Therapeutics and Clinical Risk Management

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Artificial Intelligence in the Diagnosis and Management of Pulmonary Tuberculosis: A Review of Current Applications and Future Perspectives

Yangyu Ou[1], Qi Zhang[2], Cheng Meng[3], Xingshun Zhou[4], Mengyao Na[1], Zhenni Yu[4], Wenhui Ma[4], Cong Huang[4]

[1]Department of Infectious, No. 926 Hospital, Joint Logistics Support Force of PLA, Kaiyuan, Yunnan, 661699, People's Republic of China; [2]Department of Cardiothoracic Surgery, No. 926 Hospital, Joint Logistics Support Force of PLA, Kaiyuan, Yunnan, 661699, People's Republic of China; [3]Department of Laboratory, No. 926 Hospital, Joint Logistics Support Force of PLA, Kaiyuan, Yunnan, 661699, People's Republic of China; [4]Department of Radiology, No. 926 Hospital, Joint Logistics Support Force of PLA, Kaiyuan, Yunnan, 661699, People's Republic of China

Correspondence: Wenhui Ma; Cong Huang, Email 103616150@qq.com; magichc401@163.com

Abstract: Pulmonary tuberculosis (PTB) remains a major global public health challenge, with traditional diagnostic and management approaches plagued by diagnostic delays, time-consuming drug resistance testing, and subjective efficacy assessment. Artificial intelligence (AI) has emerged as a revolutionary solution for these bottlenecks. This review comprehensively summarizes AI's current applications in PTB care: deep learning models enable automated detection, segmentation and activity differentiation of PTB lesions on chest X-ray/CT with performance comparable to or exceeding human experts (sensitivity >90% for X-ray detection); AI-driven whole-genome sequence analysis rapidly predicts anti-TB drug resistance, shortening testing time from weeks to days; multimodal AI models also show potential in dynamic treatment response monitoring and individualized outcome prediction. However, AI's clinical translation is hindered by data quality/bias, poor model generalizability, low algorithm interpretability, and regulatory/ethical issues. Future priorities include multimodal data fusion, federated learning, prospective clinical validation, and developing lightweight AI models for resource-limited settings. Interdisciplinary collaboration is critical to transform AI from a research tool into a safe, reliable and equitable clinical assistant for PTB care.

Keywords: artificial intelligence, pulmonary tuberculosis, deep learning, diagnosis, drug resistance, radiomics, clinical decision support systems

Introduction

Pulmonary Tuberculosis (PTB), caused by *Mycobacterium tuberculosis*, remains one of the most significant global public health challenges. According to the latest World Health Organization (WHO) report, there are over ten million new PTB cases annually worldwide, resulting in more than one million deaths, with the disease burden being particularly heavy in low- and middle-income countries.[1] Although "ending the tuberculosis epidemic" has become one of the global sustainable development goals, current PTB diagnosis and management practices still face multiple severe challenges.

These challenges persist throughout the entire disease management pathway. In the diagnostic phase, traditional microbiological methods, such as sputum smear microscopy, exhibit low sensitivity, while bacterial culture—the "gold standard"—requires 2–8 weeks, leading to diagnostic delays and missed opportunities for early treatment.[2] In treatment management, the prevalence of Multidrug-Resistant TB (MDR-TB) makes rapid and accurate Drug Susceptibility Testing (DST) crucial; however, phenotypic DST also involves prolonged turnaround times. Additionally, the assessment of treatment efficacy largely relies on serial chest imaging, yet the interpretation of imaging findings is prone to the subjective experience of radiologists, lacking objective and quantifiable standards.[3] These bottlenecks collectively hinder effective PTB control, creating an urgent need for innovative technologies to optimize the diagnostic and therapeutic pathway.

https://doi.org/10.2147/TCRM.S596651
Received: 16 January 2026
Accepted: 15 March 2026
Published: 31 March 2026

In recent years, Artificial Intelligence (AI), particularly its subfields of Machine Learning (ML) and Deep Learning (DL), has achieved groundbreaking progress in the medical field. DL models, such as Convolutional Neural Networks (CNNs), have demonstrated performance surpassing human experts in analyzing high-dimensional, complex medical imaging data and have been successfully applied in various scenarios, including pulmonary nodule screening and diabetic retinopathy diagnosis.[4] While these advancements are promising, a critical caveat is that many high-performance AI models are validated on curated research datasets, leading to risks of overfitting—where models perform well on training data but fail to generalize to real-world clinical data—and non-trivial false positive rates that are often underreported in initial studies. Simultaneously, AI has shown great potential in genomic data analysis, clinical prediction model construction, and mining insights from electronic health record information through Natural Language Processing (NLP). This powerful data-driven analytical capability provides unprecedented opportunities to systematically address the aforementioned challenges in PTB diagnosis and management, but only if the inherent risks of overfitting and false positives, and their subsequent clinical consequences, are rigorously evaluated and mitigated.

Based on this, this review aims to comprehensively synthesize and critically evaluate the current applications and future prospects of AI technologies in PTB diagnosis and management. We will focus on the following core questions: How effective is AI in automatically interpreting Chest X-ray (CXR) and Computed Tomography (CT) images to achieve accurate diagnosis and differential diagnosis? What are the false positive rates of AI models in real-world TB detection, and what clinical consequences do these errors impose? How can AI models utilize genomic data to accelerate drug resistance prediction? How can overfitting risks be minimized in AI-driven genomic and imaging models for TB care? What are the applications of AI in quantifying treatment response and predicting individualized treatment outcomes? Finally, this review will delve into the data, algorithmic, and ethical challenges faced during the translation of AI technologies into clinical practice—including overfitting, false positives, and their clinical impacts—and propose future research directions, aiming to provide valuable references for researchers, clinicians, and public health policymakers (Figure 1).

AI in the Radiological Diagnosis of Tuberculosis

Medical imaging, particularly chest X-ray (CXR) and computed tomography (CT), serves as a cornerstone for tuberculosis (TB) screening, diagnosis, and treatment response assessment. However, image interpretation is highly dependent on the expertise of radiologists, posing challenges such as subjectivity, high workload, and inter-observer variability. AI, especially DL technologies, offers a revolutionary solution for the automated and standardized analysis of pulmonary images, and has increasingly become a powerful auxiliary tool for radiologists.

