scientific data

Check for updates

OPEN

DATA DESCRIPTOR

An Integrated Mycobacterial CT Imaging Dataset with Multispecies Information

Zhilin Han[1,2,4], Yuyang Zhang[3,4], Wenlong Ding[1], Xiaoting Zhao[2], Bingzhen Jia[1], Tingting Liu[1], Liang Wan[2] ✉ & Zhiheng Xing[1,3] ✉

The increasing global incidence of nontuberculous mycobacterial (NTM) pulmonary disease highlights the need for rapid diagnostic methods to guide timely treatment and prevent antibiotic misuse. While bacterial culture remains the gold standard for diagnosis, its extended turnaround time compromises clinical decision-making. Computed tomography (CT), with its high sensitivity for lung lesions and rapid imaging capabilities, has emerged as a critical diagnostic tool. AI-assisted CT interpretation shows particular promise for improving NTM detection, yet progress has been hindered by limited datasets due to disease rarity. We address this gap by introducing the first comprehensive CT dataset combining 430 NTM and 871 tuberculosis cases, supplemented with clinical parameters including demographics, symptoms, and mycobacterial species data. This resource aims to catalyze AI algorithm development for differential diagnosis, ultimately enhancing precision in NTM management through advanced machine learning applications.

Background & Summary

Respiratory diseases have long been significant factors threatening public health security. Before the COVID-19 outbreak in 2020, tuberculosis (TB) was the leading cause of death from infectious diseases. In 2019, it was estimated that 10 million people worldwide were infected with tuberculosis, and 1.4 million people died from the disease[1]. In recent years, the incidence of tuberculosis has slightly decreased. However, the incidence of non-tuberculous mycobacterial (NTM) diseases, which have clinical manifestations similar to TB[2], has been increasing, posing a significant burden on public health systems[3–8]. This trend is also evident in China. Studies have shown that from 2006 to 2020, a total of 14.82 million pulmonary TB cases were reported, with an average incidence rate of 73.8 per 100,000 population. The age-standardized incidence rate (ASR) exhibited a continuous decline, decreasing from 89.1 per 100,000 in 2006 to 52.3 per 100,000 in 2020[9]. In contrast, the prevalence of NTM infections in China increased from 1.6% during 2004–2009 to 3.13% during 2012–2017, indicating a significant upward trend. Among TB-suspected patients[10], the NTM infection rate in northern China was 2.7%, significantly lower than the 8.6% reported in southern China[11]. These data highlight the declining TB burden alongside the growing clinical significance of NTM infections, mirroring the global shift in mycobacterial disease epidemiology. Some studies have also speculated that the declining incidence of TB has led to decreased immunity against NTM in both hosts and populations[12].

Non-tuberculous mycobacteria refer to a diverse group of mycobacterial species excluding the Mycobacterium tuberculosis complex and Mycobacterium leprae. NTM are ubiquitous in the environment, predominantly found in water and soil. Pathogenic NTM species can infect individuals across all age groups, leading to infections in multiple organs and tissues, such as lymph nodes, skin and soft tissues, meninges, brain parenchyma, synovial fluid, muscles, bones, and the genitourinary system. Among these, pulmonary infections are the most prevalent and severe, accounting for approximately 70% to 80% of NTM cases, manifesting as NTM pulmonary disease. A smaller proportion involves extrapulmonary tissues, presenting as extrapulmonary NTM disease[13].

[1]Department of radiology, Tianjin Haihe Hospital, TCM Key Research Laboratory for Infectious Disease Prevention for State Administration of Traditional Chinese Medicine, Tianjin Institute of Respiratory Diseases, Haihe Hospital, Tianjin University, Tianjin, China. [2]Academy of medical engineering and translational medicine, Tianjin University, Tianjin, China. [3]Haihe Clinical College, Tianjin Medical University, Tianjin, China. [4]These authors contributed equally: Zhilin Han, Yuyang Zhang. ✉e-mail: lwan@tju.edu.cn; 18920696025@189.cn

Study	Data Type	Sample Size	Task Type	Methods	Performance Metrics
Reference1[40]	Chest X-ray	301 NTM, 804 TB, 80 External Test	Disease Classification	3D-ResNet	Internal AUC: 0.86; External AUC: 0.78
Reference2[41]	CT images & T-SPOT	467 NTM, 582 TB	Disease Classification	T-SPOT+Deep Learning	Accuracy: 91.7% (NTM-PD,T-SOT -), 89.8% (TB),T-SPOT +)
Reference3[42]	Chest X-ray	2204 NTM	Prognostic Prediction	Deep Learning+Logistic Regression	AUC (10-year): 0.922; AUC (5-year): 0.942; AUC (3-year): 0.865
Reference4[43]	Chest CT	206 TB, 186 NTM	Cavity Detection & Quantification	3D nnU-Net	Mean Dice Score: 78.9; ICC (patient): 0.991; ICC (lesion): 0.933
Reference5[44]	Chest X-ray	937 NTM, 2377 TB	Disease Classification	CNN+Ensemble Learning	Internal AUC: 0.90 External AUC: 0.81
Reference6[45]	Chest X-ray	500 TB, 500 NTM, 500 Imitator	Disease Classification	Deep Neural Network (DNN)	Internal AUC:0.86; External AUC:0.64

Table 1. AI-Related Research in NTM Diagnosis.

Clinically, NTM is difficult to diagnose due to its non-specific clinical manifestations and the challenge of isolating pathogenic bacteria[14,15].Bacterial culture and strain identification are the only methods for identifying NTM, and this process is time-consuming, often requiring up to two months. As a result, while awaiting bacterial identification results, empirical anti-TB medications are often prescribed to patients with clinically suspected sputum AFB-positive pulmonary TB. Consequently, a considerable number of NTM-LD patients receive unnecessary anti-TB treatment. Misdiagnosis not only leads to unnecessary time and economic costs but also exposes patients to the risk of drug side effects[16].

