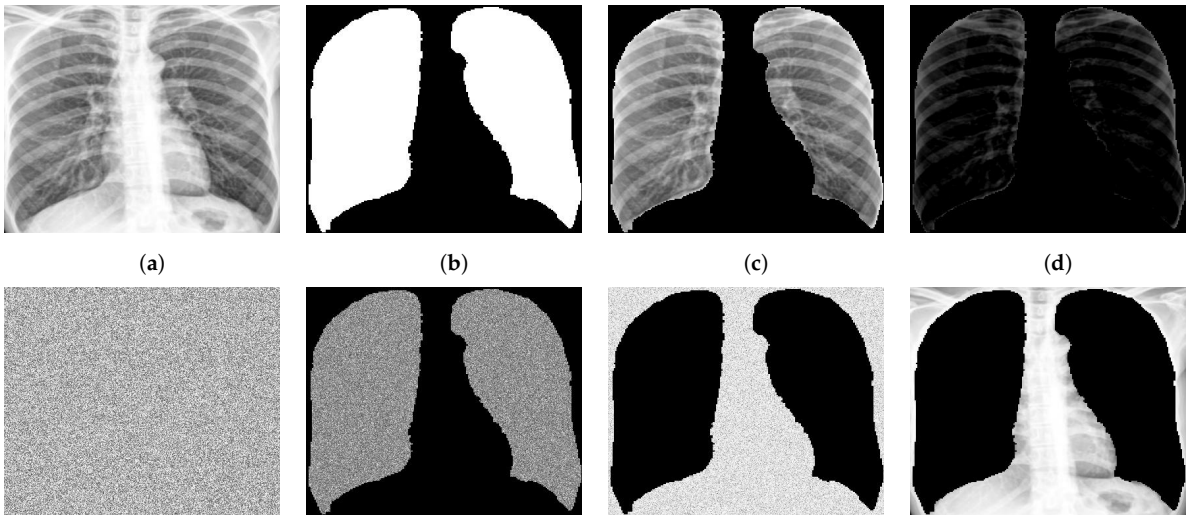


Figure 10. Different ways in which a chest X-ray image is modified to limit the information the classifier relies on. Top (left to right): (**a**) Cropped X-ray image, (**b**) lung mask, (**c**) segmented lung, (**d**) segmented lung with mean of background subtracted. Bottom (left to right): (**e**) Entire image randomized, (**f**) lung pixels randomized with non-lung area set to 0, (**g**) non-lung pixels randomized with lung pixels set to 0, and (**h**) lung pixels set to 0.

Table 6. Performance with input and model limitations.

Model Input	Validation Set	Belarus Dataset
Lung excluded	0.79	0.59
Histogram normalized (lung excluded)	0.74	0.58
Lung mask	0.61	0.55
Randomized pixels (entire image)	0.63	0.50
Randomized pixels (lung only)	0.64	0.50
Randomized pixels (lung excluded)	0.64	0.55
Frozen conv. layers (lung only)	0.66	0.62
Frozen conv. layers (lung excluded)	0.66	0.50
Background normalized (lung only)	0.75	0.61

To limit the network and its capacity, all the convolutional layers of the CNN were frozen (set to become non-trainable). These layers were used as a feature extractor and the fully-connected layers acted as a trainable classifier. When the lung pixel values were set to 0 (lung excluded), the performance on the verification test set (Belarus data) with the CNN network was 59%. Histogram normalization did not yield better results for this. As expected, when we completely randomized the pixel locations of the entire image or of the lung regions the performance was random (50%) on the balanced Belarus dataset. Using the shape of the lungs (lung masks) without the intensity values was enough to improve that performance to 55%. When intensity values of the non-lung regions (background) were provided, the performance further improved to 59%. On the same images, if the background pixels were randomized again, the performance dropped back to 55%. This series of results hints that the shape of the lung itself carries useful information. However, it was interesting that the non-lung regions had a small contribution to the identification of drug resistance. The performance on both datasets improved when the local information of the non-lung regions was retained.

Freezing the convolutional layer weights (ImageNet weights) but allowing the fully-connected layer weights to be trainable, the performance achieved was comparable at 62%. Here, the frozen convolutional layers are acting as fixed feature extractors and the dense layers are learning to interpret them. The approach and the performance were comparable to using the radiomic features. Normalizing by subtracting the mean of the non-lung regions from the lungs also had a similar performance of 61%.

5.5. Country Classification

Being able to identify the origin of the chest X-rays can be insightful in understanding the extent of bias in the data based on the data origin and the acquisition standards within a country. As seen in Table 1, the distribution of samples from different countries is not uniform and neither is the distribution of samples of each class. The chest X-ray images originate from a variety of imaging devices from hospitals in several different countries with their own imaging protocols and other variances that affect image content. These types of artifacts can often be identified by a deep neural network but may not be visible to a human observer. The entire dataset was used for the country classification experiments. Because the number of samples was imbalanced, we used weights based on the number of samples for the categorical cross entropy loss function during our training.

The mean intensities of the input images to the training network were plotted to observe the distribution differences across various countries in Figure 11. The width of each violin plot represents the frequency of the mean intensity. Generally, the wider the violin plot, the higher is the probability that the images of the respective country have the corresponding mean intensity.The histogram equalization centers the mean of the images intensities to zero for each of the countries and helps to reduce some of the bias present in the dataset due to the images' country of origin. The variance in the intensity distribution is still present.

The multi-class country-of-origin classification from the X-ray images achieves an accuracy of 85.7%. When histogram equalization is applied to the same images to remove some of the intensity-based biases, performance decreased to 82.6%. These results show that the models are very efficient in deriving the country of origin from the chest X-ray images. Even with histogram equalization, the country classification performance did not decrease sharply. This points to other biases in the data (potentially other acquisition biases) that are not accessible like the country of origin.

We also performed a multi-class country-of-origin classification based on clinical text data. Demographic features such as gender, age, and education, were included along with radiological findings from chest X-rays (such as nodules, cavities, infiltrates, collapses, etc.). The multi-class country-of-origin classification with seven demographic features and 20 radiological features resulted in an accuracy of 59.3%, and the classification with just 20 radiological features showed an accuracy of 35.2%. The confusion matrices in Figure 12

show that the image-derived features have many fewer false predictions when classifying the country of origin compared to the model trained on clinical features. The performance of this classifier is based on clinical findings whereas the deep learning classifier is using the image content which potentially includes information not related to the disease and not visible to the human observer but detectable by the network, allowing it to obtain better performance. Consequently, the DR/DS classifier also has access to this non-disease specific information which may introduce confounding features into that model, improving its performance on the trained domain but harming its generalizability.

5.6. Tuberculosis Abnormality Segmentation

Transfer learning with pretrained weights has been effectively used not only to reduce training time compared to random initialization but also to obtain a better performance. Intuitively, it makes sense to use TB abnormalities as priors to the drug resistance classification as well. Hence, we utilize the TB abnormality segmentation network to aide the classification of drug resistance, using its weights and output for the multi-task classification. We use the Shenzhen dataset with TB lung lesions [30], which has two sets of annotations for the same image.

The average segmentation overlap between the two sets of annotations was 0.538 (Dice score). For the TB abnormality segmentation network, the cross validation Dice score was 0.636. The weights for this network are used to initialize the classification and multi-task models. For the classification tasks, only the encoder weights were used, whereas for the segmentation tasks all the convolutional layer weights were used.

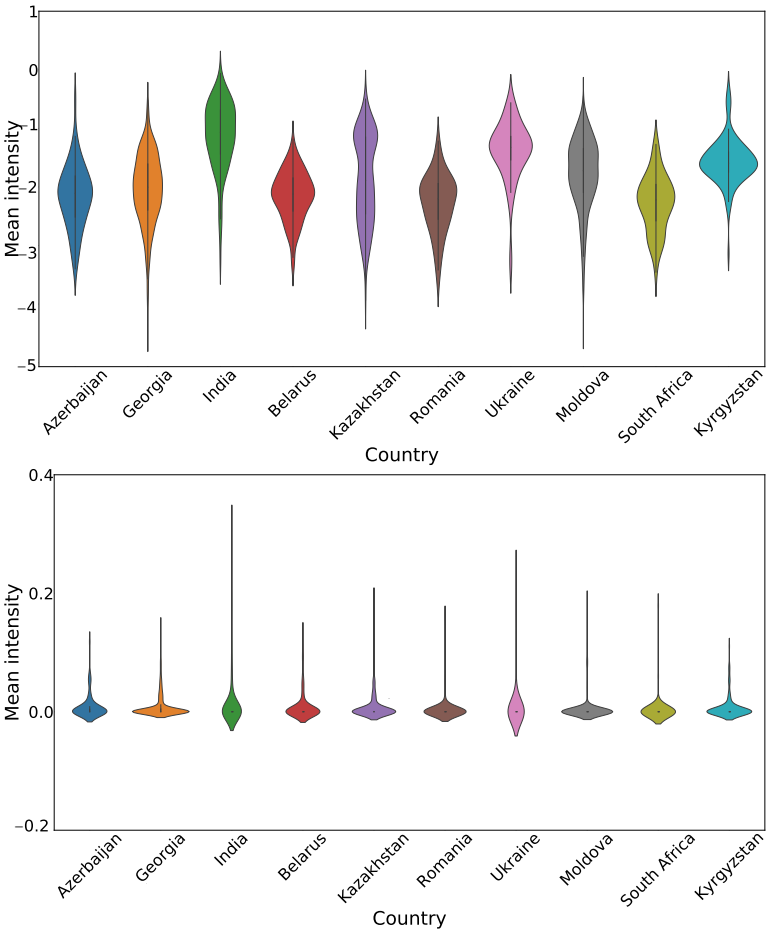


Figure 11. Mean intensity distributions per country for the normalized cropped X-ray images (**top**) and the same images after histogram equalization is applied (**bottom**).

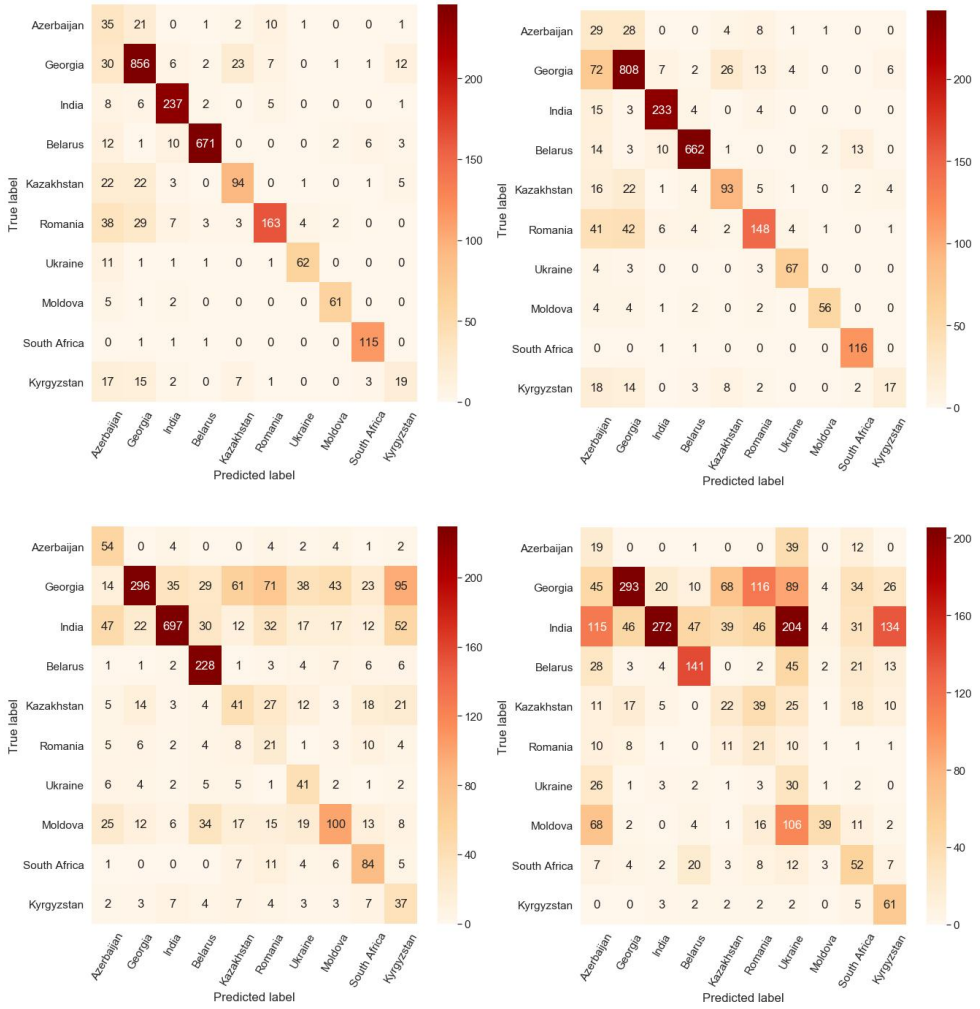


Figure 12. Summation of confusion matrices on test sets from 5 folds on the country-of-origin classification task. (**top-left**) Lung-segmented images, (**top-right**) histogram-equalized lung-segmented images, (**bottom-left**) radiological and demographic features and (**bottom-right**) radiological features. Note that the image-derived features have a significantly better performance than the disease-specific radiological findings.

Table 7 shows that the annotations by radiologists matched with the abnormalities identified by the TB abnormality segmentation model in each of the sextants and overall. The average GradCAM [41] map values for each of the sextants was poorly correlated with sextant annotations. Abnormality predictions were also derived from the mean of the GradCAM heatmaps (with a threshold of 0.7) and compared with the abnormalities from the TB segmentation model. On both of these, the top sextants had very little overlap compared with the middle and bottom sextants. This result is contrary to expectation as TB abnormalities are often seen at the top of the lung [42].

Figure 13 shows some examples of DR-TB and DS-TB. GradCAM heatmaps show the average of activations from the last layer of a CNN. In these examples, it can be seen that for the DR-TB class, activations are more frequent. This is consistent with previous results [9], where DR-TB patients are shown to have more abnormalities in their lungs.

Table 7. Comparison (AUC) of localized abnormalities from different sources tested on Belarus dataset.

Ground Truth:	**Sextant Annotations**	**Sextant Annotations**	**GradCAM (Predictions)**
Prediction:	**TB Abnormal Seg. Prob. Map**	**GradCAM (Heatmap)**	**TB Abnormal Seg. Prob. Map**
Top Right	0.788	0.469	0.518
Top Left	0.742	0.504	0.567
Middle Right	0.759	0.627	0.692
Middle Left	0.759	0.630	0.716
Bottom Right	0.808	0.642	0.699
Bottom Left	0.758	0.571	0.665
Overall	0.769	0.574	0.643

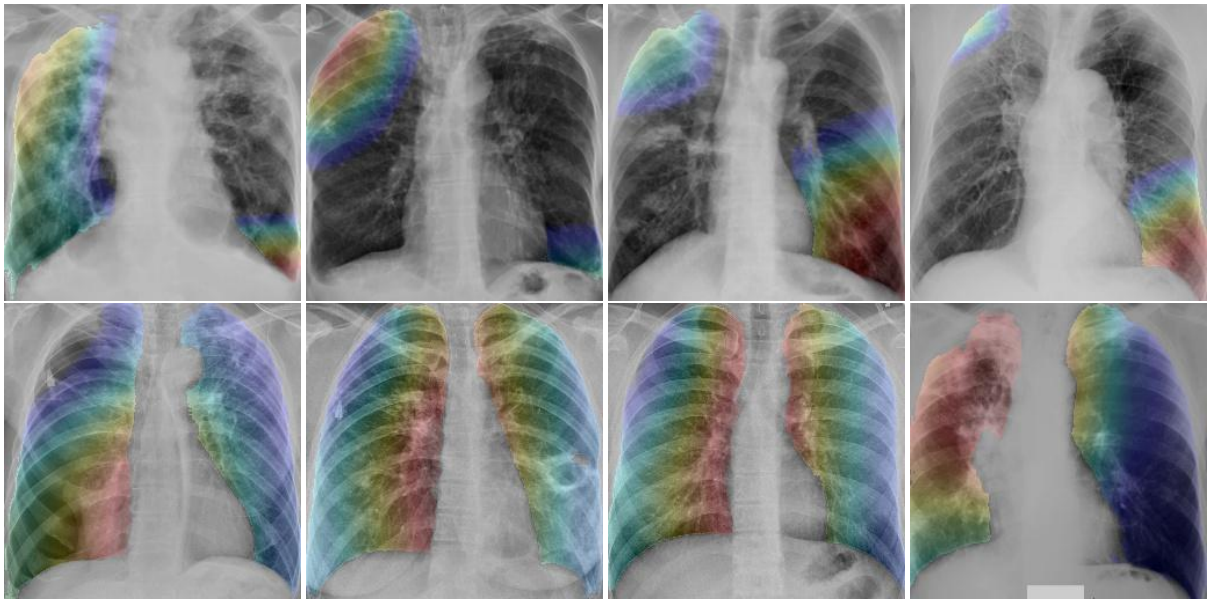


Figure 13. GradCAM visualizations on X-ray images. **Top row** shows examples of drug-sensitive TB and **bottom row** shows examples of drug-resistant TB.