Technical Basis

Current AI technologies applied to TB image analysis are predominantly based on DL models under the supervised learning paradigm, with convolutional neural networks (CNNs) and their variants forming the technical core (Figure 2).[4] CNNs can automatically extract hierarchical features from pixel-level data, ranging from low-level to high-level patterns, through multiple convolutional and pooling operations, thereby capturing complex imaging patterns. For instance, architectures such as ResNet and DenseNet address the vanishing gradient problem in deep networks, enabling deeper architectures and superior performance. Deeper network architectures, however, increase the risk of overfitting to specific training image features—such as scanner-specific artifacts or population-specific imaging patterns—rather than general TB lesion characteristics, which is a critical technical limitation that must be addressed via regularization techniques, data augmentation, and large, diverse training datasets. For pixel-level precise analysis tasks (eg, lesion segmentation), encoder-decoder architectures like U-Net are widely adopted due to their excellent performance in biomedical image segmentation. These models are typically trained on large datasets of expert-annotated images (eg, delineated lesion areas, diagnostic classifications) to learn the characteristic features of TB and distinguish them from normal anatomical structures or other pulmonary diseases, yet annotation biases in training data can further exacerbate false positive rates when models encounter atypical imaging findings in clinical practice.

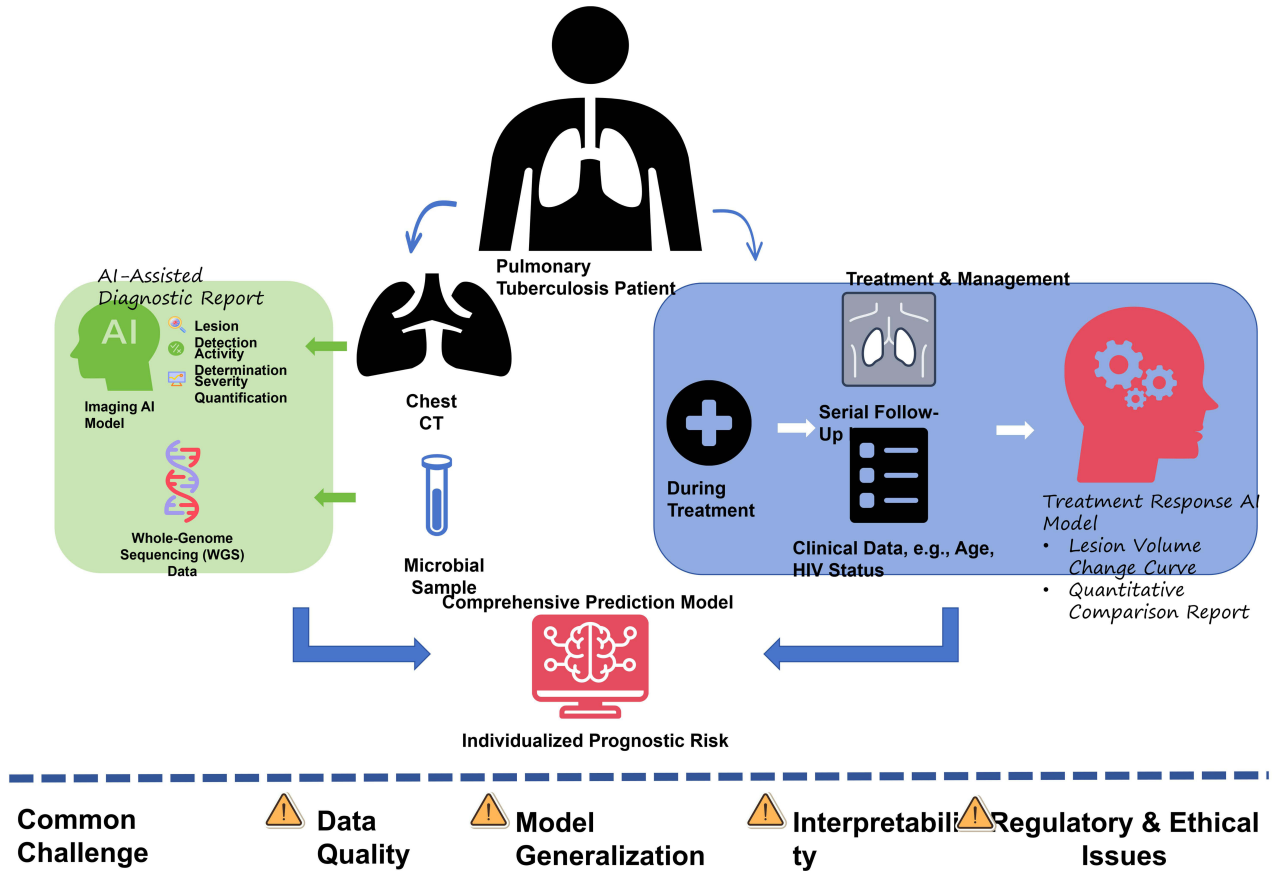

Figure 1 Schematic of an AI-integrated workflow for pulmonary tuberculosis, incorporating AI-assisted diagnosis from chest CT and microbial sequencing, treatment response monitoring via serial imaging and clinical data, and a central model for individualized prognostic risk prediction. Key implementation challenges include data quality, model generalization, interpretability, and regulatory compliance.

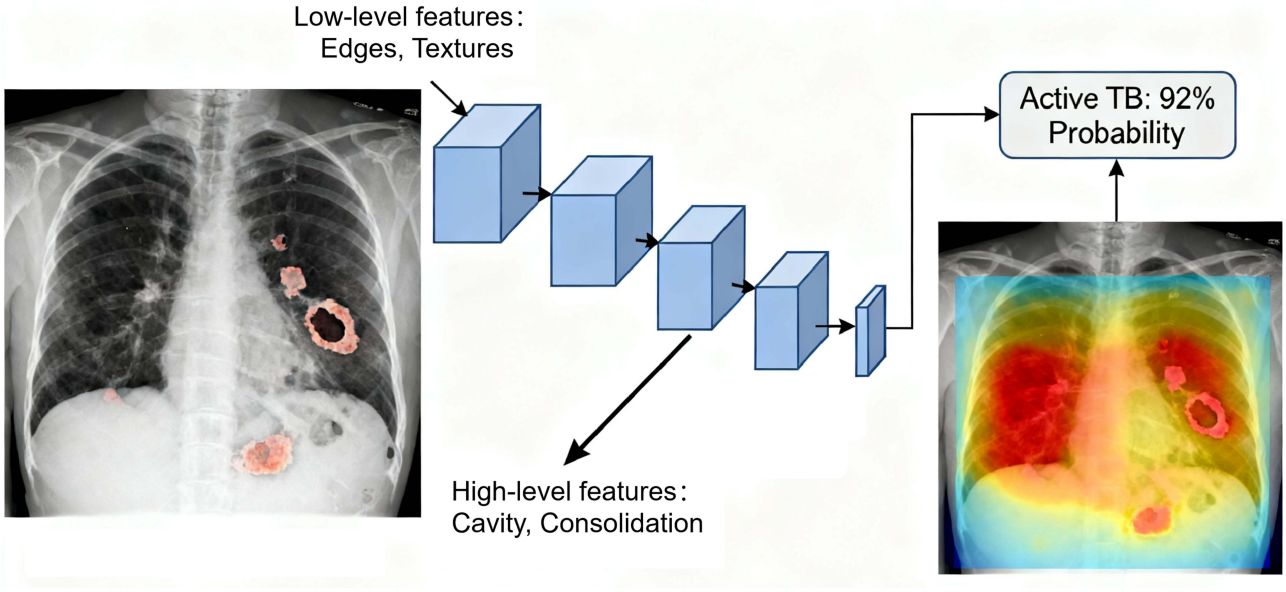

Figure 2 An illustration of a convolutional neural network (CNN) for pulmonary tuberculosis detection. The model extracts hierarchical features from input images to generate predictions and visual explanations (heatmaps).