Regarding the internal classification of non-tuberculous mycobacteria, NTM comprises various species types[17–19], many of which commonly affect the lungs, such as the *Mycobacterium avium complex* (MAC), *Mycobacterium kansasii*, and *Mycobacterium abscessus*. The treatment strategies for each strain vary accordingly[20,21]. The distribution of mycobacterial species also exhibits regional characteristics[22]. Currently, species identification and differentiation in NTM diagnosis rely heavily on bacteriological examination, which, despite its accuracy, is time-consuming and requires specific samples, such as sputum or bronchial alveolar lavage fluid[23]. Since the symptoms of NTM-LD are often indistinguishable from those of other respiratory diseases, and most NTM strains require 2–3 weeks to grow, the diagnosis of NTM-LD typically relies on imaging results for initial suspicion.

Computed tomography (CT) is one of the most effective methods for detecting various lung diseases[24,25], due to its high sensitivity to the lobes, trachea, and bronchi. This enables it to efficiently identify abnormalities within the lung fields. Moreover, the rapid scanning and imaging capabilities of CT allow for a swift assessment of structural changes in a patient's lungs. Chest CT is widely used in diagnosis, monitoring, determining the timing of treatment initiation, and assessing the response to NTM-PD treatment[26]. If valuable diagnostic information and species typing for NTM could be obtained from CT scans, it would significantly enhance the efficiency and specificity of subsequent treatments, reduce the overall treatment duration, and improve patient prognosis[27]. The imaging characteristics of NTM pulmonary disease mainly include exudation, cavities, and nodules[28] while TB primarily manifests as cavities, lung consolidation, patchy lung nodules, mediastinal lymphadenopathy, lymph node calcification, and pleural effusion[29]. Due to the similarities in imaging features between NTM and TB, the use of imaging features to differentiate NTM from TB remains a subject of debate across multiple studies[30–34].

In recent years, artificial intelligence (AI)-assisted imaging diagnostic have been extensively researched and applied for common tuberculosis (TB) and the COVID-19 pandemic[35–38]. However, studies focusing on NTM-PD remain relatively limited. Recently, we have conducted NTM classification research using radiomics, machine learning, and deep learning methods on this dataset. In radiomics tasks, we developed a radiomics model based on multiple lesions, achieving an AUC of 90.2%. For machine learning tasks, we used a linear support vector machine (SVM) to evaluate the diagnostic significance of cavities and bronchiectasis in distinguishing NTM from TB, with the bronchiectasis-based model achieving an AUC of 0.84 ± 0.06[39]. In deep learning tasks[40], we applied a 3D-ResNet model to perform binary classification between NTM and TB, achieving AUCs of 0.86 and 0.78 ($p < 0.05$), accuracies of 0.83 and 0.69, specificities of 0.57 and 0.63, and sensitivities of 0.92 and 0.75 on the test and external validation sets, respectively. Ying *et al.*[41] utilized a deep learning (DL) model to classify CT images as either NTM or TB and combined these results with each patient's T-SPOT test outcomes, offering more reliable conclusions for the early differential diagnosis of NTM-PD and PTB. Lee *et al.*[42] developed a DL model capable of predicting the mid- to long-term mortality of NTM patients using baseline chest X-rays, with predictability improving when clinical data is included. Other studies[43–45], as summarized in Table 1, have applied deep learning methods to AI-assisted tasks related to NTM. While several AI models claim to surpass clinical doctors in performance, the current DL models used for NTM diagnosis are not yet suitable for cross-institutional application and are not ready for implementation in real clinical settings.

Training AI models requires large datasets, yet collecting cases for NTM, a relatively rare disease, is challenging. To our knowledge, there are currently no publicly available NTM imaging datasets on major open-source platforms. Therefore, we plan to make our dataset publicly accessible to promote the development and application of AI technology in the field of NTM.

Our dataset includes CT data from 430 NTM patients and 871 TB patients, collected at Tianjin Hai He Hospital from 2014 to 2023. This database also includes basic patient information such as gender, age, and

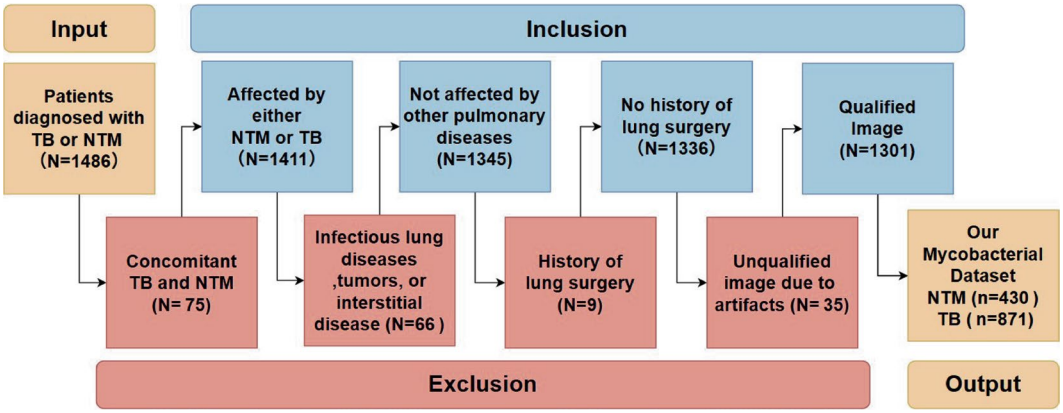

Fig. 1 Flowchart of Patient Inclusion and Exclusion for the Dataset.

clinical symptoms. Additionally, the dataset contains annotated lesion information for 240 patients, comprising 120 cases each of TB and NTM, with imaging features like ground-glass opacities (GGO) and bronchiectasis. Of the 430 NTM cases, 308 include species classification results. Compared to other chest imaging datasets, our dataset not only has a large number of cases but also offers rich supporting content. This makes it suitable for not only TB and NTM binary classification tasks but also for extended functions like image segmentation and species identification.