5.7. Multi-Task Based Classification

As an approach to assist the drug resistance classifier, a secondary task was added to further incentivize the network to focus on relevant areas. The second task acts as a regulator to constraint the neural network to focus on the interesting regions. ResNet18 was used as the primary backbone for the networks and layers were added to generate output for the secondary tasks. Instead of experimentally determining the best loss weights for the model, we allow them to be learnt by the model itself.

Initially, both tasks were given equal weights to allow the network to determine which task needs to be focused. To restrain the scope of the work, while we monitored the performance on the secondary class, we only used the performance on the main task to determine our experimental setup and hence we report the performance on the main task only.

While adding a related secondary task did not improve the drug resistance classification on the validation set, the AUC performance on the Belarus dataset improved by about 2%–3% and accuracy improved by 1%. An AUC of 68% (±1%) was the best performance achieved on the Belarus dataset as shown in Table 8.

Table 8. Classification performance with an additional task.

		Validation Set		Belarus Dataset	
Trained on	**Secondary Task**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
Sext. Dataset	Abnormal Sextant Classification	0.77 ± 0.02	**0.70 ± 0.02**	0.64 ± 0.01	0.61 ± 0.03
	Abnormal Sextant Segmentation	**0.78 ± 0.02**	**0.70 ± 0.02**	0.67 ± 0.01	**0.63 ± 0.01**
Gen. Dataset	TB Abnormalities Segmentation	0.77 ± 0.02	0.69 ± 0.02	**0.68 ± 0.01**	**0.63 ± 0.02**

6. Conclusions

This paper explores the cross validation and generalization performance achieved for drug-sensitive and drug-resistant classification on chest X-rays from different countries. By excluding data from one country of origin from training and using it for testing, we evaluate classifier performance on unseen data. The generalization performance was much lower (65% AUC) compared to the cross validation performance (79% AUC). The same CNN architecture was able to classify the country of origin from a chest X-ray image. Evaluations with radiomic features from X-ray images, and experimental limitations to the data and classifier, indicated that the model based its decisions on other artifacts present in the images. TB lesions annotated by radiologists were utilized to see if the location information was useful for discriminating between drug-resistant and drug-sensitive cases. While GradCAM heatmaps from the X-ray image-based CNN model did not overlap significantly with the TB lesions and the annotations from radiologists, adding a secondary task related to the localization of lesions did improve the classification performance to 68% AUC. Because of an imbalanced dataset, insufficient amount of samples of one of the two classes, and the lack of clinical text data describing the radiological findings for all the patients, we only excluded one single country for our generalization evaluation. A solution that does not require annotations by radiologists to improve the generalization performance would be more valuable. Procedures and methods that allow the model to pick up only the manifestations of disease are a direction for future research. In general, we believe that experiments addressing generalization to new datasets should be standard practice in medical image analysis with deep learning.

Author Contributions: Conceptualization, Z.Y. and S.J.; methodology, M.K., K.K., F.Y. and H.Y.; software, M.K., K.K., F.Y. and H.Y.; validation, M.K., K.K. and Y.X.J.W.; formal analysis, M.K., K.K., F.Y. and H.Y.; investigation, M.K., K.K. and F.Y.; resources, Y.X.J.W. and Z.Y.; data curation, Y.X.J.W. and Z.Y.; writing—original draft preparation, M.K. and K.K.; writing—review and editing, F.Y., Z.Y. and S.J.; visualization, M.K. and H.Y.; supervision, Z.Y. and S.J.; project administration, S.J.; funding acquisition, S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Data usage is exempt from local institutional review board review as it is publicly available from the TB portals program. The TB portals program participants are responsible for ensuring compliance with their countries laws, regulations, and ethics considerations.

Data Availability Statement: Links to datasets used in this study are provided in Section 3.

Acknowledgments: 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, National Institutes of Health. This project has also been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases under BCBB Support Services Contract HHSN316201300006W/HHSN27200002.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. *Global Tuberculosis Report* ; World Health Organization: Geneva, Switzerland, 2020; p. xiii.
2. Qin, Z.Z.; Sander, M.S.; Rai, B.; Titahong, C.N.; Sudrungrot, S.; Laah, S.N.; Adhikari, L.M.; Carter, E.J.; Puri, L.; Codlin, A.J.; et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci. Rep.* **2019**, *9*, 15000. [CrossRef]
3. Qin, Z.Z.; Ahmed, S.; Sarker, M.S.; Paul, K.; Adel, A.S.S.; Naheyan, T.; Barrett, R.; Banu, S.; Creswell, J. Tuberculosis detection from chest X-rays for triaging in a high tuberculosis-burden setting: An evaluation of five artificial intelligence algorithms. *Lancet Digit. Health* **2021**, *3*, e543–e554. [CrossRef]
4. Wàng, Y.X.J.; Chung, M.J.; Skrahin, A.; Rosenthal, A.; Gabrielian, A.; Tartakovsky, M. Radiological signs associated with pulmonary multi-drug resistant tuberculosis: An analysis of published evidences. *Quant. Imaging Med. Surg.* **2018**, *8*, 161–173. [CrossRef]
5. Icksan, A.G.; Napitupulu, M.R.S.; Nawas, M.A.; Nurwidya, F. Chest X-ray findings comparison between multi-drug-resistant tuberculosis and drug-sensitive tuberculosis. *J. Nat. Sci. Biol. Med.* **2018**, *9*, 42.
6. Huang, X.L.; Skrahin, A.; Lu, P.X.; Alexandru, S.; Crudu, V.; Astrovko, A.; Skrahina, A.; Taaffe, J.; Harris, M.; Long, A.; et al. Prediction of multiple drug resistant pulmonary tuberculosis against drug sensitive pulmonary tuberculosis by CT nodular consolidation sign. *bioRxiv* **2019**. [CrossRef]
7. Flores-Trevino, S.; Rodriguez-Noriega, E.; Garza-Gonzalez, E.; Gonzalez-Diaz, E.; Esparza-Ahumada, S.; Escobedo-Sanchez, R.; Perez-Gomez, H.R.; Leon-Garnica, G.; Morfin-Otero, R. Clinical predictors of drug-resistant tuberculosis in Mexico. *PLoS ONE* **2019**, *14*, e0220946.
8. Cheng, N.; Wu, S.; Luo, X.; Xu, C.; Lou, Q.; Zhu, J.; You, L.; Li, B. A Comparative Study of Chest Computed Tomography Findings: 1030 Cases of Drug-Sensitive Tuberculosis versus 516 Cases of Drug-Resistant Tuberculosis. *Infect. Drug Resist.* **2021**, *14*, 1115–1128. [CrossRef]
9. Yang, F.; Yu, H.; Kantipudi, K.; Karki, M.; Kassim, Y.M.; Rosenthal, A.; Hurt, D.E.; Yaniv, Z.; Jaeger, S. Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features. *Quant. Imaging Med. Surg.* **2022**, *12*, 675–687. [CrossRef]
10. Ionescu, B.; Müller, H.; Villegas, M.; de Herrera, A.G.S.; Eickhoff, C.; Andrearczyk, V.; Cid, Y.D.; Liauchuk, V.; Kovalev, V.; Hasan, S.A.; et al. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 309–334.
11. Gentili, A. ImageCLEF2018: Transfer Learning for Deep Learning with CNN for Tuberculosis Classification. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
12. Ishay, A.; Marques, O. ImageCLEF 2018 Tuberculosis Task: Ensemble of 3D CNNs with Multiple Inputs for Tuberculosis Type Classification. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
13. Cid, Y.D.; Müller, H. Texture-based Graph Model of the Lungs for Drug Resistance Detection, Tuberculosis Type Classification, and Severity Scoring: Participation in ImageCLEF 2018 Tuberculosis Task. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
14. Allaouzi, I.; Ahmed, M.B. A 3D-CNN and SVM for Multi-Drug Resistance Detection. In *CLEF (Working Notes), Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 11–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
15. Ureta, J.; Shrestha, A. Identifying drug-resistant tuberculosis from chest X-ray images using a simple convolutional neural network. *J. Phys. Conf. Ser.* **2021**, *2071*, 012001. [CrossRef]
16. Jaeger, S.; Juarez-Espinosa, O.H.; Candemir, S.; Poostchi, M.; Yang, F.; Kim, L.; Ding, M.; Folio, L.R.; Antani, S.; Gabrielian, A.; et al. Detecting drug-resistant tuberculosis in chest radiographs. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1915–1925. [CrossRef]
17. Karki, M.; Kantipudi, K.; Yu, H.; Yang, F.; Kassim, Y.M.; Yaniv, Z.; Jaeger, S. Identifying Drug-Resistant Tuberculosis in Chest Radiographs: Evaluation of CNN Architectures and Training Strategies. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; IEEE: Piscataway, NJ, USA, 2021.
18. Pooch, E.H.; Ballester, P.L.; Barros, R.C. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv* **2019**, arXiv:1909.01940.
19. Castro, D.C.; Walker, I.; Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **2020**, *11*, 3673. [CrossRef] [PubMed]
20. Harris, M.; Qi, A.; Jeagal, L.; Torabi, N.; Menzies, D.; Korobitsyn, A.; Pai, M.; Nathavitharana, R.R.; Ahmad Khan, F. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest X-rays for pulmonary tuberculosis. *PLoS ONE* **2019**, *14*, e0221339. [CrossRef] [PubMed]
21. Sathitratanacheewin, S.; Sunanta, P.; Pongpirul, K. Deep learning for automated classification of tuberculosis-related chest X-Ray: Dataset distribution shift limits diagnostic performance generalizability. *Heliyon* **2020**, *6*, e04614. [CrossRef]

22. Rajpurkar, P.; Joshi, A.; Pareek, A.; Chen, P.; Kiani, A.; Irvin, J.; Ng, A.Y.; Lungren, M.P. CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. *arXiv* **2020**, arXiv:2002.11379.
23. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **2018**, *15*, e1002683. [CrossRef]
24. Badgeley, M.A.; Zech, J.R.; Oakden-Rayner, L.; Glicksberg, B.S.; Liu, M.; Gale, W.; McConnell, M.V.; Percha, B.; Snyder, T.M.; Dudley, J.T. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2019**, *2*, 31. [CrossRef]
25. Ahmed, K.B.; Goldgof, G.M.; Paul, R.; Goldgof, D.B.; Hall, L.O. Discovery of a Generalization Gap of Convolutional Neural Networks on COVID-19 X-Rays Classification. *IEEE Access* **2021**, *9*, 72970–72979. [CrossRef]
26. Rosenthal, A.; Gabrielian, A.; Engle, E.; Hurt, D.E.; Alexandru, S.; Crudu, V.; Sergueev, E.; Kirichenko, V.; Lapitskii, V.; Snezhko, E.; et al. The TB portals: An open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis. *J. Clin. Microbiol.* **2017**, *55*, 3267–3282. [CrossRef] [PubMed]
27. Dodd, P.J.; Looker, C.; Plumb, I.D.; Bond, V.; Schaap, A.; Shanaube, K.; Muyoyeta, M.; Vynnycky, E.; Godfrey-Faussett, P.; Corbett, E.L.; et al. Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. *Am. J. Epidemiol.* **2015**, *183*, 156–166.
28. Yates, T.A.; Atkinson, S.H. Ironing out sex differences in tuberculosis prevalence. *Int. J. Tuberc. Lung Dis.* **2017**, *21*, 483–484. [CrossRef]
29. Hertz, D.; Schneider, B. Sex differences in tuberculosis. *Semin. Immunopathol.* **2019**, *41*, 225–237. [CrossRef]
30. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.J.; Lu, P.X.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475.
31. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.I.; Matsui, M.; Fujita, H.; Kodera, Y.; Doi, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR Am. J. Roentgenol.* **2000**, *174*, 71–74. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Williams, M.B.; Krupinski, E.A.; Strauss, K.J.; Breeden, W.K., III; Rzeszotarski, M.S.; Applegate, K.; Wyatt, M.; Bjork, S.; Seibert, J.A. Digital radiography image quality: Image acquisition. *J. Am. Coll. Radiol.* **2007**, *4*, 371–388. [CrossRef] [PubMed]
35. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]
36. Argyriou, A.; Evgeniou, T.; Pontil, M. Convex multi-task feature learning. *Mach. Learn.* **2008**, *73*, 243–272. [CrossRef]
37. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.
38. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
39. Kim, H.Y.; Song, K.S.; Goo, J.M.; Lee, J.S.; Lee, K.S.; Lim, T.H. Thoracic sequelae and complications of tuberculosis. *Radiographics* **2001**, *21*, 839–858. [CrossRef] [PubMed]
40. Nachiappan, A.C.; Rahbar, K.; Shi, X.; Guy, E.S.; Mortani Barbosa, E.J., Jr.; Shroff, G.S.; Ocazionez, D.; Schlesinger, A.E.; Katz, S.I.; Hammer, M.M. Pulmonary tuberculosis: Role of radiology in diagnosis and management. *Radiographics* **2017**, *37*, 52–72. [CrossRef] [PubMed]
41. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
42. Bennett, J.E.; Dolin, R.; Blaser, M.J. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases E-Book*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2019.

Journal of X-Ray Science and Technology 28 (2020) 939–951
DOI 10.3233/XST-200662
IOS Press

Developing and verifying automatic detection of active pulmonary tuberculosis from multi-slice spiral CT images based on deep learning

Luyao Ma[a,b,1], Yun Wang[a,b,2], Lin Guo[c,*], Yu Zhang[a], Ping Wang[a,b], Xu Pei[a,b], Lingjun Qian[c], Stefan Jaeger[d], Xiaowen Ke[c], Xiaoping Yin[a,*] and Fleming Y.M. Lure[c,e]
[a]*CT-MRI Room, Affiliated Hospital of Hebei University, Baoding, Hebei, China*
[b]*Clinical Medical College, Hebei University, Baoding, Hebei, China*
[c]*Shenzhen Zhiying Medical Imaging, Shenzhen, Guangdong, China*
[d]*National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*
[e]*MS Technologies Corp, Rockville, MD, USA*

Received 7 February 2020
Revised 28 May 2020
Accepted 15 June 2020

Abstract.
OBJECTIVE: Diagnosis of tuberculosis (TB) in multi-slice spiral computed tomography (CT) images is a difficult task in many TB prevalent locations in which experienced radiologists are lacking. To address this difficulty, we develop an automated detection system based on artificial intelligence (AI) in this study to simplify the diagnostic process of active tuberculosis (ATB) and improve the diagnostic accuracy using CT images.
DATA: A CT image dataset of 846 patients is retrospectively collected from a large teaching hospital. The gold standard for ATB patients is sputum smear, and the gold standard for normal and pneumonia patients is the CT report result. The dataset is divided into independent training and testing data subsets. The training data contains 337 ATB, 110 pneumonia, and 120 normal cases, while the testing data contains 139 ATB, 40 pneumonia, and 100 normal cases, respectively.
METHODS: A U-Net deep learning algorithm was applied for automatic detection and segmentation of ATB lesions. Image processing methods are then applied to CT layers diagnosed as ATB lesions by U-Net, which can detect potentially misdiagnosed layers, and can turn 2D ATB lesions into 3D lesions based on consecutive U-Net annotations. Finally, independent test data is used to evaluate the performance of the developed AI tool.
RESULTS: For an independent test, the AI tool yields an AUC value of 0.980. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value are 0.968, 0.964, 0.971, 0.971, and 0.964, respectively, which shows that the AI tool performs well for detection of ATB and differential diagnosis of non-ATB (i.e. pneumonia and normal cases).
CONCLUSION: An AI tool for automatic detection of ATB in chest CT is successfully developed in this study. The AI tool can accurately detect ATB patients, and distinguish between ATB and non- ATB cases, which simplifies the diagnosis process and lays a solid foundation for the next step of AI in CT diagnosis of ATB in clinical application.