https://doi.org/10.2147/TCRM.S596651

Specific Application Scenarios

Lesion Detection and Segmentation

A core capability of AI models is the automatic detection and precise delineation (segmentation) of TB-related imaging abnormalities. On CXR, models can identify various lesions such as nodules, infiltrates, consolidations, cavities, and fibrotic streaks.[5] On higher-resolution CT images, AI's segmentation capability is even more pronounced, enabling accurate quantification of morphological features of micronodules, tree-in-bud signs, ground-glass opacities, and cavities.[6] Despite this precision, AI models often generate false positive lesion detections—for example, misclassifying benign pulmonary opacities (eg, atelectasis, pneumoconiosis, or normal anatomical variants) as TB lesions—with such errors carrying direct clinical consequences, including unnecessary follow-up imaging, invasive diagnostic procedures (eg, bronchoscopy), and unwarranted antibiotic use, particularly in low-TB-prevalence regions. This functionality not only significantly reduces the workload of radiologists in initial screening but, more importantly, provides objective and reproducible quantitative data on lesions, laying a solid foundation for subsequent treatment response assessment, provided that false positive detections are minimized through rigorous model optimization.

Activity Determination

Distinguishing active TB from old fibrocalcific lesions and differentiating it from other pulmonary diseases (eg, pneumonia, lung cancer) are common challenges in clinical practice. DL models demonstrate great potential in this area by identifying subtle texture features and distribution patterns that are difficult for the human eye to perceive. Studies have shown that trained CNN models can differentiate active from inactive TB with high accuracy, performing comparably to or even surpassing experienced radiologists.[3,7] However, overfitting to training data with clear distinctions between active and inactive lesions can lead to misclassification in real-world cases with borderline or atypical imaging features, while false positive classifications of inactive lesions as active TB can result in unnecessary anti-TB treatment, with associated risks of drug toxicity, antimicrobial resistance development, and patient psychological distress. This is crucial for achieving rapid triage and initiating treatment in high TB burden areas. Notably, in the differentiation between TB and lung cancer, AI models can identify subtle differences in lesion edge characteristics and internal density distribution that are easily overlooked by humans, further improving the accuracy of differential diagnosis,[8] though false positive TB diagnoses in lung cancer patients can delay oncological treatment, representing a severe clinical consequence of AI diagnostic error.

Severity Quantification

Beyond qualitative diagnosis, AI enables the objective quantitative assessment of treatment response. By comparing changes in lesion volume, density, or cavity size on serial images before and after treatment, AI can generate precise quantitative reports, replacing traditional subjective descriptive terms such as "improved" or "resolved".[9] Overfitting to baseline lesion features in training data can compromise the model's ability to accurately quantify subtle treatment-related changes, leading to false positive assessments of treatment failure—where stable lesions are misclassified as progressive—resulting in unnecessary regimen changes and increased healthcare costs. This dynamic monitoring capability provides powerful imaging biomarkers for implementing personalized treatment, evaluating new drug efficacy, and predicting treatment outcomes, only when models are validated for generalizability and false positive risk is quantified.

Performance Evaluation

Numerous studies have validated the diagnostic performance of AI-driven models (Table 1). A review indicated that DL-based CXR screening models demonstrate high pooled sensitivity (>90%) and specificity (>80%) for detecting TB, with areas under the curve (AUC) often exceeding 0.95.[10] Critically, these metrics are often derived from internal validation or highly curated external datasets, and real-world false positive rates are frequently higher—ranging from 5% to 20% in recent clinical implementation studies—due to unaccounted for variability in image quality, patient comorbidities, and imaging artifacts. In some head-to-head comparisons, specific AI systems have outperformed the majority of junior or intermediate-level radiologists.[3] However, performance varies across different populations, equipment types, and epidemiological settings, highlighting the critical importance of external validation, as well as the need to report false positive

Table 1 Performance Summary of Selected AI Models for Pulmonary Tuberculosis Detection and Diagnosis from Chest Radiographs

Study (First Author, Year)	Model Name/ Architecture	Dataset Source & Size	Key Task	Sensitivity (%)	Specificity (%)	AUC
Qin ZZ, 2019[3]	Commercial CNN (System A)	Multi-country, 10,848 images	TB case detection	94.1	89.4	0.97
Hwang EJ, 2019[5]	Deep Learning-based CAD	National database, 54,221 studies	Major thoracic disease (incl. TB) detection	95.0*	91.0*	0.97*
Liu Y, 2021[7]	Custom Deep CNN	Single-center, 17,290 images	Active TB classification	93.8	92.1	0.98
(Representative Study from Meta-analysis)	Various DL models	Multiple cohorts, ~100,000 images (aggregated)	TB abnormality detection	90.2–96.5 (Pooled)	88.3–94.1 (Pooled)	0.95–0.99 (Range)

Notes: *Performance metrics are representative values for tuberculosis detection among multiple diseases evaluated in the study. The presented performance metrics are derived from the respective studies' internal or reported external validation sets. Variations in performance can be attributed to differences in dataset demographics, image quality, and the specific clinical task definition.
Abbreviations: AUC, Area Under the Receiver Operating Characteristic Curve; CNN, Convolutional Neural Network; CAD, Computer-Aided Detection; TB, Tuberculosis.

rates and associated clinical consequences alongside traditional sensitivity and specificity metrics to provide a more complete picture of model performance.

Limitations and Challenges

Despite promising prospects, the application of AI in TB image diagnosis faces several challenges. The primary issue is data bias; most models are trained on specific datasets and may lack generalizability to pediatric patients, those with Human Immunodeficiency Virus (HIV) coinfection, or atypical imaging presentations.[11] Closely linked to data bias is the risk of overfitting, which is exacerbated by small, homogeneous training datasets and can lead to spurious high performance in research settings that does not translate to clinical practice, alongside elevated false positive rates when models encounter unseen patient populations or imaging variations. Secondly, model interpretability remains a "black box", making it difficult for clinicians to understand the rationale behind specific decisions, which undermines clinical trust,[12] and further complicates the identification and correction of false positive predictions in real time. Additionally, requirements for computational infrastructure, regulatory approval processes, and the seamless integration of AI tools into existing clinical workflows (eg, Picture Archiving and Communication Systems (PACS)) for human-AI collaboration are practical issues that must be addressed for widespread clinical translation. The lack of standardized image acquisition protocols across different medical institutions further exacerbates the problem of poor model generalizability, as images from different devices may have differences in resolution and contrast, impairing the accuracy of model feature extraction[13] and increasing the likelihood of overfitting to device-specific artifacts and false positive lesion detections.