Methods

This section describes the data collection process, including the acquisition and annotation of imaging data, as well as subsequent standardization and anonymization steps. In addition, we provide detailed statistical information to facilitate researchers in making better use of this dataset. We also demonstrate the applicability of this dataset in deep learning approaches by applying several publicly available models. Finally, we analyze the results and discuss the potential limitation of this dataset. The data samples used in this study were approved by the Medical Ethics Committee of Tianjin Haihe Hospital (Approval No.: 2024HHSQKT-002). Data were collected from January 2014 to April 2024 and the process was divided into two phases. In the first phase (January 2014–December 2018), retrospective data were collected. During this phase, all sensitive patient information was anonymized using an automated script based on the pydicom library (https://pydicom.github.io/). This process strictly adhered to the relevant image transmission standards by masking and deleting any patient-identifiable information contained within the files, thereby ensuring that the rights and well-being of the subjects were not adversely affected. Consequently, the ethics committee waived the requirement for obtaining informed consent from the patients. In the second phase (January 2019–April 2024), prospective data were collected. During this phase, the principal investigator or project staff provided an oral explanation to participants regarding the nature, purpose, risks, and benefits of the study, and complete information was also provided in writing. Participants signed an informed consent form, thereby consenting to the use of their clinical test results and CT imaging data for AI training and analysis. Furthermore, the imaging data may be published for scientific purposes provided that no personal identity information is disclosed, and rigorous anonymization procedures were similarly applied in this phase. The ethics committee has also authorized the public release of this dataset. No animal experiments were involved in this study.

Data collection. Following standard clinical collection protocols, data were collected from patients diagnosed with NTM or TB pulmonary infections at Tianjin Haihe Hospital between January 2014 and April 2024, all of whom underwent chest CT scans before initiating NTM treatment. The NTM & TB imaging dataset was curated through a structured screening process, as illustrated in Fig. 1. Patients were excluded if they had a history of lung surgery, concurrent diagnoses of both diseases, other pulmonary conditions (including infections, tumors, and interstitial lung diseases), or poor CT image quality caused by respiratory motion or metal artifacts.

The patients in this study were scanned using two spiral CT devices (GE Healthcare's BrightSpeed CT machine and Canon Medical Systems' Aquilion Prime 128 CT machine) under the same scanning protocol. Patients were scanned in a supine position, trained to take maximum inspiration and hold their breath during scanning to ensure high-quality data. Scanning was performed from the lung apex to the base. FOV was adjusted according to the patient's body.

The BrightSpeed CT machine used the following scanning parameters: a resolution of 512×512, tube voltage of 120 kV, automatic tube current modulation, a rotation time of 0.75 seconds per rotation, a collimator width of 16×0.625 mm, and images reconstructed using standard and lung algorithms with a slice thickness and interval of 1.25 mm, resulting in a voxel size of 0.31–1.14 mm^3. For the Aquilion Prime 128 CT machine, the parameters included a resolution of 512×512, tube voltage of 120 kV, automatic tube current modulation, a rotation time of 0.5 seconds per rotation, a collimator width of 64×0.5 mm, and images reconstructed using FC 30 and FC 52 algorithms with a slice thickness of 1.0 mm and an interval of 0.8 mm. Previous studies have indicated that thicker slice images may result in the loss of important details and reduce classification accuracy. Therefore, we

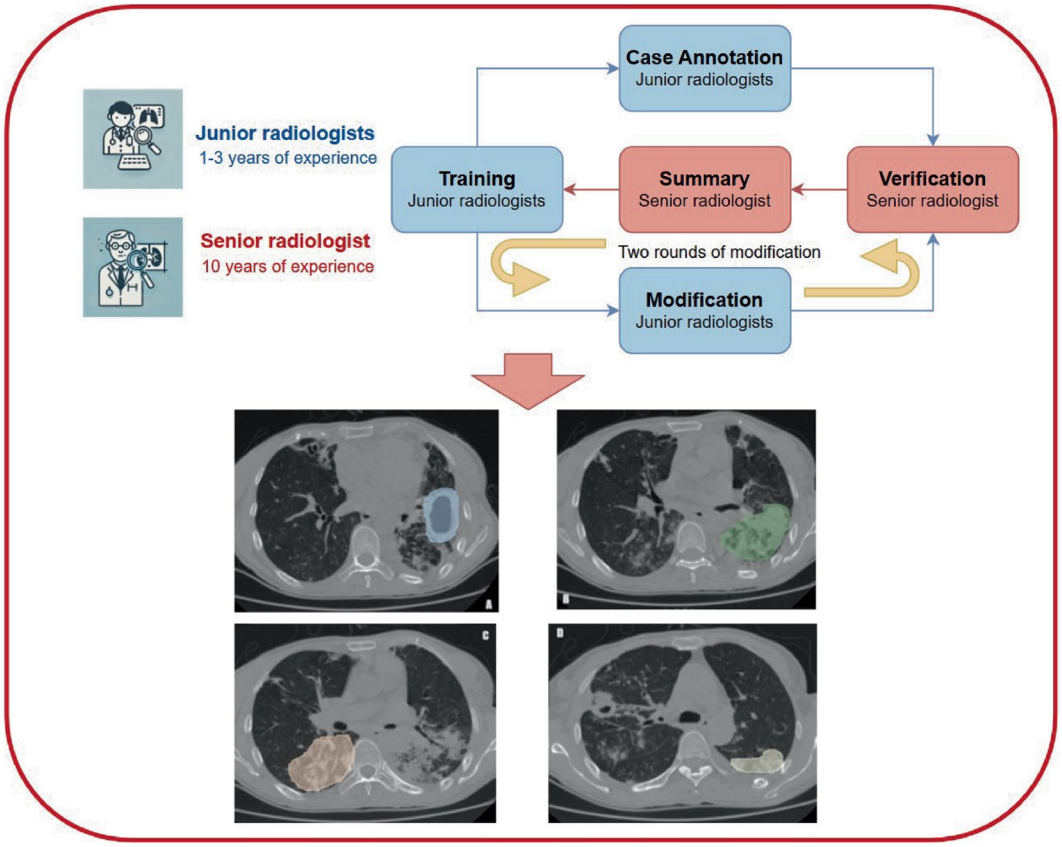

Fig. 2 Annotation Workflow and Final Annotation Results for Different Lesions. A: Cavities B: Consolidation C: Tree-in-Bud Sign D: Ground-Glass Opacities (GGO).