Keywords: Active tuberculosis (ATB), artificial intelligence (AI), deep learning

[1]First author: Luyao Ma.

[2]Joint first author: Yun Wang.

*Corresponding authors: Xiaoping Yin, E-mail: yinxiaoping78@sina.com. and Lin Guo, E-mail: guolin913@outlook.com.

1. Introduction

Worldwide, tuberculosis (TB) is one of the top 10 causes of death. Tuberculosis is also one of the most common causes of opportunistic infections and a major cause of death among people living with Human Immunodeficiency Virus (HIV). In 2017, tuberculosis killed approximately 1.3 million HIV-negative people, and about 300,000 HIV-positive people died from TB. According to the best estimate, there were about 10 million people newly infected with tuberculosis in the world in 2017 [1]. In order to reduce the global burden of this disease, the World Health Organization recommends screening for active tuberculosis (ATB) among high-risk groups for early detection and timely treatment. For the high-risk group, the incidence rate of tuberculosis is significantly higher than for the general population [2]. X-ray examination is one of the auxiliary diagnostic tests for tuberculosis screening, but there is a rate of missed cases in practice. With the popularization of computer tomography (CT) technology in small and medium-sized hospitals, CT has become very helpful in identifying chest parenchymal lesions and detecting the severity of lung disease in tuberculosis patients [3, 4]. CT images can better display the characteristics of ATB, including cavities, parenchymal abnormalities, lobular central nodules, and tree bud signs [5]. However, the diagnosis of tuberculosis in CT requires a doctor with a certain diagnostic ability, which is a difficult task for most TB prevalent locations in which experienced radiologists are often lacking [6, 7]. Consequently, the screening efficiency is quite low, which may impair timely treatment of ATB patients.

Therefore, there has been an interest in applying artificial intelligence (AI) to the detection of ATB in CT, which can help improve the screening efficiency. So far, various algorithms have been proposed based on deep learning to automatically diagnose diseases in the medical field [8]. In the Image Net Large Scale Visual Recognition Competition in 2012, deep learning techniques have achieved great success, especially in the classification of medical images [7]. In 2018, Hwang et al. developed an automatic detection software for ATB based on chest X-ray images, which showed better performance than most doctors including chest radiologists [9]. Moreover, the use of deep learning to diagnose lung nodules and pneumonia in images has demonstrated advantages and application values [10, 11]. Thus, the recent success and promising results provide great encouragement for future clinical applications of AI in assisting diagnosis of diseases using in medical imaging.

The purpose of this paper is to simplify the ATB diagnosis process and improve the diagnostic accuracy based on an automated deep learning, artificial intelligence (AI) detection system for multi-slice spiral CT images. It also improves the detection rate of early ATB for remote areas and reduces the workload of radiologists. The structure of this paper is as follows: Section II mainly introduces the data and methods, including the collection and arrangement of data, and the deep learning techniques we used for developing the automatic detection model. Section III presents the final test results and the measures for evaluating the performance of the algorithm. Section IV analyzes and discusses the results, and describes the advantages and disadvantages of the software. Section V describes the conclusions drawn from this paper. Finally, Section VI describes the outlook for the future.

2. Data and methods

2.1. Data

CT images of 868 patients (male 534, female 334; mean $\pm$ standard deviation of age: 47 ± 20 years old), from April 2016 to May 2019, in the affiliated Hospital of Hebei University of China were retrospectively collected. Each patient had a corresponding electronic image report for CT, and each ATB patient also had a corresponding tuberculosis sputum smear report. These patients were divided into 498 ATB (22 sputum smear were found negative, 476 sputum smear were found postive), 220

Table 1
Training data

Study	Total	Sex		Average±standard	Maximum	Minimum	Total number
	(case)	Male	Female	deviation age	age	age	of CT slices
ATB	337	237	100	44 ± 20	89	12	21460
Normal	120	60	60	42 ± 14	68	5	7410
Pneumonia	110	63	47	63 ± 16	89	14	6679

Table 2
Testing data

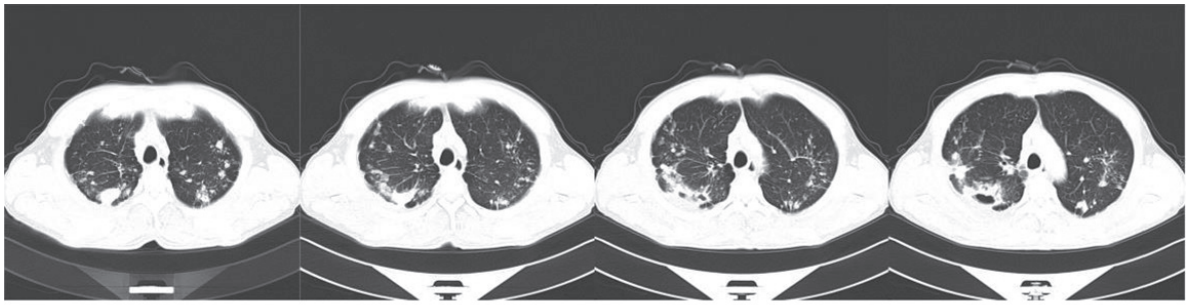
Study	Total	Sex		Average±standard	Maximum	Minimum	Total number
	(case)	Male	Female	deviation age	age	age	of CT slices
ATB	139	92	47	43 ± 20	85	7	8908
Normal	100	48	52	43 ± 16	83	2	2293
Pneumonia	40	18	22	62 ± 16	86	18	6109

normal and 150 pneumonia cases. In this study, ATB cases were diagnosed using positive tuberculosis smear as the gold standard, and normal and pneumonia cases were diagnosed using the radiologists' reports as the gold standard. Then, all the cases were divided into a training data set and a testing data set, including 337 ATB, 120 normal, and 110 pneumonia cases as the training data set to optimize network weights; and 139 ATB, 100 normal, and 40 pneumonia cases as the testing data set to test the performance of the algorithm. The training and the testing data sets are independent, with details shown in Tables 1 and 2. Because of the similar imaging findings of pneumonia and ATB in clinical diagnosis [12, 13], patients with pneumonia were enrolled in this study.

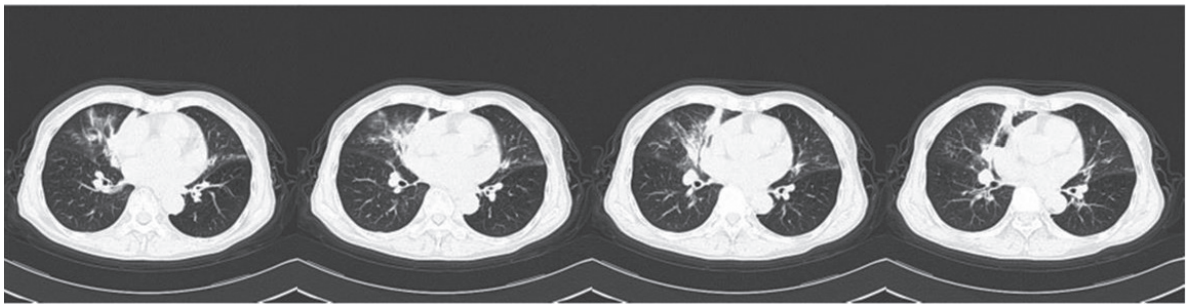
The main CT manifestations of ATB and non-tuberculous pneumonia are as follows: The CT appearance of ATB is mainly characterized by multiple lung lobes, multiple lung segments, and multiple morphologies, which are more likely to occur in the posterior segment of the upper lobes of the lungs and the dorsal segment of the lower lobes. The morphology is mainly manifested as flakes or tree bud signs, often accompanied by cord strip shadows, calcifications, voids, etc [14, 15] as shown in Fig. 1a. Pneumonia is often confined to one lobe of the lungs. In CT, it is mainly manifested as smears or ground glass shadows with blurred edges, uniform density, and air bronchial signs are often seen in the lesions [16, 17] as shown in Fig. 1b.

In the development of accurate radiology deep learning algorithms, in addition to the appropriate model structure, a large number of precise markers are needed to train the algorithm [18]. Therefore, before developing a deep-learning-based automatic detection (DLAD) algorithm, this paper uses LabelImg software to mark the lesions of 337 ATB patients in the training data. A total of 8213 CT slices are marked, with a total of 13679 labels and an average of 5917 pixels per label. Two radiologists negotiated to develop a unified standard for these markings (The standard is to circle the ATB lesions in each CT slice with a box using LabelImg software. The same focus can be marked with multiple boxes, which should not contain normal lung tissue and should exceed the abnormal lung area as little as possible). All ATB, normal and pneumonia CT images were examined by five radiologists, with more than 5 years of experience in CT reading, in the affiliated Hospital of Hebei University (Class III Class A Hospital), China.

In this study, the CT images were scanned using Siemens (SOMATOM Definition AS) 64-row 128-slice spiral CT, and 40-row 64-slice spiral CT. Before spiral CT scanning, a patient was trained in



(a) ATB cases



(b) Non-tuberculous pneumonia cases

Fig. 1. Typical CT findings of (a) ATB and (b) non-tuberculous pneumonia.

breathing. When performing a spiral CT scan, the patient takes the supine position, raises the arms and advances the head. The patient breathes in and finishes the whole lung scan within 5–8s. The scan ranges from the chest entrance to the base of the lung, in spiral scanning mode. The tube voltage is 100 kV; tube current: 100 mA; pitch: 1.3; slice thickness: 5.0 mm; field of view: 430 mm.

2.2. Method

The algorithm was primarily developed by Shenzhen Smart Imaging Healthcare Co., Ltd and CT/MRI Department, Affiliated Hospital of Hebei University in collaboration with several US and Chinese institutions listed in the affiliation of authors. In this study, CT images of ATB were used as positive samples, and CT images of pneumonia and normal patients were used as negative samples for training and testing.

The primary architecture consists of a U-Net deep learning based CNN model to process each slice of an entire CT scan to identify 2-D regions of interest (ROI) in each slice. The architecture also involves a clustering technique based on the connectivity of an ROI to convert multiple ROIs in 2-D slices into a single 3-D object of interest (OOI) in a 3-D scan, as shown in Fig. 2.

The training data process is as follows: First, the ATB lesions were labeled with rectangular boxes centered at the locations of tuberculosis findings by a board-certified radiologist based on the radiology reports, using LabelImg software developed by the authors of this paper. The training and testing data do not overlap. Second, using a U-Net segmentation network to perform image segmentation training on labeled ATB regions in the training data, the trained U-Net for automatic identification of ATB lesions was obtained. Third, when diagnosing ATB patients, image processing methods have been added that can effectively convert 2-D ROI at each slice into 3-D OOI for a CT scan. Finally, combining the steps above, the AI detection software was obtained. In order to verify the performance of the AI detection software, the testing data is used for detection evaluation.

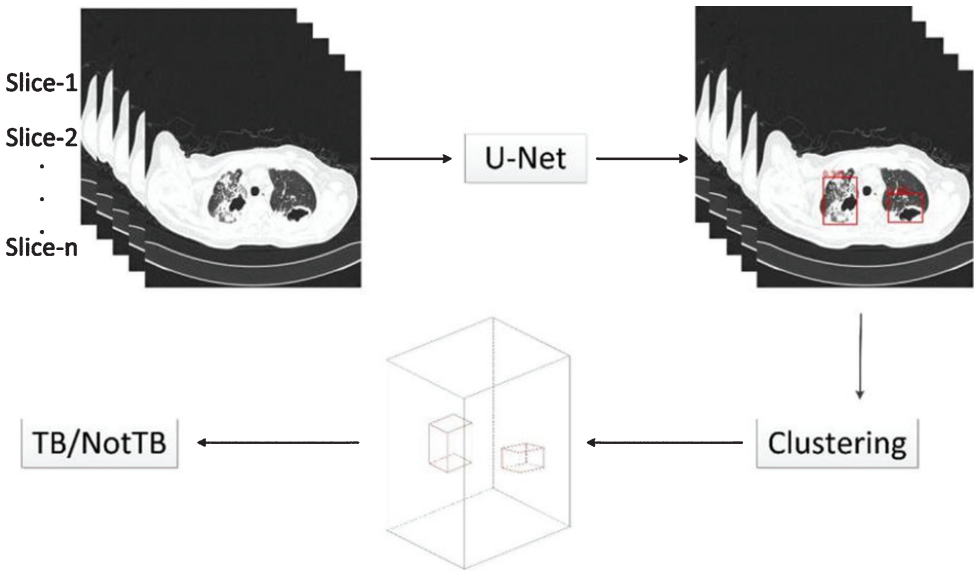


Fig. 2. Main framework of AI detection software.

2.2.1. U-Net deep learning technique to detect 2-D ROI at each CT slice

The U-Net neural network in this paper is used to predict the ATB lesions in each CT image, as shown in Fig. 3. U-Net is an image segmentation network based on a Convolutional Neural Network (CNN) for extracting image features. The basic component of U-Net is the same as the traditional convolutional network in that it performs convolution, batch normalization, activation, and maximum pooling operations in sequence [19]. However, there is a big difference between the U-Net network structure and the traditional CNN network structure. The first half adopts a downsampling pooling operation, so that the resolution of each feature map is reduced to 1/4 of the original size, and the second half adopts upsampling and merging operations, respectively improving the resolution of the feature map, and the fusion of shallow and deep information makes up for the loss of characteristic information caused by the introduction of pooling operations.

2.2.2. Clustering technique to detect 3-D OOI for an entire CT scan

After the U-Net segmentation network is used to predict lesion areas (ROI) of ATB in the testing data, the next step is to convert these 2-D ROI into 3-D OOI by combining CT slices with ATB lesions. If the contour of an ATB lesion slice is detected, and the center points of adjacent bounding boxes are within 30 voxels, the bounding boxes were fused into one, avoiding the interference of tiny regions. Then, the CT slices of ATB lesions at the joint with Intersection Over Union (IOU) >0.3 of adjacent ATB lesion slices were selected, which means the overlap of ATB lesions in adjacent CT layers is greater than 30%. Patients who meet the above conditions, and who have more than four consecutive ATB lesion slices, are classified as ATB patients, as shown in Fig. 4.

2.3. Statistical method

We use our AI software to process the testing data, and plot the ROC curve achieved by our model. To evaluate our software, we compute the AUC value, accuracy, Sensitivity (SS), Specificity (SP), Positive Predictive Value (PPV), and Negative Predictive Value (NPV).

In order to generate the ROC curve and AUC from our segmentation network, we innovatively propose a new method in this study. For each patient, the predictions at each pixel of an input CT slice (Fig. 5a) were computed with a confidence score, which can be displayed as a gray level within the

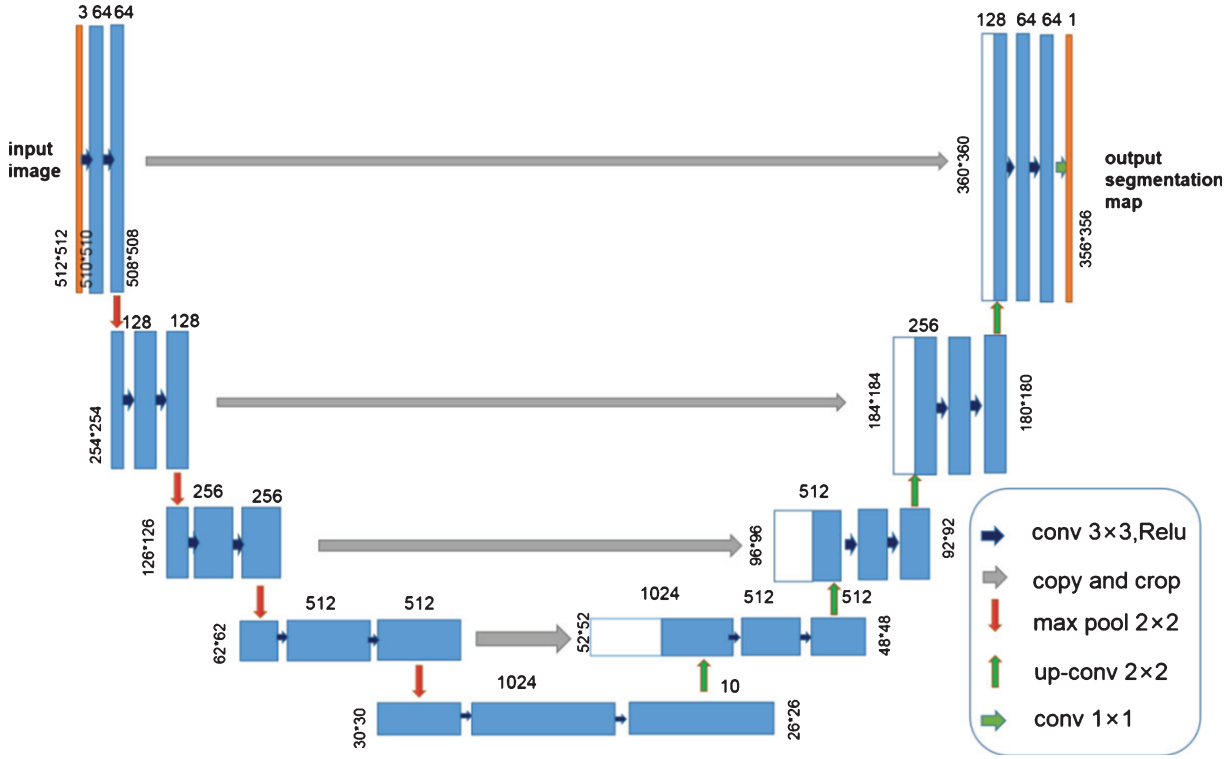


Fig. 3. Architecture of our adapted U-Net network.