AI in Drug Susceptibility Prediction and Genomic Analysis

The emergence and spread of drug resistance in *Mycobacterium tuberculosis*, particularly multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB), pose a severe threat to global TB control. While traditional phenotypic drug susceptibility testing (phenotypic DST) remains the gold standard, its process is time-consuming, requiring weeks to months, thereby causing critical delays in the initiation of effective treatment regimens. Genotypic DST based on molecular techniques (eg, GeneXpert MTB/RIF) can rapidly detect resistance to some drugs but offers limited coverage. Recently, the widespread adoption and declining cost of whole-genome sequencing (WGS) technology, combined with the powerful pattern recognition capabilities of AI, have opened new avenues for achieving rapid, comprehensive, and accurate prediction of drug resistance. Recent studies have shown that DL combined with short-read WGS can increase the Area Under the Receiver Operating Characteristic Curve (AUROC) to 0.89 for predicting rare resistance mutations in *Mycobacterium tuberculosis*;[14] however,

DL models for genomic resistance prediction are highly susceptible to overfitting to rare mutation patterns in small training datasets, leading to false positive resistance predictions for first- and second-line drugs. However, although long-read technology can accurately capture genomic structural variations, it has not yet been integrated with AI models for drug-resistance prediction, and its potential benefits remain to be verified.

Technical Basis

The core technical foundation of this application lies in the integration of WGS and ML algorithms. WGS provides nearly complete genomic information of *M. tuberculosis*, enabling the systematic capture of resistance-associated single nucleotide polymorphisms (SNPs), insertions/deletions (indels), and other genetic variations. Subsequently, various supervised ML models are employed to construct predictive mappings from genotype to drug susceptibility phenotype. Commonly used algorithms include Logistic Regression, Random Forest, Support Vector Machines (SVM), and gradient boosting machines such as XGBoost.[15] Tree-based models like Random Forest and XGBoost carry lower overfitting risks than deep learning models but can still overfit to population-specific resistance SNP patterns, while all ML approaches are vulnerable to false positive resistance predictions when confronted with novel, uncharacterized genetic variations. These models are trained on large datasets comprising known genotypes and their corresponding phenotypic DST results to learn the complex genetic mutation patterns conferring drug resistance, yet the paucity of training data for rare second-line drug resistance mutations often leads to overfitting and elevated false positive rates for these clinically critical predictions.

From Genotype to Phenotype

The core value of AI models is their ability to directly analyze WGS data to predict the susceptibility of *M. tuberculosis* to multiple first-line and second-line drugs simultaneously, thereby bypassing the lengthy culture process. Studies have shown that ML-based prediction models achieve high accuracy (sensitivity and specificity often >90%) for predicting resistance to first-line drugs like Rifampicin and Isoniazid, performing comparably to traditional phenotypic DST.[16] False positive resistance predictions—where susceptible strains are misclassified as resistant—have severe clinical consequences, including the unnecessary use of toxic, expensive second-line anti-TB regimens, increased risk of adverse drug events, and the potential for secondary resistance development due to inappropriate treatment. Crucially, this approach can reduce the turnaround time from weeks to just days, enabling clinicians to formulate precise, individualized treatment regimens more rapidly for their patients, which is vital for improving outcomes in MDR-TB cases, provided that false positive resistance predictions are rigorously minimized through model validation on diverse global genomic datasets.

Discovery of Novel Resistance-Associated Mutations

Beyond predicting known resistance mutations, the powerful data mining capability of AI allows for the discovery of new, previously unrecognized genes or intergenic regions associated with resistance from vast genomic datasets. Traditional analytical methods often focus on known candidate genes, whereas unsupervised or semi-supervised ML approaches can scan the entire genome in an unbiased manner to identify novel genetic markers significantly associated with the resistant phenotype.[17] Unsupervised ML models, however, carry a risk of overfitting to spurious genomic correlations in training data, leading to the identification of false positive resistance-associated genetic markers that lack biological or clinical validity, which can misdirect subsequent research into TB resistance mechanisms. This not only deepens our scientific understanding of the resistance mechanisms in *M. tuberculosis* but also provides valuable clues for developing new molecular diagnostic targets, only when candidate markers identified by AI are validated in independent cohorts to rule out false positive associations.

Performance and Challenges

Despite the promising outlook, the application of AI in drug resistance prediction faces key challenges. Firstly, the predictive accuracy of models heavily depends on the quality and breadth of the training data. For certain second-line drugs or rarer resistance mutations, the predictive performance may decline significantly due to the limited number of positive samples available in public databases.[18] This limited sample size is a primary driver of overfitting, and the resulting false positive resistance predictions for second-line drugs are particularly problematic, as they can lead to

inappropriate XDR-TB classification and overly aggressive treatment. To address this issue, researchers have begun to use transfer learning and data augmentation techniques—for example, simulating mutation data based on existing genomic information to expand the training set, which has been shown to improve the prediction accuracy of rare mutations by 15–20%,[19] and more importantly, reduce overfitting and false positive rates by increasing the diversity of training data. Secondly, most current models primarily identify known SNPs; their predictive power remains limited for complex resistance mechanisms involving non-coding regulatory regions, phenotypic heterogeneity, or the combined effects of multiple genes with small individual contributions. These complex mechanisms further increase the risk of false positive predictions, as models may misclassify strains with non-canonical genetic variations as resistant without clear biological justification. Finally, integrating WGS and AI analysis pipelines into routine clinical microbiology laboratories faces practical barriers related to cost, bioinformatics expertise, and standardized operational procedures. Future research needs to focus on building more diverse, higher-quality global genomic databases and developing advanced algorithms capable of interpreting complex genetic mechanisms, as well as implementing built-in regularization and cross-validation steps to mitigate overfitting and quantify false positive risks for all drug resistance predictions.

AI in Treatment Response Monitoring and Outcome Prediction

The treatment of PTB is a prolonged process, and traditional assessment of therapeutic efficacy is often characterized by lag and subjectivity. Clinicians rely on serial chest imaging to evaluate lesion "resolution" or "improvement", but this assessment lacks quantitative standards and is affected by inter-observer variability. Simultaneously, predicting the risk of treatment failure or relapse in individual patients remains highly challenging. AI is reshaping the paradigm of TB treatment response monitoring and outcome prediction by providing objective, quantifiable imaging biomarkers and constructing comprehensive predictive models. However, these predictive models are not immune to overfitting to historical treatment data, and false positive predictions of treatment failure or relapse can lead to unnecessary clinical interventions and patient anxiety.