retained only the highest-resolution thin-slice sequences with a thickness of 1.0 to 1.25 mm to ensure optimal data quality.

We utilized pathogen microbiological examination as the diagnostic basis for NTM and TB infections. Sputum samples were collected and subjected to acid-fast bacilli (AFB) staining, followed by mycobacterial culture using Löwenstein-Jensen (LJ) medium. NTM species identification was performed using matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS).Specifically, a diagnosis of NTM pulmonary disease was based on the Treatment of Nontuberculous Mycobacterial Pulmonary Disease: An Official ATS/ERS/ESCMID/IDSA Clinical Practice Guideline[23], which requires patients to meet both clinical and microbiological criteria. This includes characteristic clinical and radiographic manifestations and the repeated isolation of the same NTM species from at least two separate sputum cultures. Similarly, TB diagnosis was confirmed through sputum culture, the National Health Commission of the People's Republic of China. In: diagnostic Criteria for Pulmonary Tuberculosis (WS 288–2017)[46].

Data annotation. Lesion annotation was conducted using 3D-Slicer software (version 310.2, http://www.slicer.org/). Seven radiologists were involved in this task. To ensure accuracy, six radiologists with 1–3 years of experience underwent a training process involving trial and error. For consistency in annotation, a senior radiologist with 10 years of experience reviewed all annotations without access to clinical information. This senior radiologist also served as a supervisor, identifying errors and highlighting important considerations in the annotation process. Junior radiologists were given an opportunity to correct their mistakes based on this feedback. The entire process comprised two rounds of training, two rounds of case annotation, two rounds of corrections, two rounds of feedback summarization, and three rounds of verification to ensure that the annotations were accurate and suitable for subsequent analysis, as shown in Figure 2.

Data standardization. Our standardization process includes the following steps: First, format conversion is performed by converting multi-slice DICOM files to NIfTI format. Next, we resample the image dimensions to a consistent voxel size of 1 mm × 1 mm × 1 mm. Pixel value normalization is then applied, restricting HU values to the lung window range (window level: −600, window width: 1500) and scaling values within this range to [0,1] to meet the input requirements of deep learning models. Additionally, we standardize the naming of NTM and TB cases and consolidate clinical information, species data, and annotation details into an index file for uniform storage.

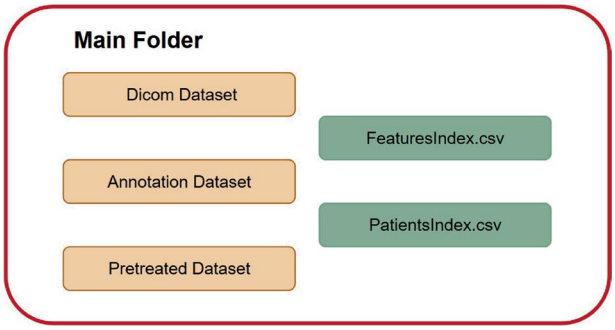

Fig. 3 Schematic Diagram of the Dataset Organization Structure.

Data Records

The dataset is publicly available on Kaggle[47,48], an open-access platform that does not require any password for entry. Detailed instructions on how to access the mycobacterial dataset can be found at the following links: https://doi.org/10.34740/kaggle/dsv/10818141 and https://doi.org/10.34740/kaggle/ds/6248246. All versions of the dataset are released under the Creative Commons Attribution (CC-BY) license.

Based on the characteristics of this dataset and the need to enhance model training efficiency, we have provided the original single-slice CT images of NTM and TB cases in DICOM (.dcm) format, stored in the "NTM" and "TB" folders within the "Dicom Dataset" respectively. The preprocessed 3D medical image files are organized in the "Pretreated Dataset" subset, while the annotated lesion files are located in the "Annotation Dataset" folder, as illustrated in Fig. 3. Within each folder, cases follow a consistent naming convention to ensure easy access and organization.

The "PatientsIndex.csv" file is an index of case information, containing two worksheets for NTM and TB. Each worksheet records patient details such as gender, age, the presence of specific clinical symptoms (indicated by 1), and whether lesion annotations were performed (indicated by 1). In the NTM worksheet, additional species subtype information is provided for some cases.

The "FeaturesIndex.csv" file serves as an index for lesion annotation information and also contains two worksheets for NTM and TB. Each case corresponds to one or more lesion identification files within the case directory, listing the label information for these lesions.

Technical Validation

Model configuration environment. In this study, we selected widely used classification models, such as ResNet101[49], BoTNet50[50], EfficientNetB4[51], and DenseNet121[52], as baseline methods for similar applications. Prior to benchmark testing, we performed standardized preprocessing and augmentation on the data. The training and testing configurations for the models followed the default settings recommended in published papers or open-source code repositories to ensure the reliability of performance evaluation. All experiments were conducted using the PyTorch framework framework[47] and executed on an NVIDIA RTX4090 GPU.