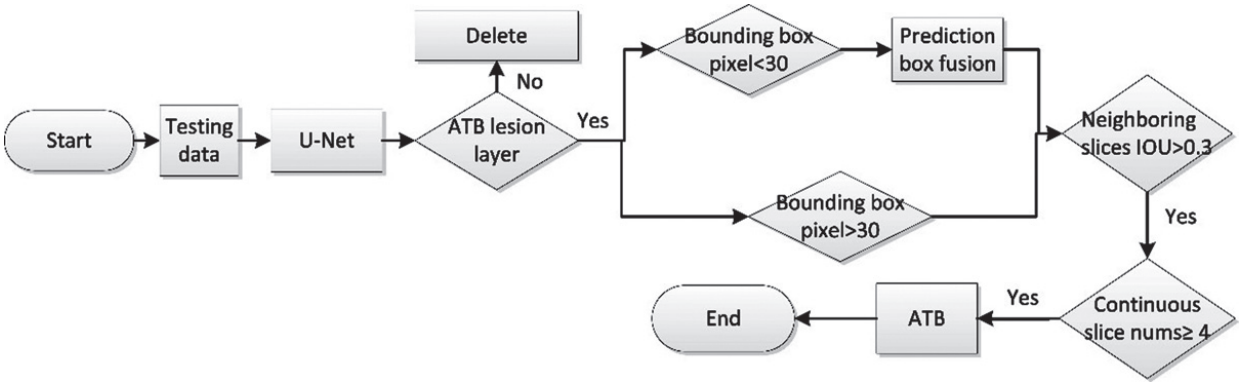


Fig. 4. AI detection software - diagnostic flowchart.

range of [0,255]. The threshold value was set to 127. Every input CT slice of the patient was then preprocessed by normalizing pixel values from the range [0, 255] to [0, 1], with a threshold value of 0.5, changing the gray image (Fig. 5b) into a binary image (Fig. 5c). The degree of confidence in a tuberculosis lesion was obtained by averaging the confidence scores for every pixel within the identified region. Therefore, each patient possesses a degree of confidence, and the patient-level predictions were then obtained to generate the ROC curve.

3. Result

Applying our AI detection software to the testing data of 279 cases, the number of correctly detected ATB cases was 134, the number of missed positive ATB diagnoses was 5, the number of correctly

(a) (b) (c)

Fig. 5. Example images for generating the ROC curve: (a) Input CT image of a patient; (b) gray image; (c) binary image.

Table 3
Software performance on testing data

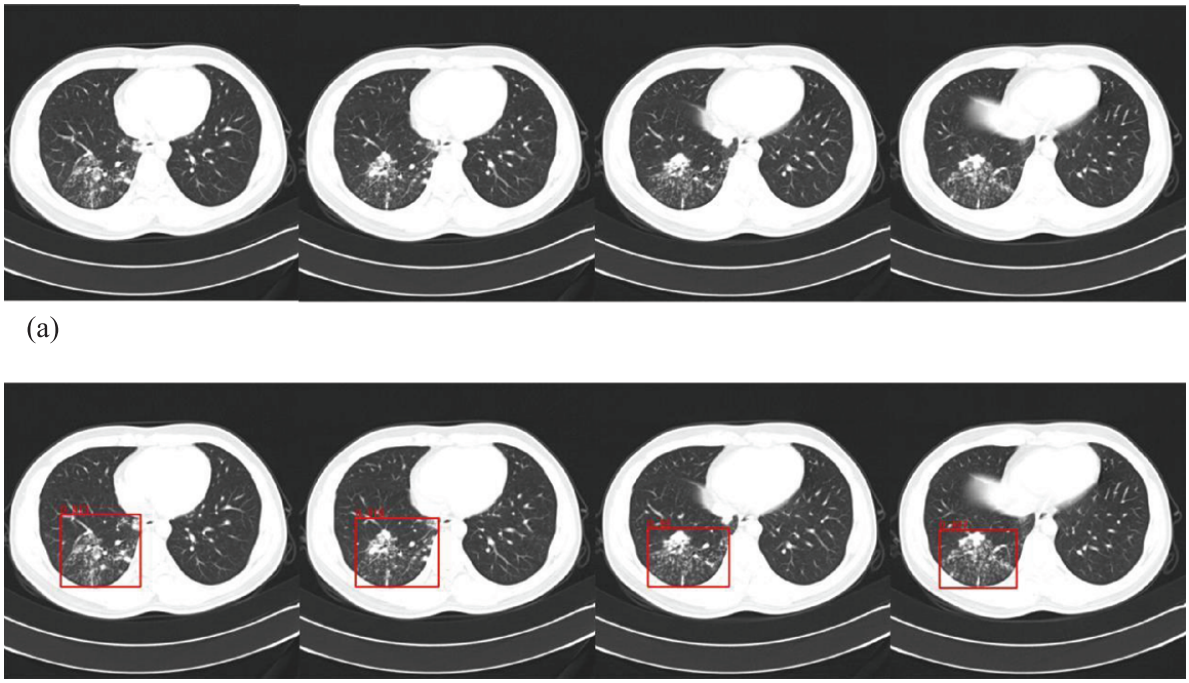
		True value			
		ATB (case)+	Not ATB (case) –		Total (case)
			Pneumonia	Normal	
Predicted value	ATB (case)+	134	2	2	138
	Not ATB (case) –	5	38	98	141
	Total (case)	139	40	100	279

detected non-ATB cases was 136, and the number of false positives was 4, as shown in Table 3. Figure 6 (a) shows the original four consecutive CT layers of an ATB patient in the testing data, and Fig. 6 (b) shows the corresponding four CT layers of the ATB patient diagnosed by our AI software. Figure 7 is a ROC curve plotted for our AI detection software on the testing data. The area under the ROC curve reached 0.980 (AUC = 0.980). We computed the following performance indices for our software: The accuracy rate is 0.968, the SS is 0.964, the SP is 0.971, the PPV is 0.971, and the NPV is 0.964, as shown in Table 4.

4. Discussion

Tuberculosis is a disease that occurs throughout the world. Because chest X-rays have limitations, CT has become important for the diagnosis of ATB [20]. Researchers have tried using various automated detection methods to diagnose tuberculosis [6, 21–24]. However, to the best of our knowledge, chest radiographs are commonly used for detecting ATB, and no studies have been conducted for detecting ATB in CT images using AI. A study on the detection of pulmonary tuberculosis using AI in CT has reported the prediction of multi-drug resistant tuberculosis in CT lung images based on deep learning technology, with an accuracy rate of 91.11% [25].

In the data analysis, we found that the minimum ages of ATB, normal, and pneumonia patients were 12, 5, and 14 respectively, and the maximum ages were 89, 68, and 89, respectively. In the testing data, the minimum ages of ATB, normal, and pneumonia patients were 7, 2, and 18 respectively, and the maximum ages were 85, 83, and 86, respectively. In each group of cases, there are minors (under 18 years old) and elderly (over 60 years old), which means that the data contains immature young lungs,



(a)

(b)

Fig. 6. CT slices of an ATB patient in the testing data. (a) Original slices and (b) abnormalities detected by the software.

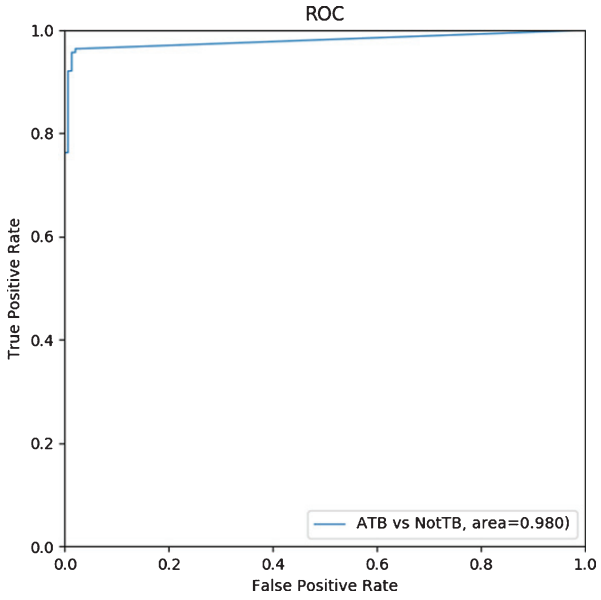


Fig. 7. ROC curve for ATB vs no-ATB.

Table 4
Software performance on testing data

Evaluation index	AUC	Accuracy	SS	SP	PPV	NPV
Result	0.980	0.968	0.964	0.971	0.971	0.964

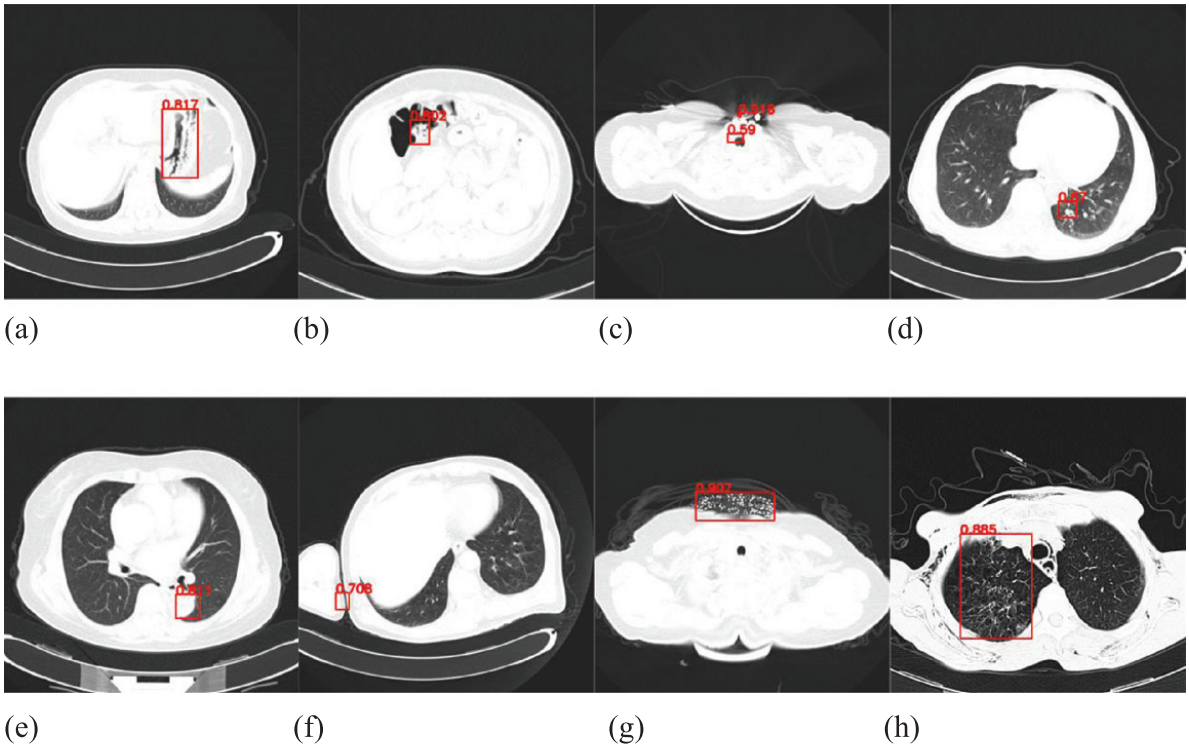


Fig. 8. Misdiagnosed CT images of normal and pneumonia patients by U-Net: (a) stomach; (b) intestinal; (c) sternal angle and trachea; (d) left pulmonary vascular cross section; (e) thoracic aorta; (f) right humerus; (g) metal clothing *in vitro*; (h) right-sided pneumonia lesion.

which may be smaller than the lungs of adults, while the elderly may have atrophy and other factors leading to smaller lungs. The factor of age has little effect on the AI detection software.

In contrast to previous AI research in medical imaging diagnosis, in the development process, we adopted the recommendations of radiologists. Radiologists in the CT room need to handle patients with different diseases during their daily work, which requires AI to identify diseases like radiologists. Accordingly, the AI detection software of this study first tries to distinguish ATB, pneumonia, and normal CTs. Unlike chest X-rays, CT diagnosis requires several consecutive slices to be inspected. Therefore, when diagnosing ATB patients, image processing methods are proposed in this study that can effectively convert the 2D ROI in each CT slice into a 3D OOI, which has been described previously in Section 2.2.2. The recommendations of doctors and radiologists were a tremendous help to our image processing method for AI diagnosis of ATB. The AI detection software in this paper not only relies on neural networks to make the diagnosis, but also lays the foundation for adding other diseases later and for applying ATB detection software at a clinic in the future.

In this paper, U-Net is used to perform image segmentation prediction on the testing data. As a result, we found that some slices in normal and pneumonia patients in the testing data were misjudged as ATB regions, as shown in Fig. 8. The causes of misdiagnosis include: stomach, intestine, sternal angle and trachea, pulmonary blood vessels, thoracic aorta, humerus, wearing of metal clothes, and pneumonia lesions mistaken for ATB lesions. Therefore, we adopted an image processing method to eliminate the ATB lesion areas incorrectly predicted by U-Net in the CT image.