Imaging-Based Treatment Response Assessment

The most direct application of AI lies in the automated, precise quantitative comparison of serial images (eg, sequential chest CT scans). DL segmentation models (eg, U-Net) can automatically delineate the precise three-dimensional volumes of lesions such as lung infiltrates, nodules, and cavities on baseline and follow-up images.[20] By calculating metrics like the percentage reduction in total lesion volume, changes in cavity volume, or alterations in pixel density values of affected lung regions, AI generates continuous, objective quantitative reports.[9] Overfitting to baseline lesion volume and density features can cause AI models to generate false positive assessments of poor treatment response—for example, misclassifying minor, non-clinically significant lesion volume fluctuations as treatment failure—resulting in unnecessary regimen adjustments, additional diagnostic testing, and increased healthcare utilization. This approach significantly reduces subjective interpretation bias by radiologists, sensitively captures subtle changes imperceptible to the human eye, and provides unprecedented data support for clinical decisions (eg, determining treatment effectiveness or need for regimen adjustment), when model predictions are contextualized with clinical data to reduce false positive interventions. Studies have shown that the AI-quantified rate of lesion volume reduction early in treatment (eg, at 2 months) is significantly correlated with the final treatment outcome.[21] In a multicentre retrospective–prospective cohort (n = 493), an AI system that quantified CT lesion-volume changes at 2 months achieved an AUC of 0.82 (95% CI 0.77–0.87) for predicting subsequent treatment failure, outperforming the AUC of 0.74 (95% CI 0.68–0.80) obtained with conventional sputum-smear conversion.[22] Notably, this study did not fully report the false positive rate of treatment failure predictions, a critical omission that limits clinical translation, as false positive predictions can have a substantial impact on patient care and healthcare resource allocation.

Integrated Predictive Models

The complexity of outcome prediction necessitates moving beyond single data sources. A more advanced application of AI involves integrating multimodal data to build comprehensive predictive models. ML algorithms can fuse clinical

parameters (eg, age, HIV co-infection status, baseline bacterial load), serial imaging quantitative features, laboratory results, and sociodemographic factors to predict individualized treatment outcomes, such as the risk of treatment failure, relapse, or mortality.[23] High-dimensional multimodal data increases the risk of overfitting, as models may learn spurious correlations between non-clinically relevant features and treatment outcomes, leading to false positive risk stratification —where low-risk patients are misclassified as high-risk, resulting in unnecessary intensive monitoring and treatment. For instance, ensemble learning models like Gradient Boosting Trees can process these high-dimensional, heterogeneous data, identify key predictors and their weights, and generate personalized risk scores, and these models require rigorous regularization and external validation to mitigate overfitting and reduce false positive risk stratification. This facilitates the identification of high-risk patients, enabling closer monitoring or more intensive therapeutic interventions, ultimately improving overall treatment success rates, provided that false positive high-risk classifications are minimized to avoid over-medicalization of low-risk patients.

Challenges

Despite its considerable potential, applications in this field face significant challenges. The primary obstacle is the acquisition of high-quality, standardized longitudinal data. Building robust predictive models requires complete treatment cycle data from numerous patients, including imaging and clinical data at standardized time points, which is difficult to obtain in real-world clinical settings. Small or incomplete longitudinal datasets are a major driver of overfitting in treatment outcome prediction models, leading to elevated false positive rates for predicting treatment failure or relapse, and these errors are often unrecognized due to the lack of independent external validation. Secondly, model generalizability is a major test. A model developed in one cohort may exhibit degraded performance when applied to another cohort from different regions, populations, or healthcare settings,[24] with this degradation often manifesting as increased overfitting and higher false positive rates in the new cohort. Finally, integrating such predictive models into clinical workflows and implementing prospective interventions based on their predictions require rigorous prospective clinical validation and cost-effectiveness analyses, which are currently relatively scarce, and these analyses must explicitly evaluate false positive rates and their associated clinical and economic consequences to guide rational clinical implementation.

Clinical Decision Support and Implementation Challenges

The ultimate value of various AI applications in TB diagnosis and treatment lies in their seamless integration into routine clinical practice as Clinical Decision Support Systems (CDSS), supporting rather than replacing clinicians' judgment. However, the path from a high-performance research model to a routinely used medical tool is fraught with challenges. A central, underrecognized challenge is the mitigation of overfitting risks and the management of false positive consequences, which are critical to ensuring the safety and clinical utility of AI-CDSS in TB care. Systematically identifying and addressing these challenges is a key prerequisite for realizing the revolutionary potential of AI.

AI as a Clinical Decision Support System (CDSS)

A successful AI-CDSS should be designed to enhance the clinical workflow, not add complexity. In an ideal scenario, when a radiologist reviews the chest images of a suspected TB patient in a PACS system, an AI algorithm could automatically run an analysis and display annotations—such as suspicious lesion areas, a probability score for activity, or quantitative comparisons with prior images—in a conspicuous yet non-intrusive manner alongside the images.[25] Displaying probability scores for AI predictions is a critical strategy to mitigate the impact of false positives, as it allows clinicians to contextualize AI results and avoid over-reliance on binary positive/negative classifications that can lead to unnecessary interventions. Similarly, upon obtaining WGS data of *M. tuberculosis* in the microbiology lab, an AI prediction model could automatically generate a report detailing the drug resistance profile with confidence intervals, sending it directly to the treating physician's electronic health record (EHR) system. Including confidence intervals for drug resistance predictions helps clinicians identify high-risk false positive cases—eg, predictions with low confidence—and avoid inappropriate treatment decisions based on unsubstantiated AI outputs. This deep integration provides clinicians with a crucial "second opinion", improving diagnostic efficiency and consistency, a value particularly

pronounced in resource-limited settings with a scarcity of experts.[26] In clinical practice, there are two main models for integrating AI-CDSS into existing clinical pathways. One is the "pre-screening tool" model, in which AI first screens for positive cases and physicians then review the results. This model is particularly vulnerable to the impact of false positives, as a high false positive rate can increase rather than reduce clinician workload by requiring review of numerous non-TB cases, undermining the utility of the AI tool in resource-limited settings. This approach has been shown in scenarios such as Alzheimer's disease screening to significantly reduce physicians' workload by automating data analysis, allowing clinicians to devote more time to patient care;[27] the other is the "auxiliary diagnosis tool" mode, where doctors and AI read images simultaneously and make decisions together—this mode is more suitable for complex cases such as TB combined with other lung diseases, and can improve the diagnostic accuracy by 10–15%,[28] and also provides an opportunity for real-time correction of AI false positives by clinicians, reducing the clinical consequences of algorithmic errors.