Image preprocessing. One of the main challenges faced by this dataset is class imbalance, with a significantly larger number of TB cases compared to NTM cases. This imbalance can lead to model overfitting during training and testing, manifesting as an excessive focus on the TB class and insufficient predictive ability for the NTM class. To address this issue, we adopted a class-balanced training strategy and implemented various data preprocessing and augmentation techniques. By applying random transformations to the training data, we enhanced its diversity, enabling the model to learn richer feature representations. Specific operations included random flipping and 90-degree rotation, aimed at improving the model's invariance to image orientation changes, thereby increasing its robustness. Random contrast adjustments and the addition of Gaussian noise introduced varying levels of image changes, allowing the model to learn a broader range of image features. These augmentation techniques effectively expanded the training set, alleviated the class imbalance problem, and improved the model's ability to capture features from different classes, ultimately enhancing the prediction performance for NTM samples.

Baseline classification. Table 2 and Figs. 4, 5 present the performance evaluation results of four models (ResNet101, BoTNet50, EfficientNetB4, and DenseNet121), including key metrics such as Accuracy, Precision, Recall, F1 Score, MCC (Matthews Correlation Coefficient), and AUC (Area Under the Curve). The excellent performance of BoTNet50 can be attributed to its combination of convolutional networks and self-attention mechanisms, allowing it to better capture both global and local features. In contrast, ResNet101's traditional architecture struggled with capturing complex features and addressing data imbalance, which may be the main reason for its lower performance. EfficientNetB4 and DenseNet121 exhibited intermediate performance, reflecting the importance of efficient network depth scaling and dense connections in classification tasks. However, while EfficientNetB4 had an advantage in Recall, its Precision and AUC were slightly inferior to BoTNet50. DenseNet121 achieved higher Precision but still did not outperform BoTNet50 in overall metrics such as F1 Score and MCC. These results further highlight the models' limitations in addressing class imbalance and optimizing the classification of positive and negative samples. Although BoTNet50 has significantly improved classification

Model	Accuracy	Precision	Recall	F1 Score	MCC	AUC
ResNet101	0.744	0.524	0.512	0.518	0.251	0.61
BoTNet50	0.860	0.700	0.651	0.675	0.593	0.71
EfficientNetB4	0.789	0.591	0.605	0.598	0.394	0.68
DenseNet121	0.774	0.556	0.581	0.568	0.353	0.66

Table 2. Performance Results of Various Benchmark Models.

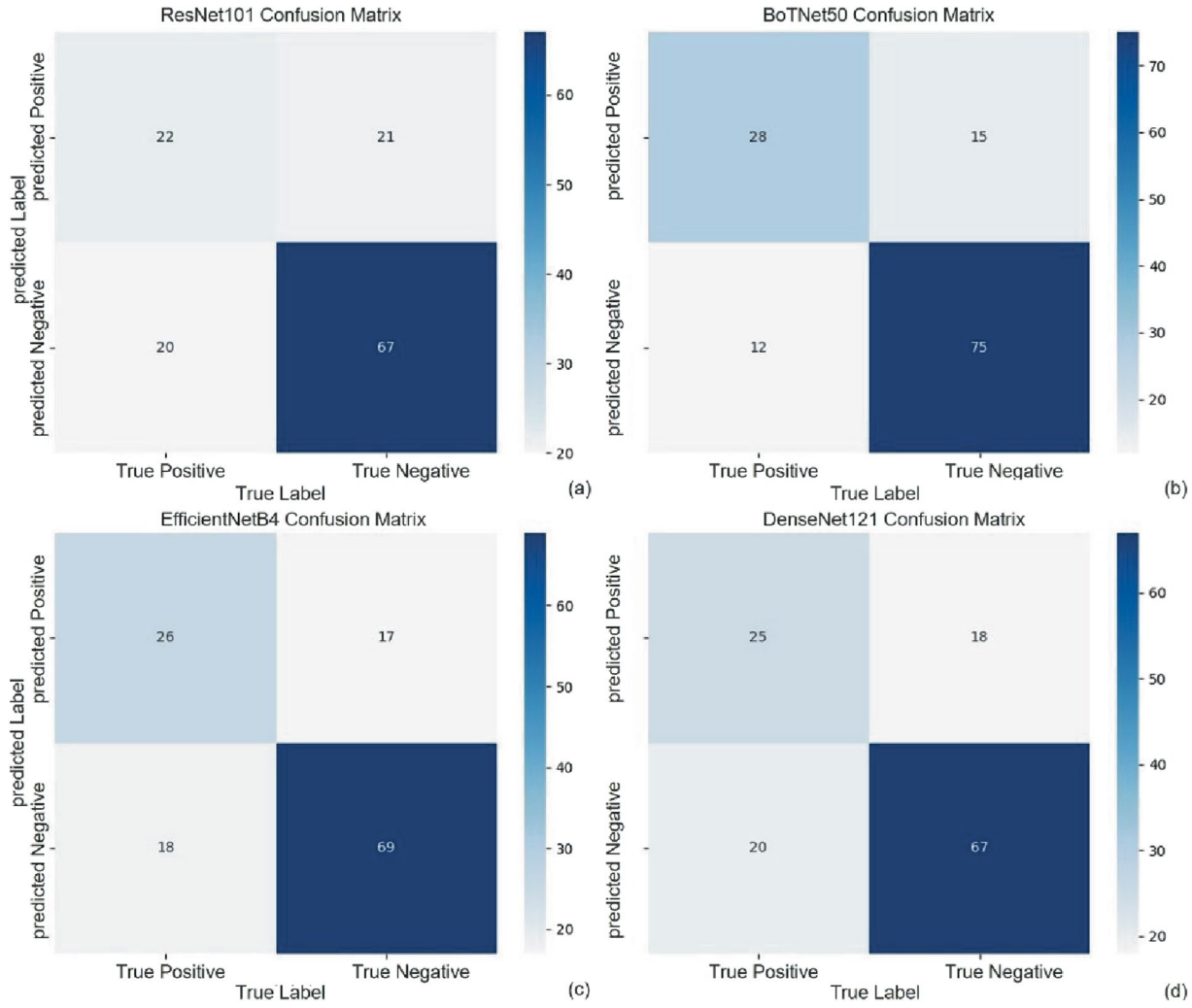

Fig. 4 Confusion Matrices of Several Benchmark Models, Showing Lower Classification Accuracy for Positive Samples Compared to Negative Samples.

performance for positive samples, the overall AUC and MCC levels suggest that there is still room for improvement in the differentiation between positive and negative samples, and the model's performance has potential for further optimization.