Four cases of misdiagnosis are shown in Fig. 9 (a), (b): Two cases of pneumonia were misdiagnosed as ATB, because the two cases of pneumonia lesions affected more than two lung lobes. The other 2 cases of normal patients were misdiagnosed as ATB slices, as shown in Fig. 9 (c) and (d). Figure 9 (c) shows how the AI software incorrectly responded to the heightened cross section of the blood vessels

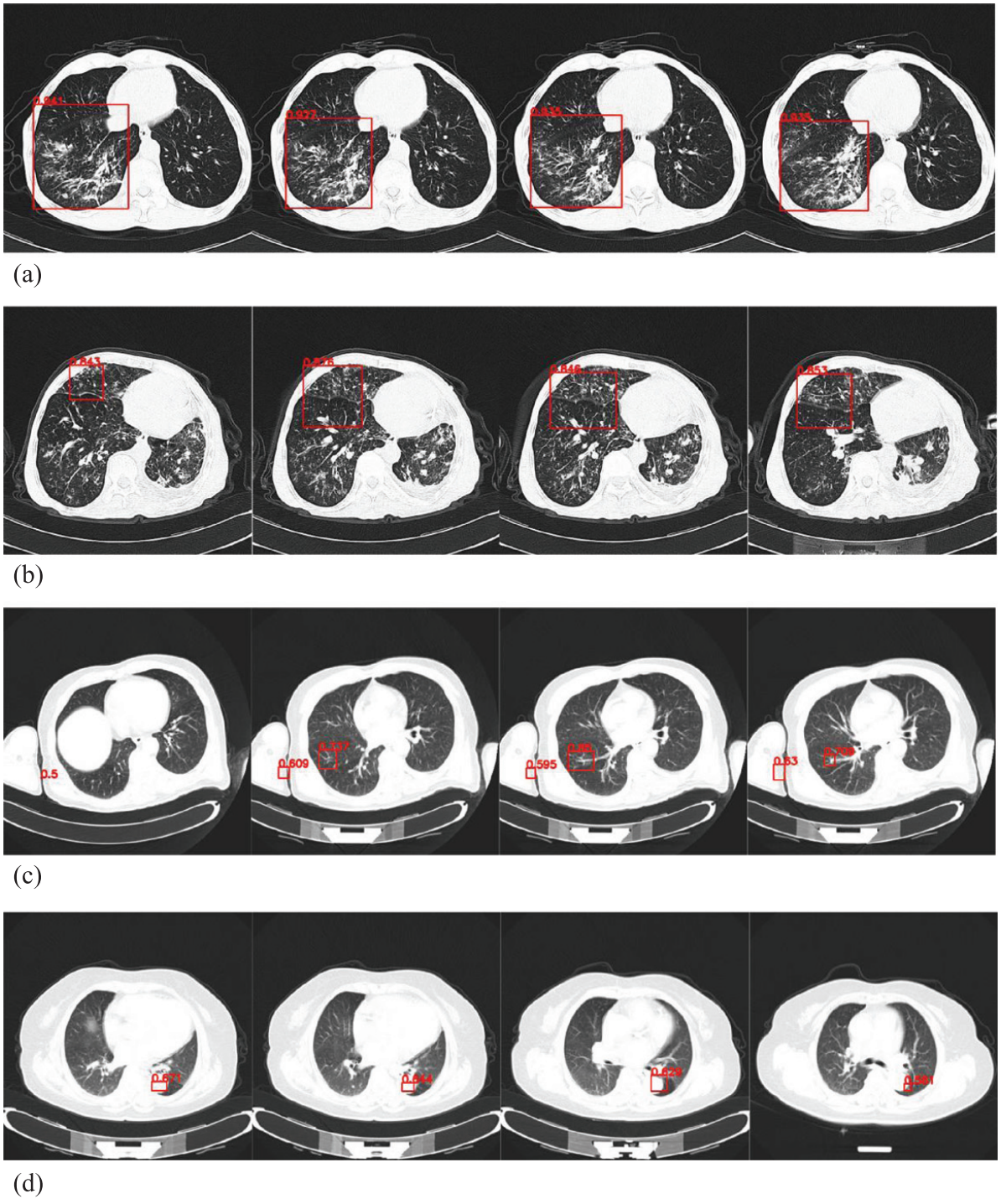
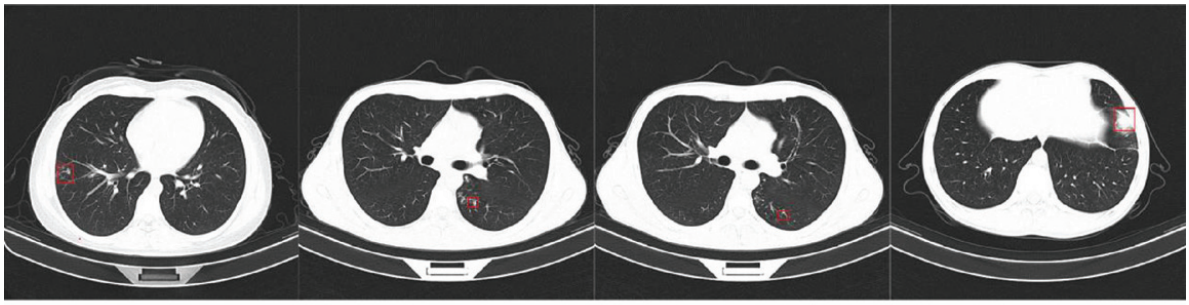


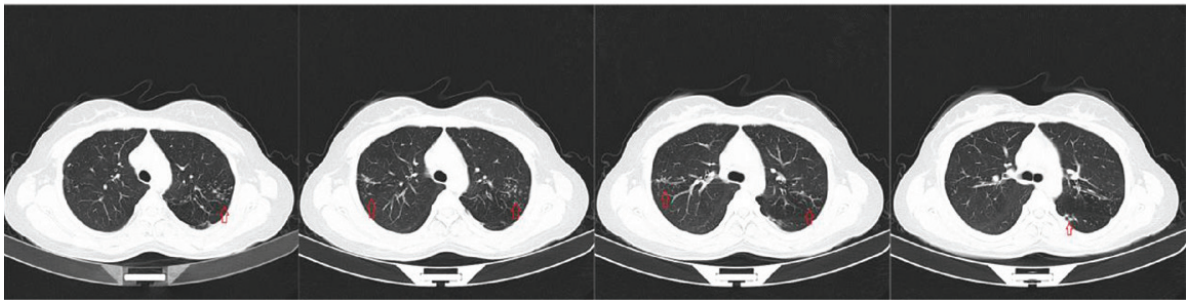
Fig. 9. AI detection software was used to diagnose images of normal and pneumonia patients in the testing data (a) right-sided pneumonia lesion; (b) right-sided pneumonia lesion; (c) images of right humeral head and right lung in normal patient; (d) normal thoracic aortic images.

in the right lung and the right humerus. Density shadows are misdiagnosed as ATB lesions; Fig. 9 (d) shows that the software mistakes a high-density shadow of a normal thoracic aorta as an ATB lesion.

Figure 10 shows cases missed by the AI detection software. We found that five cases of ATB were missed. After further analysis, we found that the common imaging characteristics of the five missed patients were small and atypical lesions. The ATB lesions in some slices were relatively



(a)



(b)

Fig. 10. CT slices of an ATB patient missed by the software.

small and not predicted by U-Net. The ATB lesions were predicted in some slices; however, less than 4 consecutive slices with predicted ATB lesions were generated by the software, so the software eventually misdiagnosed the patient, as shown in Fig. 10.

The AI detection software developed in this study has following advantages and potentially clinical application value. First, ATB has a high incidence in resource-poor areas and areas where people's quality of life is not high. These areas are often lacking radiologists [26]. The accuracy, sensitivity, and specificity of the AI software developed in this paper can help to reduce the workload of radiologists in remote areas. Second, this AI detection software has laid a foundation for distinguishing ATB from other lung diseases in future research. Third, image processing methods have been added to effectively convert a 2-D region of interest into a 3-D object of interest for a CT scan, see Section 2.2.2. This method can delete false positive slices and effectively improve the accuracy of diagnosis.

This study also has limitations. First, our training data does not contain other diseases and abnormalities, such as lung abscess, obsolete pulmonary tuberculosis, lung tumors, multiple nodules and other lung diseases. We plan to add other lung diseases in the next step to widen the scope of the algorithm. Second, because this study assumes that more than four consecutive lesion slices define an ATB patient, this may cause lesions of less than two centimeter in size to be missed (CT slice thickness is 5 mm); however, considering that most ATB lesions are larger, the impact is relatively small.

5. Conclusion and future work

This study successfully developed an AI software tool for automatic detection of ATB in chest CT. The software has a high performance, with an AUC value of 0.980. The accuracy is 0.968, the SS is 0.964, the SP is 0.971, the PPV is 0.971, and the NPV is 0.964. In summary, our study shows that using AI detection software can not only diagnose ATB patients, and distinguish between ATB and non-ATB (normal and pneumonia), but also simplify the diagnosis process for doctors. Therefore, this

AI detection software lays a foundation for the application of AI in CT diagnosis of ATB in clinical medicine.

In order to improve the completeness and practicality of this software, future research should continue adding training data for other lung diseases. At present, drug-resistant tuberculosis has become a problem. In our next study, we will use CT data of drug-resistant tuberculosis to train and develop an AI detection system based on deep learning to predict drug-resistant tuberculosis in CT images.

Acknowledgments

This work was funded by Shenzhen Science and Technology Program (Grant No. KQTD2017033 110081833). This work was partially supported by the Intramural Research Program of the Lister Hill National Center for Biomedical Communications (LHNCBC) at the U.S. National Library of Medicine (NLM), National Institutes of Health (NIH).

References

[1] World Health Organization. Global tuberculosis report 2018. [2018-09-18], https://www.who.int/tb/Publications/global_report/en/. 2018

[2] World Health Organization. Systematic screening for active tuberculosis: principles and recommendations, World Health Organization, 2013.

[3] S.W. Lee, Y.S. Jang, C.M. Park, et al. The role of chest CT scanning in TB outbreak investigation, *Chest* **137**(5) (2010), 1057–1064.

[4] A.S. Bhalla, A. Goyal, R. Guleria, et al. Chest tuberculosis: Radiological review and imaging recommendations, *The Indian Journal of Radiology & Imaging* **25**(3) (2015), 213.

[5] C. Lange and T. Mori, Advances in the diagnosis of tuberculosis, *Respirology* **15**(2) (2010), 220–240.

[6] J. Melendez, C.I. Sánchez, R. Philipsen, et al. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information, *Scientific Reports* **6** (2016), 25265.

[7] A.H. Van't Hoog, H.K. Meme, H. Van Deutekom, et al. High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey, *The International Journal of Tuberculosis and Lung Disease* **15**(10) (2011), 1308–1314.

[8] R.O. Panicker, B. Soman, G. Saini, et al. A review of automatic methods based on image processing techniques for tuberculosis detection from microscopic sputum smear images, *Journal of Medical Systems* **40**(1) (2016), 17.

[9] E.J. Hwang, S. Park, K.N. Jin, et al. Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs, *Clinical Infectious Diseases* **69**(5) (2018), 739–747.

[10] X. Zhao, S. Qi, B. Zhang, et al., Deep CNN models for pulmonary nodule classification: model modification, model integration, and transfer learning, *Journal of X-ray Science and Technology* **27**(4) (2019), 615–629.

[11] J.R. Zech, M.A. Badgeley, M. Liu, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, *PLoS Medicine* **15**(11) (2018), e1002683.

[12] V.V. Dantsev, A.S. Golota, V.G. Karpuschenko, et al. The modern state and improvement prospects of community-acquired pneumonia and pulmonary tuberculosis differential diagnostics, *Voenno-meditsinskii Zhurnal* **336**(5) (2015), 29–36.

[13] H. Qu, W. Zhang, J. Yang, et al. The value of the air bronchogram sign on CT image in the identification of different solitary pulmonary consolidation lesions, *Medicine* **97**(35) (2018), e11985.

[14] J.M. Goo, J.G. Im, CT of tuberculosis and nontuberculous mycobacterial infections, *Radiologic Clinics of North America* **40**(1) (2002), 73–87.

[15] H.P. McAdams, Radiological manifestations of pulmonary tuberculosis, *Radiol Clin Morth Am* **33**(4) (1995), 655–678.

[16] A. Nambu, A. Saito, T. Araki, et al. Chlamydia Pneumoniae: Comparison with Findings of Mycoplasma Pneumoniae and Streptococcus Pneumoniae at Thin-Section CT, *Radiology* **238**(1) (2006), 330–338.

[17] A. Nambu, K. Ozawa, N. Kobayashi, et al. Imaging of community-acquired pneumonia: Roles of imaging examinations, imaging diagnosis of specific pathogens and discrimination from noninfectious diseases, *World Journal of Radiology* **6**(10) (2014), 779–793.

[18] L. Song, J. Liu, B. Qian, et al. A deep multi-modal CNN for multi-instance multi-label image classification, *IEEE Transactions on Image Processing* **27**(12) (2018), 6025–6038.
[19] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham (2015), 234–241.
[20] R.L. Eisenberg and N.R. Pollock, Low yield of chest radiography in a large tuberculosis screening program, *Radiology* **256**(3) (2010), 998–1004.
[21] S. Jaeger, A. Karargyris, S. Candemir, et al. Automatic screening for tuberculosis in chest radiographs: a survey, *Quantitative Imaging in Medicine and Surgery* **3**(2) (2013), 89.
[22] S. Jaeger, A. Karargyris, S. Candemir, et al. Automatic tuberculosis screening using chest radiographs, *IEEE Transactions on Medical Imaging* **33**(2) (2013), 233–245.
[23] S. Vajda, A. Karargyris, S. Jaeger, et al. Feature selection for automatic tuberculosis screening in frontal chest radiographs, *Journal of Medical Systems* **42**(8) (2018), 146.
[24] L. Hogeweg, C.I. Sánchez, P. Maduskar, et al. Automatic detection of tuberculosis in chest radiographs using a combination of textural, focal, and shape abnormality analysis, *IEEE Transactions on Medical Imaging* **34**(12) (2015), 2429–2442.
[25] X.W. Gao, Y. Qian, Prediction of multidrug-resistant TB from CT pulmonary images based on deep learning techniques, *Molecular Pharmaceutics* **15**(10) (2017), 4326–4335.
[26] C. Qin, D. Yao, Y. Shi, et al. Computer-aided detection in chest radiography based on artificial intelligence: a survey, *Biomedical Engineering Online* **17**(1) (2018), 113.

Journal of X-Ray Science and Technology 28 (2020) 391–404
DOI 10.3233/XST-200687
IOS Press

Clinical and radiological features of novel coronavirus pneumonia

Qiuting Zheng[a], Yibo Lu[b], Fleming Lure[c,d,*], Stefan Jaeger[e] and Puxuan Lu[a,*]
[a]*Department of Medical Imaging, Shenzhen Center for Chronic Disease Control, Guangdong Shenzhen 518020, China*
[b]*Department of Medical Imaging, The Fourth People's Hospital of Nanning, Guangxi Nanning 530023, China*
[c]*MS Technologies, 10110 Molecular Dr., Suite 305, Rockville, MD 20850, USA*
[d]*Shenzhen Zhiying Medical Co., Ltd, Guangdong Shenzhen 518020, China*
[e]*National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA*

Received 7 April 2020
Revised 18 April 2020
Accepted 5 May 2020

Abstract. Recently, COVID-19 has spread in more than 100 countries and regions around the world, raising grave global concerns. COVID-19 transmits mainly through respiratory droplets and close contacts, causing cluster infections. The symptoms are dominantly fever, fatigue, and dry cough, and can be complicated with tiredness, sore throat, and headache. A few patients have symptoms such as stuffy nose, runny nose, and diarrhea. The severe disease can progress rapidly into the acute respiratory distress syndrome (ARDS). Reverse transcription polymerase chain reaction (RT-PCR) and Next-generation sequencing (NGS) are the gold standard for diagnosing COVID-19. Chest imaging is used for cross validation. Chest CT is highly recommended as the preferred imaging diagnosis method for COVID-19 due to its high density and high spatial resolution. The common CT manifestation of COVID-19 includes multiple segmental ground glass opacities (GGOs) distributed dominantly in extrapulmonary/subpleural zones and along bronchovascular bundles with crazy paving sign and interlobular septal thickening and consolidation. Pleural effusion or mediastinal lymphadenopathy is rarely seen. In CT imaging, COVID-19 manifests differently in its various stages including the early stage, the progression (consolidation) stage, and the absorption stage. In its early stage, it manifests as scattered flaky GGOs in various sizes, dominated by peripheral pulmonary zone/subpleural distributions. In the progression state, GGOs increase in number and/or size, and lung consolidations may become visible. The main manifestation in the absorption stage is interstitial change of both lungs, such as fibrous cords and reticular opacities. Differentiation between COVID-19 pneumonia and other viral pneumonias are also analyzed. Thus, CT examination can help reduce false negatives of nucleic acid tests.

Keywords: 2019-novel Coronaviruses (2019-nCoV), SARS-CoV-2, COVID-19, Computed Tomography (CT), ground glass opacity (GGO), differentiation of pneumonia

1. Background and current situation of the epidemics

Since December 2019, an epidemic of pneumonia of unknown cause has spread rapidly to the whole Hubei Province, the whole country, and multiple regions and countries around the world. The pathogen

*Corresponding author: Fleming Lure, MS Technologies, 10110 Molecular Dr., Suite 305, Rockville, MD 20850, USA; Shenzhen Zhiying Medical Co., Ltd, Guangdong Shenzhen 518020, China. E-mail: f.lure@hotmail.com; Puxuan Lu, Department of Medical Imaging, Shenzhen Center for Chronic Disease Control, Guangdong Shenzhen 518020, China. E-mail: lupuxuan@126.com.

of this pneumonia with unknown cause was isolated successfully by Chinese scientists on January 7th, 2020 and confirmed to be a newly-discovered novel coronavirus through whole-genome sequencing. The Coronaviridae Study Group (CSG) of the International Committee on Taxonomy of Viruses, World Health Organization (WHO), has assessed the placement of the human pathogen, tentatively named 2019-nCoV, within the Coronaviridae. CSG recognizes this virus as forming a sister clade to the prototype human and bat severe acute respiratory syndrome coronaviruses (SARS-CoVs) of the species Severe acute respiratory syndrome-related coronavirus, and designates it as SARS-CoV-2. Sequencing was completed for the virus on January 10th, 2020. China then shared the data globally (https://www.gisaid.org/), and shortly after a diagnostic method based on fluorescence quantitative RT-PCR was developed for 2019-nCov [1]. The epidemic of novel coronavirus pneumonia in China has become a "Public Health Emergency of International Concern (PHEIC)", as declared by Tedros Adhanom Ghebreyesus, director general of WHO, in Geneva, on January 30th, 2020. The National Health Commission issued Announcement No. 1 on January 20th, 2020, listing novel coronavirus pneumonia as one of the Statutory Class-B Infectious Diseases, for which Class-A Measures would be adopted [2]. On February 11th, 2020, WHO named novel coronavirus pneumonia as "COVID-19", with CO standing for corona, VI for virus, and D for disease.