Common Implementation Challenges

The translation of AI technologies into routine TB clinical practice is hindered by interconnected core challenges. These primarily include data-related issues such as selection bias, standardization deficits, and privacy-driven data silos; regulatory and ethical hurdles involving lengthy approval processes, strict privacy compliance requirements, and the need to mitigate algorithmic fairness; as well as technical barriers centered on model interpretability and generalizability across diverse populations and healthcare settings, and critical algorithmic challenges of overfitting and false positive-associated clinical harms (Table 2). Table 2 has been updated to include overfitting and false positives as key implementation challenges, with corresponding mitigation strategies. Addressing these multifaceted challenges is essential to ensure AI tools can be safely, equitably, and effectively integrated into clinical workflows.

Data Quality and Bias

The classic ML adage "garbage in, garbage out" holds true. The performance of AI models is highly dependent on the quality, scale, and representativeness of their training data. Currently, many publicly available medical imaging and genomic datasets suffer from significant selection bias, which can be divided into the following aspects: (1) Data representativeness bias: For example, datasets over-represent specific demographics (eg, adult males), disease stages, or regional bacterial strains. A study found that a certain AI model had an accuracy of 94% on adult data but dropped to 78% on pediatric data.[11] This drop in accuracy is often accompanied by a sharp increase in overfitting and false positive rates, as models fail to generalize to underrepresented populations. (2) Data standardization issues: Variations in imaging equipment, scanning protocols, and annotation standards across different healthcare institutions. For instance, images from different CT scanners may have differences in layer thickness and reconstruction algorithms, leading to inconsistent feature extraction by the model,[29] and increased overfitting to scanner-specific features, resulting in higher false positive rates when models are applied to images

Table 2 Key Challenges and Mitigation Strategies for AI Implementation in TB Care

Challenge Category	Specific Challenges	Mitigation Strategies
Data-Related	Selection bias, poor standardization, data silos	Federated learning, standardized imaging/genomic protocols, data augmentation
Regulatory-Ethical	Lengthy approval, privacy concerns, algorithmic bias	Adherence to FDA/GDPR/HIPAA guidelines, prospective clinical validation, fairness audits
Technical	Model interpretability, generalizability	Explainable AI (XAI), external validation across diverse populations
Clinical	Workflow integration, clinician acceptance	Integration with PACS/EHR systems, human-AI collaborative training

Notes: This table summarizes the core challenges encountered in the clinical implementation of AI technologies for tuberculosis (TB) care, categorized into four key dimensions: data-related, regulatory-ethical, technical, and clinical. Corresponding mitigation strategies are provided for each challenge to address practical barriers, enhance model reliability, and promote seamless integration of AI into routine TB diagnosis and management workflows.

Abbreviations: FDA, US Food and Drug Administration; GDPR, General Data Protection Regulation; HIPAA, Health Insurance Portability and Accountability Act; XAI, Explainable AI; PACS, Picture Archiving and Communication Systems; HER, Electronic Health Record.

from untested scanners. (3) Data privacy and sharing contradictions: While model training requires a large amount of data, privacy protection regulations restrict data sharing, resulting in "data silos" that make it difficult to obtain diverse training data.[30] Data silos limit the size and diversity of training datasets, a primary cause of overfitting, and the resulting lack of external validation leads to unrecognized false positive risks in clinical implementation. These issues can lead to a precipitous drop in model performance when faced with cases involving children, patients with HIV co-infection, or rare resistance types,[11,29] and are the root cause of most overfitting and false positive problems in AI models for TB care.

Regulatory and Ethical Hurdles

Regulating AI software as a medical device is a global consensus. Regulatory agencies like the US Food and Drug Administration (FDA) and those granting the CE mark in the EU have established evaluation frameworks for AI/ML-enabled software, though these pathways are continually evolving.[31] A critical gap in current regulatory frameworks is the lack of explicit requirements for reporting overfitting risks and false positive rates, alongside their clinical consequences, in AI model validation studies, which is essential to ensure the safety of AI medical devices for TB care. The lengthy approval cycles and the requirement for rigorous clinical validation evidence (often demanding prospective trials) constitute high translational costs. Ethically, patient data privacy is paramount; model training and use must strictly comply with regulations like the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). Additionally, algorithmic fairness is critical; it must be ensured that AI tools do not perpetuate or exacerbate systematic discrimination against certain racial, gender, or socioeconomic groups due to biases in the training data.[30] Algorithmic bias is often compounded by overfitting to biased training data, leading to disproportionately high false positive rates in marginalized populations—for example, higher false positive TB detection rates in low-income or ethnic minority patients—which exacerbates health disparities and represents a major ethical concern. For example, if the training data mostly includes patients from high-income groups, the AI model may have lower diagnostic accuracy for patients from low-income groups who have different living environments and disease manifestations,[32] and this reduced accuracy is frequently associated with higher overfitting and false positive rates, leading to inequitable clinical care.

Interpretability (XAI)

DL models are often criticized as "black boxes", meaning their internal decision-making logic is opaque. However, in life-critical medical decisions, clinicians must understand why a model made a specific judgment to build trust and assume ultimate responsibility. For instance, if an AI suggests "active tuberculosis", the physician needs to know which specific imaging features (eg, cavity wall, surrounding infiltrates) drove this conclusion.[12] Explainable AI (XAI) techniques are not only critical for building trust but also for identifying and correcting false positive AI predictions—by highlighting the features that drove a positive prediction, clinicians can quickly determine if the prediction is spurious (eg, based on an imaging artifact rather than a true TB lesion) and avoid unnecessary clinical actions. Developing Explainable AI (XAI) techniques—such as generating heatmaps to highlight critical regions on an image or providing a list of top genetic mutations that informed a decision—is crucial for fostering human-AI collaboration and ensuring clinical safety,[33] and for mitigating the clinical consequences of false positives by enabling real-time clinician oversight. A recent study applied XAI technology to TB imaging diagnosis, generating heatmaps that accurately marked the key areas of lesions, and the consistency between the heatmap prompts and the radiologists' judgment reached 89%, markedly enhancing clinicians' trust in the AI model and their ability to identify and reject false positive AI predictions.[34]

Future Directions

Multimodal Fusion Models

Future AI systems will inevitably transcend single-data-source analysis and advance toward multimodal data fusion. By integrating imaging features, genomic variation information, transcriptomic data, serum biomarkers, and clinical electronic health record text, more comprehensive digital phenotypes of patients can be constructed. While multimodal fusion improves predictive power, it also increases the risk of overfitting to high-dimensional data; future models must incorporate robust regularization techniques (eg, L1/L2 regularization, dropout) and cross-validation to minimize overfitting and false positive rates, while also developing XAI techniques to interpret multimodal predictions and identify

false positives. This approach holds promise for achieving earlier and more accurate diagnosis, as well as more precise prediction of treatment efficacy.[35] For instance, a comprehensive model integrating CT imaging characteristics of cavities, drug resistance mutation profiles, and inflammatory markers would likely demonstrate far superior predictive power for treatment failure compared to any single-modal model, provided that overfitting is mitigated and false positive prediction rates are quantified and minimized.