Usage Notes

Artificial intelligence (AI) methods, especially deep learning based on neural networks, have played an important role in advancing the timely diagnosis of nontuberculous mycobacteria (NTM). Our dataset aims to facilitate the application of AI methods in NTM-related tasks, specifically in the following areas:

Binary classification of pulmonary mycobacterial diseases. This dataset includes 430 NTM cases and 871 TB cases, making it suitable for binary classification tasks between NTM and TB.

Multi-class classification of mycobacterial subtypes. Comprising 308 cases, this dataset primarily includes four species: *Mycobacterium intracellulare*, *Mycobacterium abscessus*, *Mycobacterium kansasii*, and *Mycobacterium avium*. It supports multi-class classification tasks involving different mycobacterial species.

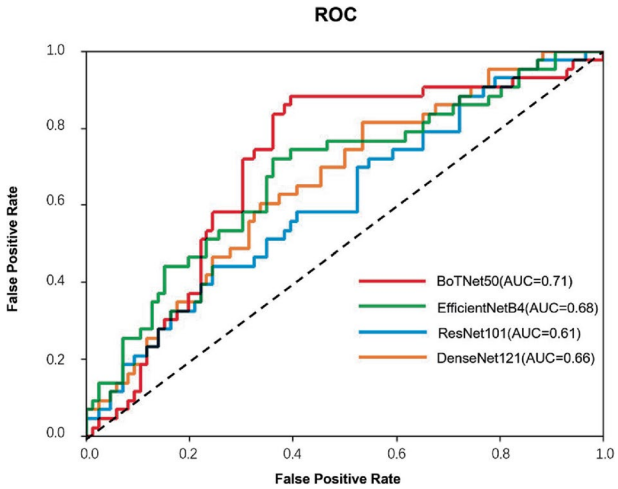

Fig. 5 Receiver Operating Characteristic (ROC) Curves of Four Models on the Test Set.

Segmentation and identification of lung lesions. This dataset contains 240 annotated cases, with lesions marked for 120 NTM and 120 TB cases. Radiologists have labeled lesions such as ground-glass opacities (GGO), bronchiectasis, consolidation, and the tree-in-bud sign, providing valuable data for training segmentation models.

Through these datasets, we aim to promote AI applications in the NTM field, fostering more efficient disease diagnosis and management.

Limitations

The data for this study were obtained from chest CT scans of patients at Tianjin Haihe Hospital in China, representing a subset of the Asian population and potentially not fully capturing the imaging characteristics of all patients with nontuberculous mycobacteria (NTM) infections. Since all data were collected from a single hospital, this may limit the model's applicability to data from other healthcare centers. Additionally, the dataset spans a significant time period and includes scans from different types of imaging devices, which may result in variations in scanning and display parameters among some original images. However, we randomized the order of data labeling to increase the internal distribution randomness across multiple devices and time periods within the dataset. The absence of data on other types of respiratory infections in this study also represents a limitation, as it restricts the generalizability of our findings to broader diagnostic scenarios.

Our future research will prioritize systematic dataset expansion through three synergistic dimensions: diversifying patient demographics and multi-center imaging protocols to enhance population representativeness, expanding lesion annotations to enhance annotation diversity, and integrating additional pulmonary diseases (e.g., bacterial pneumonia, fungal infections) to address critical challenges in differentiating NTM from TB and other respiratory pathologies.

Code availability

The code for image standardization and preprocessing has been released along with the dataset and can be accessed through our Kaggle dataset link.

Received: 16 December 2024; Accepted: 14 March 2025;
Published online: 29 March 2025

References

1. Chakaya, J. *et al.* Global Tuberculosis Report 2020 - Reflections on the Global TB burden, treatment and prevention efforts. *Int J Infect Dis* **113** (Suppl 1), S7–S12, https://doi.org/10.1016/j.ijid.2021.02.107 (2021).
2. Fowler, S. *et al.* Nontuberculous mycobacteria in bronchiectasis: Prevalence and patient characteristics. *European Respiratory Journal* **28**, 1204–1210 (2006).
3. Prevots, D. R. & Marras, T. K. Epidemiology of human pulmonary infection with nontuberculous mycobacteria: a review. *Clinics in chest medicine* **36**, 13–34 (2015).
4. Ratnatunga, C. N. *et al.* The rise of non-tuberculosis mycobacterial lung disease. *Frontiers in immunology* **11**, 303 (2020).
5. Donohue, M. J. Increasing nontuberculous mycobacteria reporting rates and species diversity identified in clinical laboratory reports. *BMC Infect. Dis.* **18**, 1–9 (2018).
6. Cowman, S., van Ingen, J., Griffith, D. E. & Loebinger, M. R. Non-tuberculous mycobacterial pulmonary disease. *Eur Respir J* **54** https://doi.org/10.1183/13993003.00250-2019 (2019).
7. Lee, H., Myung, W., Koh, W. J., Moon, S. M. & Jhun, B. W. Epidemiology of Nontuberculous Mycobacterial Infection, South Korea, 2007-2016. *Emerg Infect Dis* **25**, 569–572, https://doi.org/10.3201/eid2503.181597 (2019).
8. Winthrop, K. L. *et al.* Incidence and Prevalence of Nontuberculous Mycobacterial Lung Disease in a Large US Managed Care Health Plan, 2008-2015. *Ann. Am. Thoracic Society* **17**, 178–185, https://doi.org/10.1513/AnnalsATS.201804-236OC (2020).
9. Dong, Z. *et al.* Age-period-cohort analysis of pulmonary tuberculosis reported incidence, China, 2006-2020. *Infect. Dis. Poverty* **11**, 10, https://doi.org/10.1186/s40249-022-01009-4 (2022).