As of April 5th, 2020, there are 1,299 existing confirmed cases (including 265 severe cases), 77,078 accumulative cases cured and discharged, 3,331 accumulative deaths, 81,708 accumulative confirmed cases, and 88 existing suspected cases, as reported by 31 provinces (autonomous regions or municipalities directly under the central government) and Xinjiang Production and Construction Corps in China. Globally, the epidemic is rampant in the Western Pacific, Europe, Southeastern Asia, the Middle East, the Americas, and Africa, manifesting most severely in USA, Spain, and Italy, while the number of infected people continues to rise. The global impact by the epidemic is far greater than that induced by SARS in 2003 [3].

2. Etiological features and pathogenic mechanism

The literature has attributed the outbreak of the COVID-19 Pneumonia to a place called Huanan Seafood Market in Wuhan City, Hubei Province, China, so far, where live animals such as poultry, bats, marmots and other wild animals are sold, suggesting the possible transmission of pathogens from animals to humans [4]. The present whole-genome phylogenetic analysis on 2019-nCoV shows that it has a much closer phylogenetic relationship to the SARS-like coronaviruses bat-SL-CoV ZC45 and bat-SL-CoV ZXC21 from Rhinolophus Sinicus (a species of Chinese Horseshoe Bat) [5–7]. Therefore, bats are considered the possible main host of 2019-nCoV [8]. It has not been defined, however, whether 2019-nCoV pneumonia is transmitted through Rhinolophus Sinicus directly or via an intermediate host. Most scholars believe that Rhinolophus Sinicus is the most primitive host of the 2019-nCoV, and that the virus might be spread to humans through a yet unknown animal host, possibly pangolin [9, 10]. Recently, 149 mutational sites in 2019-nCoV have been detected, as suggested by the most recent Chinese research. The virus has evolved into two subtypes, L- and S-subtype, with the latter having stronger aggressiveness and infectiousness [11].

The results of Coronavirus gene sequencing performed by Brazilian researchers for the first confirmed case in Brazil, on February 26th, showed three differences in the Novel Coronavirus (Brazil/SPBR1/2020) compared with the virus gene (Hu-1 Reference Strain) published in Wuhan; that is to say, mutations might have occurred in the virus during its transmission [12].

Corona viruses belong to the subfamily of Ortho coronavirus under the family of Coronaviridae and catalog of Nidovirales. This subfamily comprises four attributes including Alpha-coronavirus (α), Beta-coronavirus (β), Gamma-coronavirus (γ), and Delta-coronavirus (δ). Attributes α and β tend

to infect mammals, and contain seven species of coronavirus causing disease in humans, including HCoV-OC43, HCoV-229E, SARS-CoV, HCoV-NL63, HCoV-HKU1, MERS-CoV, and SARS-CoV-2 (2019-nCoV), while attributes γ and δ infect mainly birds [13]. 2019-nCoV, with a genome structure typical of coronavirus, is a positive single-strand RNA virus particle capsulated in a diameter of about 60–140 nm and a size of 30 kb. With spike spines on its capsule, the whole virus looks like a corona [14–16].

The genome of coronavirus usually encodes four structural proteins, including spike protein (S), membrane protein (M), envelope protein (E), and nucleocapsid protein (N). Some coronaviruses of attribute β also encode hemagglutinin esterase protein (HE). S protein mediates the attachment of the virus to a receptor on a cellular surface and is one of the key factors achieving effective interpersonal transmission. Protein S, belonging to Type-I transmembrane glycoprotein, consists of two domains: a receptor-binding subunit (S1) and a membrane fusion subunit (S2). In the process of virus invasion, S1 is responsible for binding with the receptor on the surface of the host cell followed by viral attachment, and S2 is responsible for fusing the cell membrane of the host and viral envelope, making the virus genome enter the host cell to form a stable binding compound. The whole process of infection consists of four steps: adsorption invasion, gene synthesis, packaging of the mature virus, and virus release [17, 18]. Through the pathway of S protein binding ACE2 receptor in human cells, 2019-nCoV constitutes a major risk for public health due to human transmission [6]. In addition, ACE2 is mainly situated at alveolar Type II epithelial cells in the lower lungs, so 2019-nCoV is more likely to cause severe diseases such as pneumonia.

3. Epidemiological features

Most of the early patients infected with COVID-19 pneumonia were in Wuhan, indicating a local outbreak [1]. Later, most of the patients had been to Wuhan or in close contact with patients coming from Wuhan [19–21]. Meanwhile, confirmed cases have appeared in other regions of China and have been reported in many countries and regions outside of China [22]. Infection of medical staff and family clusters has shown that SARS-CoV-2 spreads in population clusters [23] with stronger infectiousness than SARS-CoV or MERS-CoV [19], and that the epidemic has developed into community transmission. The epidemic has expanded rapidly and spread from Hubei Province to other areas of China with the immigration of infected people, while the number of cases increased gradually all over the world.

The main source of infection is the population of COVID-19 patients, including asymptomatic infected people. Identification of the transmission chain and subsequent tracing of the contacts become more complex when several infected people seem asymptomatic or only mildly symptomatic [24]. The main routes of transmission include respiratory droplets and close contact. Transmission via aerosols seems possible in case of prolonged exposure to aerosols in high concentration in a relatively closed environment. SARS-CoV-2 has been isolated from urine and feces of patients in multiple regions; therefore, transmission via aerosol or contact due to environmental contamination should be considered [14, 20]. Mother-to-child transmission need to be confirmed, as well as other transmissions [10]. The latent period lasts 1–14 days, mostly 3–7 days. The longest reported latent period is 24 days [20], although in an individual case. People are generally susceptible, and patients are concentrated in a population aged between 30 and 79 years [19] and are correlated with exposure to viral load. Conditions after infection are more serious in the older patients and those with underlying diseases, while severe diseases seem rare among children and infants [22, 25].

4. Clinical manifestation and laboratory examination

The symptoms of SARS-CoV-2 infection seem to be non-specific, as far as clinical manifestations are being concerned. They may be very similar to influenza and manifest mainly in fever, fatigue, dry cough, sore throat, headache, and occasionally nasal congestion, runny nose, and diarrhea in a few patients. Clinically, cases can be classified into mild, common, severe, and critical. Dyspnea and/or hypoxemia occur(s) usually in severe cases within one week. Serious cases can progress rapidly into acute respiratory distress syndrome (ARDS), septic shock, refractory metabolic acidosis, and coagulation dysfunction. It is noteworthy that patients with severe or critical diseases might manifest only moderate to low fever, or even no obvious fever during the course of the disease. Patients with mild disease may manifest only low fever and mild fatigue without manifestations of pulmonary inflammation and can recover after one week in most cases.

RNA, the hereditary substance of the virus, may become detectable post systemic infection of SARS-CoV-2. Nucleic acid detection aims to find RNA of SARS-CoV-2 in samples from the patient, so it is also the "Golden Standard" and an important approach in clinical diagnosis. There are mainly two methods in detecting nucleic acids of SARS-CoV-2, including NGS (Next-generation sequencing) and reverse transcription polymerase chain reaction (RT-PCR) [26]. NGS, the next generation of sequencing technology, was the first method that succeeded in detecting the new pathogen at the initial stage of the epidemic and defined soon the sequences of nucleic acids of SARS-CoV-2. The simplest and fastest strategy for detection will be undoubtedly the targeted fluorescence quantitative RT-PCR using primer probes designed for the conserved domains of the nucleic acid sequence that have been defined for the virus. This method involves a proliferation of specific RNA sequences in the sample post to their reverse transcription. Theoretically, the quantity of genome segments of the virus post to each amplification will be multiplied, and the visual detection will become feasible when segments of the targeted genes reach a certain number after more than 30 amplifications. Targets detected for nucleic acids of SARS-CoV-2 consist of three conserved sequences in the viral genome, including open reading frame 1ab (ORF1ab), nucleocapsid protein (N), and Envelope gene [27].

According to the Laboratory Guidelines for Novel Coronavirus Infections issued by the China Health Commission [28], the prerequisite for a laboratory confirmation of a positive case should be the positiveness of real-time fluorescence RT-PCR results in two specific targeted genes including ORF1ab and gene N, or single-target positiveness of RT-PCR either in two types of specimens simultaneously or in two separately-sampled specimens of the same category. Re-sampling and re-test are needed, in case of a positiveness of a single target. The test should be performed at least for ORF1ab region, the most conserved and the most specific region of 2019-nCoV, if only one target gene is detectable because of limited availability of the selected laboratory kit. COVID-19 cannot be ruled out, even if there is a negative result. Factors inducing false negative should be excluded, such as poor quality of the specimen (e.g., specimen from respiratory tract such as oropharynx), specimens collected too early or too late, incorrect conservation, transportation or treatment, reasons of the technology per se (e.g., viral mutation, PCR inhibition) [29].

The leukocyte count in peripheral blood is normal or decreased in early stage, with a low lymphocyte count, especially low numbers of T lymphocytes, which are in an over-activated state that can cause severe immune damage in patients [30]. C-reactive protein (CRP) and ESR increase in most patients, and procalcitonin remains normal. In quite a few patients, increase of liver enzymes, muscle enzymes, and myoglobin are observed. In severe cases, D-dimer increases and lymphocytes in peripheral blood decrease progressively. SARS-CoV-2 can be positive in specimens including nasopharynx swab, sputum, secretion frp, secretion of lower respiratory tract, blood, urine, and feces. Importantly, symptoms last longer in cases of SARS-CoV-2 than in most cases of non-complicated influenza [31, 32]. Despite non-typical symptoms occurring in a few cases, fever is still the typical symptom of a

SARS-CoV-2 infection. Patients with underlying chronic diseases generally have prolonged courses, severer conditions, more rapid progressions, and worse prognoses [21, 31].

5. Pathological and radiological manifestations

The autopsy of the first death due to COVID-19 in China showed obvious lung injuries, with patchy grayish lesions and dark red hemorrhage discernable by the naked eye. The texture of the lung tissue became hard, losing the sponginess inherent to lungs. A large amount of thick secretions could be seen escaping from the section of alveoli, and fiber cords were observable [33].

Histological examination: under the light microscope, pulmonary interstitial blood vessels were congested and edematous, with inflammatory infiltration of lymphocytes and monocytes as well as clear thrombosis in the vessels. Serous fluid, fibrin exudate, and formation of hyaline membrane were observed in the alveolus cavities. The exudate cells were mainly monocytes and macrophages, and multinucleated giant cells were also common. In addition, atypical enlarged alveolar cells could be seen, in which the atypical enlarged alveolar cells had large nuclei, amphiphilic cytoplasmic granules and obvious nucleoli, showing viral cytopathic-like changes [30]. Some alveolar epithelial cells were exfoliated, with inclusion bodies inside. Partial alveolar exudate organization and interstitial fibrosis occurs. Local hemorrhage, necrosis, and hemorrhagic infarction might occur in lung tissue. Part of the mucous epithelium of intrapulmonary bronchi was exfoliated, with mucus and mucus thrombus observable in the lumen. A few alveoli were overinflated, with broken alveoli septum or formation of cysts. Under an electron microscope, coronavirus particles could be seen in the cytoplasm of bronchial epithelial cells and Type II alveolar epithelial cells. Immunohistochemistry staining proved the presence of SARS-CoV-2 antigen in some of the alveolar epithelial cells and macrophages, and SARS-CoV-2 nucleic acid was positive as shown by RT-PCR [14]. Local pathological changes suggested a high similarity of COVID-19 with SARS-CoV and MERS-CoV infection.

DR imaging is a convenient and fast examination method for patients with lung diseases, but low sensitivity and specificity due to the influence of an overlap of anatomic structure on the observation of lesions make it easy to miss the image manifestation of COVID-19. COVID-19 usually has no abnormal change or manifests as bronchitis in the early stage. In the progression stage, it may manifest limited or multiple segmental patchy opacities in the middle and peripheral zones/subpleural areas of both lungs. In severe patients, multiple consolidations and ground-glass opacities (GGO) are seen in both lungs, and some of them are fused into large consolidations, with a small amount of pleural effusion or pleural thickening. The progression of lesions into the critical illness may manifest multiple diffusive consolidations in both lungs, appearing "white lungs" with a small amount of pleural effusion. Manifestations in the absorption stage include dissipation or decreased density of the previous lesions, or evolution into fiber or cord-like opacities. Therefore, DR is suitable for only primary hospitals without CT, or for severe or critically severe patients [34].

It has been reported that abnormal CT images may occur in 60% to 93% of the cases prior to a positive nucleic acid antibody test (or both occur simultaneously) [35]. The sensitivity of chest CT is greater than that of RT-PCR, which supports the use of chest CT to screen for COVD-19, especially when patients consistent with clinical and epidemiological profiles of COVD-19 but negative in RT-PCR tests are screened [36]. Therefore, chest CT scans with an appropriate scanning plan and parameters for COVID-19 patients will have a synergistic diagnostic effect in evaluating patients. The present consensus about COVID-19 among Chinese radiological experts is that high-resolution CT (HRCT) should be the main modality in screening and diagnosing this disease, with 16 layers or above, routine dosage, plain scans, a reconstruction layer thickness of not more than 3 mm and a layer thickness of 1 mm [37].

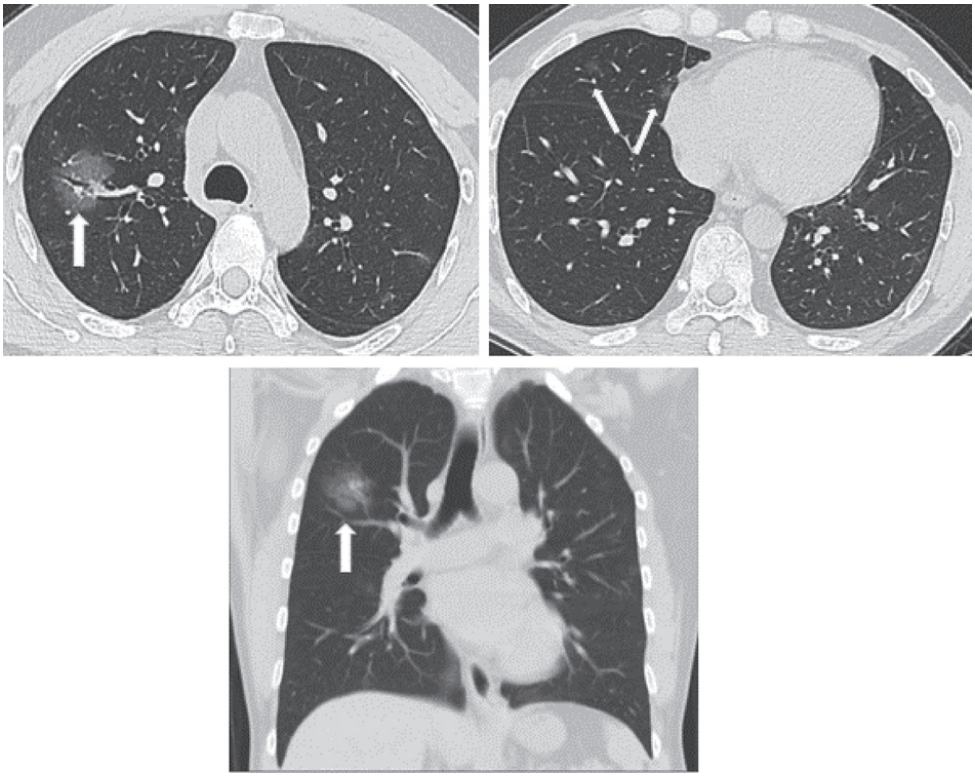


Fig. 1. Example images show a 52-year male patient who had onset of fever up to 39.1° on January 26th, 2020, complicated with fatigue and dry cough with a positive nucleic acid test. He underwent the first CT scan on January 27th, 2020, day 2 after the onset, which showed in the upper right lung a patchy GGO containing thickened blood vessels and air bronchogram. Small patches of GGOs are scattered in the middle lobe of the right lung.