Federated Learning and Privacy Preservation

To resolve the conflict between data silos and privacy protection, Federated Learning—a distributed ML paradigm—will emerge as a key solution. It enables model training on local data within individual healthcare institutions, exchanging only model parameters instead of raw data. Federated learning is a powerful strategy to mitigate overfitting, as it leverages large, diverse, real-world clinical datasets from multiple institutions without compromising data privacy, leading to more generalizable models with lower false positive rates in clinical practice. This approach facilitates the collaborative refinement of model performance using multi-center, diverse datasets while strictly protecting patient privacy, thereby effectively enhancing algorithm generalizability,[36] and reducing overfitting and false positive rates by eliminating the limitations of small, single-institution training datasets. Future research should prioritize the development of federated learning frameworks specifically tailored for TB AI models, with a focus on minimizing overfitting and quantifying false positive rates across diverse global populations.

Prospective Clinical Validation

The majority of published AI research constitutes retrospective validation. The next critical step must shift toward conducting rigorously designed, large-scale prospective randomized controlled trials. These trials should validate the efficacy of AI tools within real-world clinical workflows and evaluate their ultimate impact on patient outcomes—such as treatment success rates and survival—as well as on healthcare cost-effectiveness.[37] Crucially, these prospective trials must explicitly report overfitting risks, false positive rates, and their associated clinical consequences (eg, unnecessary interventions, drug toxicity, delayed care) as key primary and secondary outcomes, rather than solely focusing on traditional performance metrics like sensitivity and specificity. This transition is an indispensable pathway for AI to achieve clinical certification and widespread adoption, and only prospective trials that evaluate false positive consequences can provide the evidence needed to guide safe, rational clinical implementation of AI tools for TB care.

Solutions for Resource-Limited Settings

Regions bearing the highest global TB burden often face the most severe healthcare resource constraints. Developing lightweight, computationally efficient AI models capable of offline deployment, and integrating them with portable diagnostic devices—such as handheld ultrasound units or smartphone-connected compact X-ray systems—carries significant public health significance for achieving widespread TB screening and rapid diagnosis in resource-limited areas.[38] Lightweight models must be specifically optimized to minimize overfitting and false positive rates, as resource-limited settings lack the clinical expertise and infrastructure to correct AI errors or manage the clinical consequences of false positives (eg, unnecessary anti-TB treatment in settings with limited drug access). Additionally, future work should focus on developing simple, user-friendly AI outputs that include probability scores and confidence intervals for predictions, enabling frontline healthcare workers in resource-limited settings to contextualize AI results and reduce the impact of false positives.

Conclusion

AI undoubtedly provides powerful new momentum toward achieving the global goal of ending tuberculosis. Its applications across radiology, genomics, and other domains are advancing pulmonary TB diagnosis and treatment toward greater speed, objectivity, and precision. However, the field has historically overemphasized high performance metrics from curated research datasets, while underappreciating the critical risks of algorithmic overfitting and non-trivial false positive rates in real-world clinical practice—risks that carry substantial clinical consequences, including unnecessary diagnostic procedures, inappropriate treatment, drug toxicity, delayed care, increased healthcare costs, and exacerbated health disparities. Technological maturity represents only half the success. Future development must adhere to the principle of clinical value orientation,

with success depending on close interdisciplinary collaboration—requiring deep engagement from computer scientists, clinicians (including pulmonologists, radiologists, and microbiologists), public health experts, and regulatory bodies, with a central focus on mitigating overfitting through diverse, large-scale training datasets and robust regularization, and addressing false positive consequences through XAI, probability-based predictions, and rigorous clinical validation, and regulatory bodies, with a central focus on mitigating overfitting through diverse, large-scale training datasets and robust regularization, and addressing false positive consequences through XAI, probability-based predictions, and rigorous clinical validation.

There are significant differences in AI application priorities between high-income and low-income countries: high-income countries focus more on the development of multimodal fusion models and the improvement of model interpretability, and on the rigorous quantification and mitigation of overfitting and false positive rates to ensure the safety and precision of AI tools in clinical practice;[38] while low-income countries prioritize the promotion of lightweight models and portable device integration to solve the problem of insufficient medical resources, and on the optimization of these models to minimize false positive rates—given the lack of infrastructure to manage their clinical consequences—and the development of simple AI outputs that enable frontline workers to contextualize predictions.[38] Only through collective efforts to address systemic challenges—including data bias, model generalizability, interpretability, regulatory-ethical considerations, algorithmic overfitting, and the clinical harms associated with false positive AI predictions—can we safely translate AI from a "high-performance" laboratory model into a "safe and reliable" clinical assistant worldwide, ultimately benefiting every TB patient.

Funding

There is no funding to report.

Disclosure

The authors declare that they have no conflicts of interest in this work.