10. Xu, D., Han, C., Wang, M. S. & Wang, J. L. Increasing prevalence of non-tuberculous mycobacterial infection from 2004-2009 to 2012-2017: A laboratory-based surveillance in China. *J. Infect.* **76**, 422–424, https://doi.org/10.1016/j.jinf.2017.12.007 (2018).
11. Dahl, V. N. *et al.* Global trends of pulmonary infections with nontuberculous mycobacteria: a systematic review. *Int. J. Infect. Dis.* **125**, 120–131, https://doi.org/10.1016/j.ijid.2022.10.013 (2022).
12. Brode, S. K., Daley, C. L. & Marras, T. K. The epidemiologic relationship between tuberculosis and non-tuberculous mycobacterial disease: a systematic review. *Int. J. Tuberc. Lung Dis.* **18**, 1370–1377, https://doi.org/10.5588/ijtld.14.0120 (2014).
13. Grigg, C. *et al.* Epidemiology of Pulmonary and Extrapulmonary Nontuberculous Mycobacteria Infections at 4 US Emerging Infections Program Sites: A 6-Month Pilot. *Clinical Infectious Diseases* **77**, 629–637, https://doi.org/10.1093/cid/ciad214 (2023).
14. Haworth, C. S. *et al.* British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD). *Thorax* **72**, ii1–ii64, https://doi.org/10.1136/thoraxjnl-2017-210927 (2017).
15. Kim, Y. K. *et al.* Comparable characteristics of tuberculous and non-tuberculous mycobacterial cavitary lung diseases. *Int. J. Tuberc. Lung Dis.* **18**, 725–729, https://doi.org/10.5588/ijtld.13.0871 (2014).
16. Maiga, M. *et al.* Failure to recognize nontuberculous mycobacteria leads to misdiagnosis of chronic pulmonary tuberculosis. *PLoS One* **7**, e36902, https://doi.org/10.1371/journal.pone.0036902 (2012).
17. Tortoli, E. The new mycobacteria: an update. *FEMS Immunol Med Microbiol* **48**, 159–178, https://doi.org/10.1111/j.1574-695X.2006.00123.x (2006).
18. Griffith, D. E. & Daley, C. L. Treatment of Mycobacterium abscessus Pulmonary Disease. *Chest* **161**, 64–75, https://doi.org/10.1016/j.chest.2021.07.035 (2022).
19. Gupta, R. S., Lo, B. & Son, J. Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus Mycobacterium into an Emended Genus Mycobacterium and Four Novel Genera. *Front Microbiol* **9**, 67, https://doi.org/10.3389/fmicb.2018.00067 (2018).
20. Johansen, M. D., Herrmann, J. L. & Kremer, L. Non-tuberculous mycobacteria and the rise of Mycobacterium abscessus. *Nat Rev Microbiol* **18**, 392–407, https://doi.org/10.1038/s41579-020-0331-1 (2020).
21. Ryu, Y. J., Koh, W.-J. & Daley, C. L. Diagnosis and treatment of nontuberculous mycobacterial lung disease: clinicians' perspectives. *Tuberculosis and respiratory diseases* **79**, 74–84 (2016).
22. Hoefsloot, W. *et al.* The geographic diversity of nontuberculous mycobacteria isolated from pulmonary samples: an NTM-NET collaborative study. *Eur Respir J* **42**, 1604–1613, https://doi.org/10.1183/09031936.00149212 (2013).
23. Daley, C. L. *et al.* Treatment of Nontuberculous Mycobacterial Pulmonary Disease: An Official ATS/ERS/ESCMID/IDSA Clinical Practice Guideline. *Clin Infect Dis* **71**, 905–913, https://doi.org/10.1093/cid/ciaa1125 (2020).
24. Gould, M. K. *et al.* Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **143**, e93S–e120S, https://doi.org/10.1378/chest.12-2351 (2013).
25. Li, Y. & Xia, L. M. Coronavirus Disease 2019 (COVID-19): Role of Chest CT in Diagnosis and Management. *Am. J. Roentgenol.* **214**, 1280–1286, https://doi.org/10.2214/ajr.20.22954 (2020).
26. Lee, A.-R. *et al.* Phenotypic, immunologic, and clinical characteristics of patients with nontuberculous mycobacterial lung disease in Korea. *BMC Infect. Dis.* **13**, 1–7 (2013).
27. Im, Y. *et al.* Impact of Time Between Diagnosis and Treatment for Nontuberculous Mycobacterial Pulmonary Disease on Culture Conversion and All-Cause Mortality. *Chest* **161**, 1192–1200, https://doi.org/10.1016/j.chest.2021.10.048 (2022).
28. Johnson, M. M. & Odell, J. A. Nontuberculous mycobacterial pulmonary infections. *J Thorac Dis* **6**, 210–220, https://doi.org/10.3978/j.issn.2072-1439.2013.12.24 (2014).
29. Burrill, J. *et al.* Tuberculosis: a radiologic review. *Radiographics* **27**, 1255–1273 (2007).
30. Chang, C. Y. & Chang, S. C. Comparative chest computed tomography findings of non-tuberculous mycobacterial lung diseases and pulmonary tuberculosis in patients with afb smear-positive sputum. *Respirology* **18**, 122–122 (2013).
31. Chu, H.-Q. *et al.* Chest imaging comparison between non-tuberculous and tuberculosis mycobacteria in sputum acid fast bacilli smear-positive patients. *European Review for Medical & Pharmacological Sciences* **19** (2015).
32. Kahkouee, S. *et al.* Multidrug resistant tuberculosis versus non-tuberculous mycobacterial infections: a CT-scan challenge. *Braz. J. Infect. Dis.* **17**, 137–142, https://doi.org/10.1016/j.bjid.2012.10.011 (2013).
33. Kwak, N. *et al.* Non-tuberculous mycobacterial lung disease: diagnosis based on computed tomography of the chest. *European Radiology* **26**, 4449–4456, https://doi.org/10.1007/s00330-016-4286-6 (2016).
34. Matveychuk, A., Fuks, L., Priess, R., Hahim, I. & Shitrit, D. Clinical and radiological features of Mycobacterium kansasii and other NTM infections. *Respiratory Medicine* **106**, 1472–1477, https://doi.org/10.1016/j.rmed.2012.06.023 (2012).
35. Li, C. *et al.* Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Computer methods and programs in biomedicine* **153**, 211–225 (2018).
36. Wu, Y. H. *et al.* JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. *IEEE Trans Image Process* **30**, 3113–3126, https://doi.org/10.1109/TIP.2021.3058783 (2021).
37. Shi, F. *et al.* Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. *IEEE Rev Biomed Eng* **14**, 4–15, https://doi.org/10.1109/RBME.2020.2987975 (2021).
38. Albahri, O. S. *et al.* Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects. *J Infect Public Health* **13**, 1381–1396, https://doi.org/10.1016/j.jiph.2020.06.028 (2020).
39. Xing, Z. *et al.* Machine Learning-Based Differentiation of Nontuberculous Mycobacteria Lung Disease and Pulmonary Tuberculosis Using CT Images. *Biomed Res Int* **2020**, 6287545, https://doi.org/10.1155/2020/6287545 (2020).
40. Wang, L. *et al.* Distinguishing nontuberculous mycobacteria from Mycobacterium tuberculosis lung disease from CT images using a deep learning framework. *European Journal of Nuclear Medicine and Molecular Imaging* **48**, 4293–4306, https://doi.org/10.1007/s00259-021-05432-x (2021).
41. Ying, C. *et al.* T-SPOT with CT image analysis based on deep learning for early differential diagnosis of nontuberculous mycobacteria pulmonary disease and pulmonary tuberculosis. *Int. J. Infect. Dis.* **125**, 42–50, https://doi.org/10.1016/j.ijid.2022.09.031 (2022).
42. Lee, S. *et al.* Deep Learning-Based Prediction Model Using Radiography in Nontuberculous Mycobacterial Pulmonary Disease. *Chest* **162**, 995–1005, https://doi.org/10.1016/j.chest.2022.06.018 (2022).
43. Liu, C.-J. *et al.* A deep learning model using chest X-ray for identifying TB and NTM-LD patients: a cross-sectional study. *Insights Imaging* **14** https://doi.org/10.1186/s13244-023-01395-9 (2023).
44. Park, M. *et al.* Distinguishing nontuberculous mycobacterial lung disease and Mycobacterium tuberculosis lung disease on X-ray images using deep transfer learning. *BMC Infect. Dis.* **23** https://doi.org/10.1186/s12879-023-07996-5 (2023).
45. Yoon, I. *et al.* Mycobacterial cavity on chest computed tomography: clinical implications and deep learning-based automatic detection with quantification. *Quantitative Imaging in Medicine and Surgery* **13**, 747-+, https://doi.org/10.21037/qims-22-620 (2023).
46. National Health Commission of the People's Republic of China. Diagnostic Criteria for Pulmonary Tuberculosis (WS 288-2017). Available at: http://www.nhc.gov.cn/ewebeditor/uploadfile/2017/11/20171128164254246.pdf (2017).
47. Han, Z. L. *et al.* Mycobacterial CT Imaging Dataset. *Kaggle* https://doi.org/10.34740/kaggle/dsv/10818141 (2025).
48. Han, Z. L. *et al.* Mycobacterial CT Imaging Dataset. *Kaggle* https://doi.org/10.34740/kaggle/ds/6248246 (2025).