The CT manifestations of COVID-19 depend on a patient's age, immune status, disease stage at the time of scanning, underlying diseases, and drug intervention. Main radiological features include: (1) distribution: dominantly in peripheral pulmonary/subpleural bands, and distributed along bronchovascular bundles, (2) quantity: mostly three or more lesions, and rarely individual or pairs, (3) shape: patchy, bulky, or nodular, or in a shape of blocks, cellular texture, or reticular, (4) density: mostly uneven, manifesting GGOs, signs of paving stone, thickened interlobular septum with consolidation, and thickened bronchial wall, and (5) various complicated symptoms, but rare signs of pleural effusion or mediastinal lymphadenopathy.

According to guidelines issued by the Radiology Branch of the Chinese Medical Association (*Imaging Diagnostic Guideline for Novel Coronavirus Pneumonia - Edition 2020),* the Infectious Disease Group of the Chinese Society of Radiology (*Adjuvant Imaging Diagnostic Guideline for Novel Coronavirus Pneumonia)*, the Project Group of Prevention and Control of Novel Coronavirus Pneumonia in Zhongnan Hospital of Wuhan University *(Rapid Guideline for Diagnosis and Treatment of Novel Coronavirus Pneumonia (SARS-CoV-2))*, as well as domestic and international literature, CT manifestations for COVID-19 can be classified into three stages. Specifically, based on the onset timepoint and the body's response to the virus, manifestations can be classified into the following three stages:

5.1. Early stage

This stage lasts 1–3 days after the onset of clinical manifestations, including fever, and dry cough. The pathological changes in this stage include dilation and congestion of alveolar septal capillaries, exudate in alveolar cavity, and interstitial edema in interlobular septum. CT scans manifest single

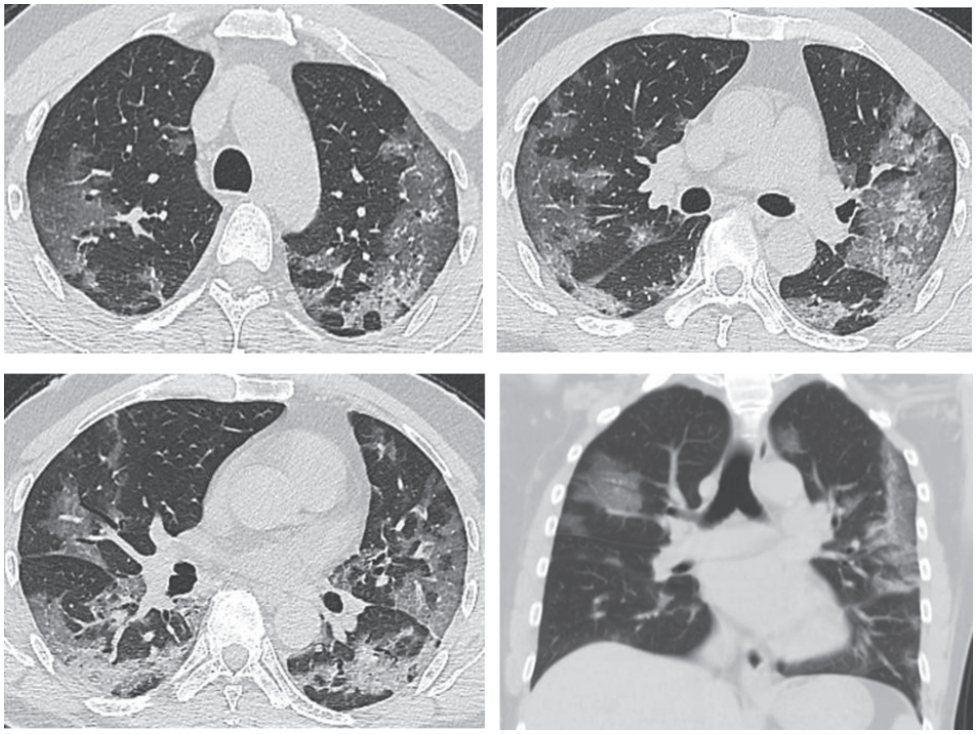


Fig. 2. Images of the same patient as shown by Fig. 1, which manifest an obviously increased number of lesions with progression in both lungs in CT scans on February 3rd, 2020, day 9 after the onset. Multiple large patchy GGOs appear in both lungs with consolidations containing air bronchogram inside and obvious thickening of interlobular septum.

or multiple scattered patchy or bulky GGOs, separated by lobular septum in honeycomb or reticular thickening [38–44] (Fig. 1).

Generally, lesions in the early stage do not involve an entire pulmonary section. A patient without any other pulmonary disease will not show lymphadenectasis in the mediastinum or hilum, pleural thickening, or pleural effusion. The pathological changes are considered to be correlated with pathological mechanisms including the vulnerability of terminal bronchioles and the pulmonary parenchyma around the respiratory bronchioles for viral pneumonia in the early stage, subsequent involvement of the whole pulmonary lobule, and diffusive lobular injuries.

5.2. Progressive stage

This stage refers to day 3–14 after the onset of clinical manifestations. The pathological feature during this stage is that the interstitial vascular dilation seems more obvious than before, and the exudate containing a large number of cells accumulates in the alveolus cavity, both of which will further aggravate the alveolus and interstitial edema. Cellulosic exudation enters every alveolus through intercellular connection to form a fusion state. CT scans show faded GGOs and consolidations appearing more distributed in comparison with the early stage and maintaining the dominance in peripheral pulmonary/subpleural distribution, with enlargement and fusion of some lesions invading subsequently multiple pulmonary lobes. The lesions are irregular, wedge-shaped or fan-shaped, with unclear borders, scattered in multiple foci or even diffuse, showing bilateral asymmetry. Opacities of soft tissues are observed, including broncho-vascular bundle thickening or subpleural multifocal pulmonary consolidation, showing rapid progression and changes, with great morphological variation within short-term reexamination. This can be complicated with tissue necrosis to form small cavities, and air bronchogram is commonly seen. It may also be accompanied by thickening of interlobular septum, manifesting

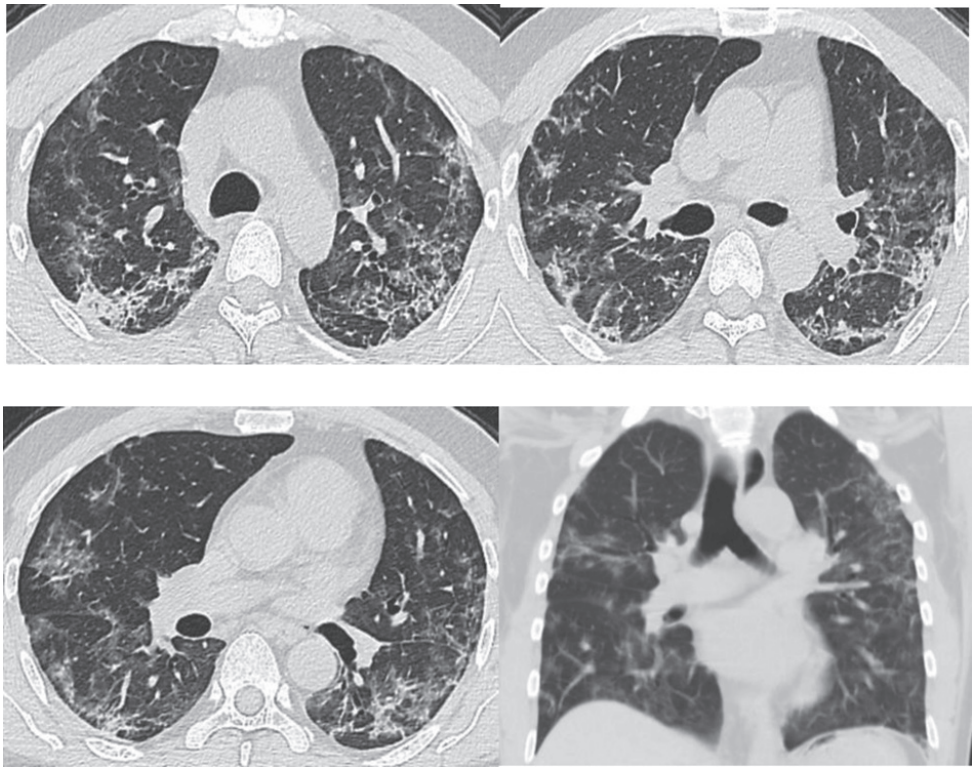


Fig. 3. Images of the same patient as shown by Figs. 1 and 2. CT scans on February 21st, 2020, day 24 after onset (during the absorption stage), show a gradual decrease and absorption of large patches of GGOs in both lungs, which are scattered with cord-like, reticulate and patchy opacities, with obvious thickening of interlobular septum.

"paving stone" sign, or complicated with fibrotic focus. Some cases manifest GGOs, usually without pleural effusion. Very few cases are complicated with lymphadenectasis in mediastinum and hilum. The disease in the progression stage often evolves abruptly and needs vigorous treatment. One should be alert of the occurrence of acute respiratory distress syndrome (ARDS) (Fig. 2).

If the patient does not receive effective treatment in time, pulmonary consolidation may occur on basis of the existing GGOs, in addition to the enlargement of GGOs and increase of their number. The main pathological feature in this stage is the cellulosic exudation in alveolar cavities and the gradual disappearance of capillary hyperemia of alveolar walls. Diffusive lesions in both lungs are observed in CT scans, and a few patients manifest "white lungs" [45, 46]. The lesions may increase by 50% in their sizes within 48 hours, dominated by consolidations and complicated with GGO, air bronchogram, and multiple cord-like opacities, which is called clinically severe pneumonia and may develop clinically into the critically severe type if the disease keeps progressing.

5.3. Absorption stage

This stage refers to the 2-3 weeks' time period after the onset of clinical symptoms. The area of lesions shrinks after effective treatment. CT scans manifest patchy consolidations or cord-like opacities. The exudates undergo absorption or organization by the body over time with decreased density, and the pulmonary consolidations dissipate gradually. Along with the gradual absorption of the lesions, reticulate interlobular septal thickening is observable, with thickened and twisted bronchial walls in the image of cord-like changes, and a few scattered patchy high-density opacities being visible [38–46] (Fig. 3).

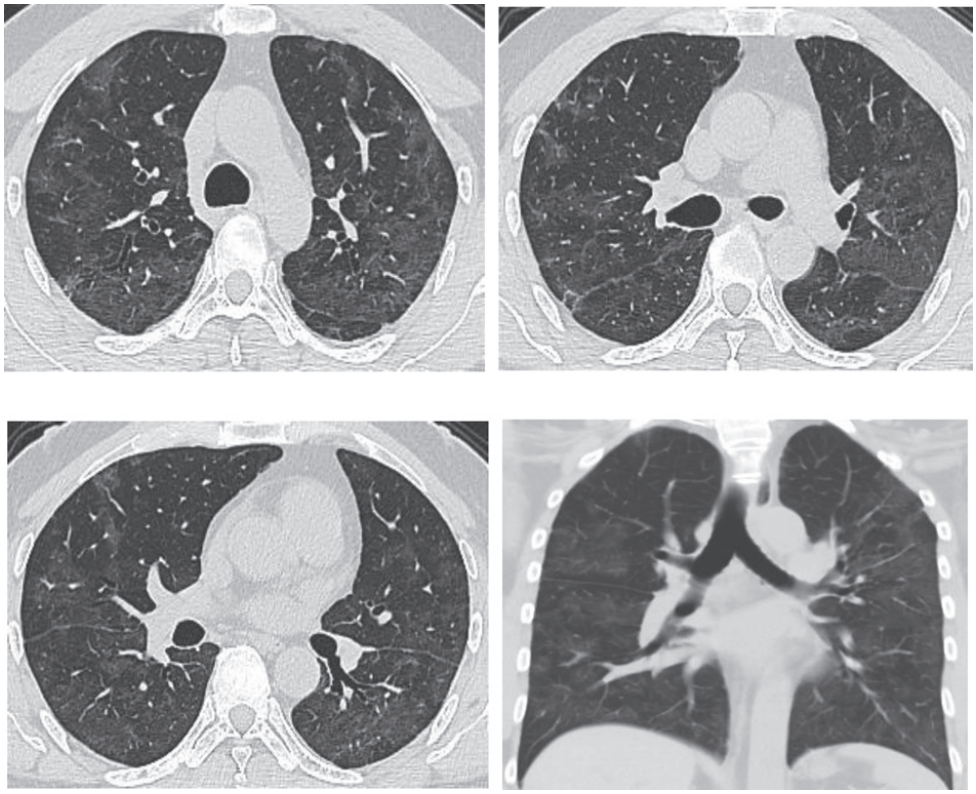


Fig. 4. Chest CT scans on March 20th, 2020, for the same patient shown by Figs. 1–3, after treatment, showing obvious absorption of lesions (decreased lesions with partial fibrosis). After admission, the patient received antiviral and atomizing therapies, but symptoms aggravated with repeated fever, cough, expectoration, and gasp. A written notice of danger was given on February 2nd 2020, and the patient was given oxygen inhalation, continuous ECG monitoring, finger pulse oxygen monitoring, anti-infection treatment, anti-virus therapy, atomization of recombinant human interferon α2b and intravenous injection of human immunoglobulin to improve immunity, as well as symptomatic treatments to reduce inflammation. The symptoms improved after treatment. Chest CT scans on March 20th showed the obvious absorption of the lesions. Re-test of nucleic acid of novel coronavirus after an interval of 24 hours was negative. The patient was discharged in agreement after consulting with the Municipal Novel Coronavirus Pneumonia Expert Group of the 4th People's Hospital of Nanning City.

The patient of Fig. 3 had a follow-up chest CT examination four weeks later. The images showed that the patient's symptoms significantly improved after treatment with the obvious absorption of lesions (Fig. 4).

The characteristic feature of COVID-19 is GGO, which is a radiological concept manifesting a minor increase of density and a cloudy opacity, but with observable textures of blood vessels and bronchi inside [47]. GGO is a non-specific term and highly suggests that the disease of lung tissue is active or reversible. Lesions of GGOs in lungs of patients with COVID-19 decrease gradually over time, while fibrous cord-like opacities increase gradually, which becomes the most common radiological manifestation. Research has shown that the most obvious changes of intrapulmonary lesions are observable during day 6–9 after admission in 75.0% of the patients, and relatively obvious absorption can be seen during day 10–14 of the hospital stay in 76.9% of the patients [48]. In addition, some research work implies that the dandelion fruit sign is the characteristic change in COVID-19 patients [49]. The longer the duration of a SARS-CoV-2 infection, the more changes can be seen in CT images [50]. The radiological changes in late sequela remain unknown [51].