References

1. World Health Organization. Global tuberculosis report 2025. India: World Health Organization; 2025.
2. Kontsevaya I, Cabibbe AM, Cirillo DM, et al. Update on the diagnosis of tuberculosis. *Clin Microbiol Infect*. 2024;30(9):1115–1122. doi:10.1016/j.cmi.2023.07.014
3. Hansun S, Argha A, Liaw ST, et al. Machine and deep learning for tuberculosis detection on Chest X-Rays: systematic literature review. *J Med Internet Res*. 2023;25:e43154. doi:10.2196/43154
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118. doi:10.1038/nature21056
5. Tavaziva G, Harris M, Abidi SK, et al. Chest X-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: an individual patient data meta-analysis of diagnostic accuracy. *IClin Infect Dis*. 2022;74(8):1390–1400. doi:10.1093/cid/ciab639
6. Said Y, Ayachi R, Afif M, et al. AI-driven genetic algorithm-optimized lung segmentation for precision in early lung cancer diagnosis. *Sci Rep*. 2025;15(1):23058. doi:10.1038/s41598-025-08116-w
7. Acharya V, Dhiman G, Prakasha K, et al. AI-assisted tuberculosis detection and classification from Chest X-Rays using a deep learning normalization-free network model. *Comput Intell Neurosci*. 2022;2022:2399428. doi:10.1155/2022/2399428
8. Munjal YP, Mahrooqi AA, Rajan R, et al. Population-scale cross-sectional observational study for AI-powered TB screening on one million CXRs. *NPJ Digit Med*. 2025;8(1):418. doi:10.1038/s41746-025-01832-7
9. Tu'ersun A, Abulizi A, Patiman Maimaiti P, et al. Predicting pulmonary tuberculosis treatment outcomes using longitudinal chest CT radiomics and deep learning. *Chin J Antituberculosis*. 2025;47(8):1044–1052. doi:10.19982/j.issn.1000-6621.20250047
10. Hansun S, Argha A, Liaw ST, et al. Deep learning for tuberculosis detection: a systematic review and meta-analysis. *J Med Internet Res*. 2023;25:e43154.
11. Kazemzadeh S, Yu J, Jamshy S, et al. Deep learning detection of active pulmonary tuberculosis at chest radiography matched the clinical performance of radiologists. *Radiology*. 2023;306(1):124–137. doi:10.1148/radiol.212213
12. Han X, Hu P, Ding JE, et al. No black boxes: interpretable and interactable predictive healthcare with knowledge - enhanced agentic causal discovery. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; November 5 – 9, 2025.; Suzhou, China.
13. Cobo M, Menéndez Fernández-Miranda P, Bastarrika G, et al. Enhancing radiomics and deep learning systems through the standardization of medical imaging workflows. *Sci Data*. 2023;10(1):732. doi:10.1038/s41597-023-02641-x
14. Ren Y, Chakraborty T, Doijad S, et al. Deep transfer learning enables robust prediction of antimicrobial resistance for novel antibiotics. *Antibiotics*. 2022;11(11):1611. doi:10.3390/antibiotics11111611
15. Hatherell HA, Colijn C, Stagg HR, et al. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med*. 2016;14(1):21. doi:10.1186/s12916-016-0566-x

16. Wang L, Yang J, Chen L, et al. Whole-genome sequencing of Mycobacterium tuberculosis for prediction of drug resistance. *Epidemiol Infect*. 2022;150:e22. doi:10.1017/S095026882100279X
17. Saliba JG, Zheng W, Shu Q, et al. Enhanced diagnosis of multi-drug-resistant microbes using group association modeling and machine learning. *Nat Commun*. 2025;16(1):2933. doi:10.1038/s41467-025-58214-6
18. Goossens SN, Sampson SL, Van Rie A. Mechanisms of drug-induced tolerance in mycobacterium tuberculosis. *Clin Microbiol Rev*. 2020;34(1): e00141–20. doi:10.1128/CMR.00141-20
19. Nijiati M, Guo L, Tuersun A, et al. Deep learning on longitudinal CT scans: automated prediction of treatment outcomes in hospitalized tuberculosis patients. *iScience*. 2023;26(11):108326. doi:10.1016/j.isci.2023.108326
20. Li Y, Huang F, Chen D, et al. Deep learning models for CT segmentation of invasive pulmonary aspergillosis, mucormycosis, bacterial pneumonia and tuberculosis: a multicentre study. *Mycoses*. 2024;67(11):e70084. doi:10.1111/myc.70084
21. Du Plessis T, Rae WID, Ramkilawon G, et al. Quantitative chest X-ray radiomics for therapy response monitoring in patients with pulmonary tuberculosis. *Diagnostics*. 2023;13(17):2842. doi:10.3390/diagnostics13172842
22. Kim H-J, Kwak N, Yoon SH, et al. Artificial intelligence-based radiographic extent analysis to predict tuberculosis treatment outcomes: a multicentre cohort study. *Sci Rep*. 2024;14(1):13162. doi:10.1038/s41598-024-63885-0
23. Peetluk LS, Ridolfi FM, Rebeiro PF, et al. Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults. *BMJ Open*. 2021;11(3):e044687. doi:10.1136/bmjopen-2020-044687
24. Cohen B, Vawdrey DK, Liu J, et al. Challenges associated with using large data sets for quality assessment and research in clinical settings. *Policy Polit Nurs Pract*. 2015;16(3–4):117–124. doi:10.1177/1527154415603358
25. Zaidi SMA, Mahfooz A, Latif A, et al. Geographical targeting of active case finding for tuberculosis in Pakistan using hotspots identified by artificial intelligence software (SPOT-TB): study protocol for a pragmatic stepped wedge cluster randomised control trial. *BMJ Open Respir Res*. 2024;11(1):e002079. doi:10.1136/bmjresp-2023-002079
26. Siddharth G, Ambekar A, Jayakumar N. Enhanced CoAtNet based hybrid deep learning architecture for automated tuberculosis detection in human chest X-rays. *BMC Med Imag*. 2025;25(1):379. doi:10.1186/s12880-025-01901-z
27. Elhaddad M, Hamam S. AI-driven clinical decision support systems: an ongoing pursuit of potential. *Cureus*. 2024. doi:10.7759/cureus.57728
28. Liu J, Zhao L, Han X, et al. Estimation of malignancy of pulmonary nodules at CT scans: effect of computer-aided diagnosis on diagnostic performance of radiologists. *Asia Pac J Clin Oncol*. 2021;17(3):216–221. doi:10.1111/ajco.13362
29. Mackin D, Fave X, Zhang L, et al. Measuring CT scanner variability of radiomics features. *Investigative Radiol*. 2015;50(11):757–765. doi:10.1097/RLI.0000000000000180
30. Koutsoubis N, Waqas A, Yilmaz Y, et al. Privacy-preserving federated learning and uncertainty quantification in medical imaging. *Radiol Artif Intell*. 2025;7(4):e240637. doi:10.1148/ryai.240637
31. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med*. 2018;378 (11):981–983. doi:10.1056/NEJMp1714229
32. Zhang Y, Wang L, Liu H. AI-driven healthcare: a review on ensuring fairness and mitigating bias. arXiv preprint arXiv:2407.19655. 2024. Available from: https://arxiv.org/pdf/2407.19655v2. Accessed March 26, 2026.
33. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1
34. Bhandari M, Shahi TB, Siku B, et al. Explanatory classification of CXR images into COVID-19, pneumonia and tuberculosis using deep learning and XAI. *Comput Biol Med*. 2022;150:106156. doi:10.1016/j.compbiomed.2022.106156
35. Teoh JR, Dong J, Zuo X, et al. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *PeerJ Comput Sci*. 2024;10:e2298. doi:10.7717/peerj-cs.2298
36. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24–29. doi:10.1038/s41591-018-0316-z
37. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374. doi:10.1038/s41591-020-1034-x
38. Geric C, Qin ZZ, Denkinger CM, et al. The rise of artificial intelligence reading of chest X-rays for enhanced TB diagnosis and elimination. *Int J Tuberc Lung Dis*. 2023;27(5):367–372. doi:10.5588/ijtld.22.0687