49. He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J. & Ieee. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778 (Ieee, 2016).
50. Srinivas, A. *et al*. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16514–16524 (Ieee Computer Soc, 2021).
51. Tan, M. & Le, Q. in *International conference on machine learning*. 6105–6114 (PMLR).
52. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

Acknowledgements

This study was supported by the following grants and funding sources: the Tianjin Science and Technology Project, "Research on NTM Diagnosis Applications Based on CT-Labeled Big Data Resources" (Project No. 21JCYBJC00510); the Tianjin Haihe Hospital Science and Technology Fund Project, "AI-Assisted NTM-ID Diagnosis Applications Based on CT-Labeled Big Data Resources" (Project No. HHYY-202007); and the Tianjin Key Medical Discipline (Specialty) Construction Projects (Project Nos. TJYXZDXK067C and TJYXZDXK063B).

Author contributions

Zhilin Han: Responsible for data organization, analysis, and model testing; wrote the manuscript and summarized the results; Yuyang Zhang: Responsible for segmentation annotation and assisted with data collection, assisted with manuscript revisions; Wenlong Ding: Participated in data organization and segmentation annotation; assisted with manuscript revisions; Xiaoting Zhao: Responsible for segmentation annotation and assisted with data collection; Bingzhen Jia: Responsible for segmentation annotation and assisted with data collection; Tingting Liu: Responsible for segmentation annotation and assisted with data collection; Liang Wan: Supervised the research and participated in the study design; Zhiheng Xing: Supervised the research, responsible for segmentation annotation oversight, study design, and manuscript revisions.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.W. or Z.X.