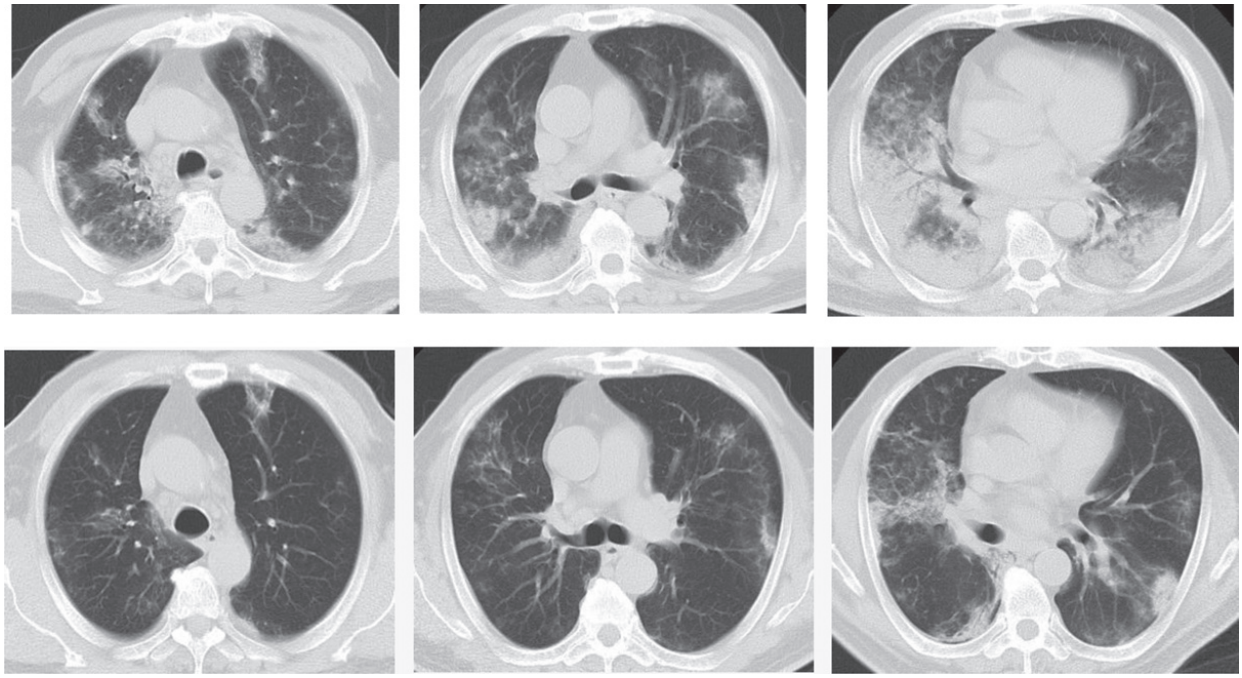


Fig. 5. Example CT images of a 63-year-old male patient diagnosed with H7N9 avian pneumonia. Top row shows chest CT scans taken on January 17th, 2014, day 11 after the onset, in the progression phase. CT scans show patches of ground glass opacities in the upper right lung and the upper left lung, dominantly distributed in the subpleural zone. Consolidations can be seen in the lower lobes of both lungs, with air bronchogram inside dominantly at the lower right lung. Bottom row shows chest CT taken on January 24th, 2014, day 18 after the onset during the absorption period. CT shows the obvious absorption of ground glass opacities and consolidations in both lungs, which have been dominantly interstitial changes and small patches of consolidations.

6. Differential diagnosis

6.1. Differentiation between COVID-19 pneumonia and other viral pneumonias

Other viral pneumonia manifest generally diffuse large patches of GGOs in both lungs with thickening of interlobular septum, which makes it hard to differentiate them from COVID-19 pneumonia in images and clinically. Definite epidemic history provides suggestive efficacy for differentiation for all kinds of viral pneumonias, especially COVID-19, while confirmed diagnosis should be based on etiological detection.

6.1.1. Differentiation with pneumonia induced by SARS and MERS

Pathogens for SARS, MERS, and COVID-19 all belong to the coronavirus family, with similar pathogenic mechanisms among them all. Given the diversity in imaging, without obvious specificity, it is hard to differentiate merely based on imaging. The incorporation of epidemic features and pathogenic tests may facilitate the discrimination.

6.1.2. Differentiation with pneumonia by avian influenza virus

Avian influenza virus pneumonia in human is an infectious disease of the acute respiratory tract induced by several subtypes of the avian influenza virus such as H7N9 and H5N1. It spreads mainly through contact with poultry, in contrast to the contamination through wild animals for the novel coronavirus pneumonia showing an obvious human-human transmission. As for imaging, lesions seem wider in lungs for avian influenza pneumonia in human with severe pulmonary consolidation, although

both pneumonias manifest patches or large sheets of ground glass opacities and consolidations in the lung without major difference (Fig. 5). Therefore, some difficulties exist in differentiating them.

6.2. Differentiation with mycoplasma pneumonia

Mycoplasma pneumonia, observed commonly in children and young people, is characterized clinically by paroxysmal and irritant dry cough and by the chest X-ray showing either interstitial infiltration (e.g., increased, thickened or gridded lung markings) or consolidations of segments or lobes manifesting patchy or fan-shaped infiltration. Imaging of CT scans is characterized by thickening of the bronchial wall, centrilobular nodules, consolidations distributed along lobes, segments or subsegments of the lung, and enlargement of mediastinal lymph nodes. In comparison, enlarged mediastinal lymph nodes are rarely seen in novel coronavirus pneumonia. Positiveness in mycoplasma antibody can be observed in laboratory tests, as a response to Macrolides antibiotics. A differentiation can be made, therefore, based on clinical features and serological examination.

6.3. Differentiation between COVID-19 pneumonia and bacterial pneumonia

Bacterial pneumonia often manifests shivering, high fever, and bloody or rusty expectoration, with laboratory tests showing an increase of leukocytes and neutrophils. Imaging usually shows small patches of consolidations distributed along the bronchi and even fused into the distribution along subsegments, segments, or lobes, which show good response to antibiotics and are easily differentiated with COVID-19 pneumonia.

7. Summary

Currently, detection of viral nucleic acid is still the gold standard for diagnosing COVID-19 [52], although it has high specificity but low sensitivity. Temporal difference exists in clinical manifestations, between nucleic acid detection and CT imaging during the early stage of COVID-19, which might be the critical factor for misdiagnosing the disease, and a factor for transmission of the disease [53]. A cross validation should be performed by clinicians, based on clinical manifestations and CT imaging, to make the clinical diagnosis of COVID.

It is also very likely that outbreaks of new diseases caused by coronavirus might occur in the future, given our limited understanding of SARS-CoV-2, climate and ecological changes possibly affecting the course of outbreaks, and continuing interactions between humans and animals. Therefore, future research should focus on developing effective drugs and vaccines through strengthened international cooperation to investigate and contain infectious diseases caused by coronavirus as a vital public health issue [54].

Funding Acknowledgments

This study was partially funded by Shenzhen Science and Technology Program (Grant No. KQTD2017033110081833), as well as partially supported by the Intramural Research Program of the Lister Hill National Center for Biomedical Communications (LHNCBC), the U.S. National Library of Medicine (NLM), and the National Institutes of Health (NIH).

References

[1] C. Wang, et al., A novel coronavirus outbreak of global health concern, *Lancet* **395**(10223) (2020), 470–473.
[2] National Health Commission. An Announcement by National Health Commission of People's Republic of China (No. 1 in 2020) [EB/OL]. [2020-01-21] [2020-04-06]. http://www.gov.cn/xinwen/2020-01/21/content_5471158.htm.
[3] World Health Organization (WHO). Summary table of SARS cases by country, 1 November 2002 – 7 August 2003[EB/OL]. [2003-08-15] [2020-04-06]. https://www.who.int/csr/sars/country/2003_08_15/en/.
[4] N.S. Chen, et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study, *Lancet* **395**(10223) (2020), 507–513.
[5] Y. Chen, et al., Emerging Coronaviruses: genome structure, replication, and pathogenesis, *J Med Virol* **92**(4) (2020), 418–423.
[6] X. Xu, et al., Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission, *Sci China Life Sci* **63**(3) (2020), 457–460.
[7] Y.Z. Zhou, et al., Analysis of the variation and evolution of coronavirus SARS CoV-2, *Journal of Southern Medical University* **40**(02) (2020), 152–158.
[8] S. Perlman, Another decade, another coronavirus, *N Engl J Med* **382**(8) (2020), 760–762.
[9] South China Agricultural University. Press conference of research on novel coronavirus pneumonia epidemic in Guangzhou, February 7th, 2020.
[10] Expert Team of Control of Novel Coronavirus Pneumonia under Chinese Society of Preventive Medicine, The latest knowledge of epidemiology of novel coronavirus pneumonia, *Chinese Journal of Viral Diseases* (2020), 1–7. https://doi.org/10.16505/j.2095-0136.2020.0015.
[11] X.L. Tang, et al., On the origin and continuing evolution of SARS-CoV-2, National Science Review, nwaa036, https://doi.org/10.1093/nsr/nwaa036.
[12] J. G. de Jesus, et al, First cases of coronavirus disease (COVID-19) in Brazil, South America Virological, 2020. http://virological.org/t/first-cases-of-coronavirus-disease-covid-19-in-brazil-south-america-2-genomes-3rd-march-2020/409
[13] Z. Ma, G.J. Cao, M. Guan, Status and progress in researches on human coronavirus, *International Journal of Laboratory Medicine* **41**(05) (2020), 518–522.
[14] Notification of printing and distributing therapeutic regimens for novel coronavirus pneumonia (Trial Version VII), General Office of National Health Commission, and Office of State Administration of Traditional Chinese Medicine. [2020-03-03] [2020-04-06]. http://www.gov.cn/zhengce/zhengceku/2020-03/04/content_5486705.htm.
[15] Y. Gong, et al., Present status in researches on coronavirus, *Chinese Journal of Bioengineering* **40**(Z1) (2020), 1–20.
[16] B.L. Xu, et al., Progress in researches on novel coronavirus COVID-19, *Chinese Journal of Nosocomial Infection* **30**(6) (2020), 806–811.
[17] T.Y. Qiu, et al., Identification of potential cross-protective epitope between 2019-nCoV and SARS virus, *J Genet Genomics* **47**(2) (2020), 115–117.
[18] F. Li, Structure, function, and evolution of coronavirus spike proteins, *Annu Rev Virol* **3**(1) (2016), 237–261.
[19] Epidemiology Team in Emergency Response Mechanism for Novel Corona virus Pneumonia in Chinese Center for Disease Control and Prevention, Analysis of epidemiological features of novel coronavirus pneumonia, *Chinese Journal of Epidemiology* **41**(2) (2020), 145–151.
[20] W.J. Guan, et al., Clinical characteristics of 2019 novel coronavirus infection in China. NEJM. Published online February 28, 2020. DOI: 10.1056/NEJMoa2002032.
[21] Y.J. Zhuang, et al., Clinical and epidemiological characteristics of 26 confirmed cases of novel coronavirus pneumonia, *Chinese Journal of Nosocomial Infection* **30**(6) (2020), 817–820.
[22] Y.H. Jin, et al., A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version), *Mil Med Res* **7**(1) (2020), 4.
[23] R.C. Wang, et al., Emergence of SARS-like Corona virus poses new challenge in China, *Journal of Infection* **80**(3) (2020), 350–371.
[24] V.J. Munster, et al., A Novel Corona virus Emerging in China – Key Questions for Impact Assessment, *N Engl J Med* **382**(8) (2020), 692–694.
[25] H.S. Shi, et al., Clinical features and imaging findings of novel coronavirus (2019-nCoV) pneumonia, *Clinical Radiology* **39**(01) (2020), 8–11.
[26] V.M. Corman, et al., Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* **25**(3) (2020), 2000045.
[27] F. Wu, et al., A new coronavirus associated with human respiratory disease in China, *Nature* **579**(7798) (2020), 265–269.

[28] Notice of Prevention and Control Plan for Novel Corona virus Pneumonia issued by the General Office of National Health Committee of People's Republic of China (Fifth Edition) [EB/OL]. [2020-02-21] [2020-04-06]. http://www.nhc.gov.cn/jkj/s3577/202002/a5d6f7b8c48c451c87dba14889b30147.shtml.
[29] W. Da, et al., Misunderstanding of nucleic acid testing for 2019-nCoV, *Chinese Journal of Nosocomiology* **30**(8) (2020), 1153–1156.
[30] Z. Xu, et al., Pathological findings of COVID-19 associated with acute respiratory distress syndrome, *Lancet Respir Med* **8**(4) (2020), 420–422.
[31] D. Wang, et al., Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Corona virus-Infected Pneumonia in Wuhan, China, *JAMA* **323**(11) (2020), 1061–1069.
[32] J.W. Tang, et al., Emergence of a novel coronavirus causing respiratory illness from Wuhan, China, *J Infect* **80**(3) (2020), 350–371.
[33] Q. Liu, et al., Report of gross findings of systemic anatomy in autopsy of a death due to novel coronavirus pneumonia, *Journal of Forensic Medicine* **36**(1) (2020), 19–21.
[34] Infectious Diseases Group of Radiology Branch of Chinese Medical Association, A radiological diagnostic guideline for novel coronavirus pneumonia Version I, 2020, 2020-1-24.
[35] T. Ai, et al., Correlation of Chest CT and RT-PCR Testing in Corona Virus Disease 2019 (COVID-19) in China: A Report of 1014 Cases, *Radiology*, 2020, 200642.
[36] Y.C. Fang, et al., Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR, *Radiology*, 2020, 200432.
[37] H.X. Guan, et al., Preliminary study of clinical imaging features of 2019 novel coronavirus (2019-nCoV) pneumonia, *Radiological Practice* (2020), **1–6**. https://doi.org/10.13609/j.cnki.1000-0313.2020.02.001.
[38] Professional Committee of Infection and Inflammation Radiology of Chinese Research Hospital Association, Guideline for adjuvant imaging diagnosis for novel coronavirus pneumonia, *Chinese Medical Imaging Technology* **36**(3) (2020), 1–11.
[39] Radiology Branch of Chinese Medical Association, Radiological diagnosis of novel coronal pneumonia: Expert Recommendation by Radiology Branch of Chinese Medical Association (Version I), *Chinese Journal of Radiology* (2020), 54. DOI: 10.3760/cma.j.issn.1005-1201.2020.0001.
[40] Prevention and Control Group of Novel Corona virus Pneumonia of Zhongnan Hospital of Wuhan University) Evidence-Based Medicine Branch of China Healthcare Association for International Exchange and Promotion, Rapid guideline for diagnosis and treatment of novel coronavirus pneumonia (2019-nCoV) 2020-02-02.
[41] Q. Liang, Imaging examination, diagnosis and prevention and control of nosocomial infection in patients with novel coronavirus pneumonia: Consensus of radiologists in Hunan Province, *Journal of Central South University (Medical Version)* **45**(03) (2020), 221–228.
[42] C.Y. Liu, et al., A comparative study on CT findings between novel coronavirus pneumonia manifesting ground-glass opacities and early-stage lung tumor, *Chinese Journal of Thoracic and Cardiovascular Surgery* **27**(04) (2020), 376–380.
[43] X.M. Gong, et al., Preliminary discussion of CT findings of novel coronavirus pneumonia (COVID-19), *Radiological Practice* **35**(03) (2020), 261–265.
[44] F.M. Liu, et al., Chest CT findings and clinical features of novel coronavirus pneumonia (COVID-19), *Radiological Practice* **35**(03) (2020), 266–268.
[45] K. Wang, et al., An analysis on features of chest CT findings of novel coronavirus pneumonia, *Chinese Clinical Medicine* **27**(01) (2020), 27–31.
[46] G.Z. Fu, et al., The application of chest CT scans to screening of novel coronavirus pneumonia, *Journal of Wenzhou Medical University* (2020), 1–9. http://kns.cnki.net/kcms/detail/33.1386.r.20200219.1451.002.html.
[47] K. Ikeda, et al., Differential diagnosis of ground-glass opacity nodules: CT number analysis by three dimensional computerized quantification, *Chest* **132**(3) (2007), 984–990.
[48] J.C. Wang, et al., Dynamic changes of chest CT findings in patients with 2019 coronavirus disease (COVID-19), *Journal of Zhejiang University (Medical Version)* (2020), 1–13. http://kns.cnki.net/kcms/detail/33.1248.R.20200225.1528.004.html.
[49] X.B. Fu, et al., Dandelion fruit sign: a CT sign for diagnosing novel coronavirus pneumonia, *Journal of Southern Medical University* (2020), 1–5. http://kns.cnki.net/kcms/detail/44.1627.R.20200228.1823.004.html.
[50] A. Bernheim, et al., Chest CT Findings in Corona virus Disease-19 (COVID-19): Relationship to Duration of Infection, *Radiology* 2020, 200463. doi:10.1148/radiol.2020200463.
[51] J. Chen, et al., An analysis on clinical features of 29 cases of 2019 novel coronavirus pneumonia, *Chinese Journal of Tuberculosis and Respiration* **43** (2020), E005.
[52] Disease Prevention and Control Bureau of the National Health Commission. Prevention and control plan for novel coronavirus pneumonia (Version II) (2020), 1–22.

[53] Y.H. Du, et al., Preliminary discussion on clinical features and CT findings in early stage of family cluster of novel coronavirus pneumonia, *Journal of Xi'an Jiaotong University (Medical Version)* **41**(02) (2020), 215.
[54] R. Min, et al., Progress in researches on clinical features and pathogenesis of novel coronavirus pneumonia, *Chinese Journal of Nosocomial Infection* **30**(8) (2020), 1136–1141.