Can We Triage for Pulmonary Tuberculosis from the Sound of a Cough? A Comprehensive Technical Review of Artificial Intelligence-Based Approaches

Gutta J. Chowdary[1 †], Shams Nafisa Ali[1,2,3 †], Varunya Sakpuntoon[1], Adrienne E. Shapiro[4], Blanca I. Restrepo[5], Stephen J. Pont[6], Lisa Y. Armitige[7,8], Leonard Kingwara[9], Umberto E. Villa[10, 1], and Nuttada Panpradist[1,11 *]

[1]Department of Biomedical Engineering, The University of Texas at Austin, Austin, Texas, USA
[2]Department of Biomedical Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[3]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland, USA
[4]Departments of Global Health and Medicine (Division of Allergy & Infectious Diseases), University of Washington, Seattle, Washington, USA
[5]Department of Epidemiology, UTHealth School of Public Health in Brownsville, Brownsville, Texas, USA
[6]Texas Department of State Health Services; Departments of Pediatrics and Population Health, Dell Medical School, University of Texas at Austin, Austin, Texas, USA
[7]Heartland National TB Center, San Antonio, Texas, USA
[8]Medicine/Pediatrics Division of Adult Infectious Diseases, The University of Texas Health Northeast, San Antonio, Texas, USA
[9]Biomedical Testing and Analytical Services, Kenya National Public Health Institute, Ministry of Health (KNPHI-MoH), Nairobi, Kenya
[10]Oden Institute for Computational Engineering and Sciences, Austin, Texas, USA
[11]LaMontagne Center for Infectious Disease, College of Natural Sciences, The University of Texas at Austin, USA
[†]These authors contributed equally to this work.
[*]Corresponding author email: nuttadap@utexas.edu

ABSTRACT

Tuberculosis (TB) remains the world's deadliest infectious disease, and progress toward elimination is hindered by persistent gaps in early diagnosis. Current reliance on sputum-based tests excludes many vulnerable groups and constrains community-level detection. Recent advances in artificial intelligence (AI) and mobile health technologies have renewed interest in cough, the most universal symptom of pulmonary TB, as a scalable acoustic biomarker. Emerging studies suggest that cough carries disease-specific signatures that can be captured by digital devices and interpreted through AI models, raising the possibility of rapid, non-invasive, and widely deployable triage tools. However, the field is still constrained by small and geographically skewed datasets, inconsistent data collection methods, and limited validation in real-world populations. In this technical review, we synthesize evidence from existing studies and situate cough analysis within the broader landscape of non-sputum diagnostics. We highlight methodological and clinical challenges, examine the roles of diverse stakeholders in development and deployment, and outline a roadmap toward equitable translation. If these challenges are addressed, AI-assisted cough diagnostics could redefine TB case finding by moving testing from centralized laboratories to community settings worldwide.

1 Introduction

Tuberculosis (TB) remains the leading cause of death from a single infectious pathogen, despite being both preventable and curable. In 2023, an estimated 10.8 million people developed TB and 1.25 million died, including 161,000 among people living with HIV (PLH)[1–3]. Following years of decline, global incidence and mortality rose during the COVID-19 pandemic and have not returned to pre-2020 trajectories[1]. Although the greatest burden falls on low- and middle-income countries (LMICs), rising incidence in high-income regions[4,5] and outbreaks in low-incidence settings such as Kansas[6] underscore that TB remains a persistent global threat. These patterns, combined with widening inequities in access to care, highlight the fragility of current control strategies.

A central barrier to TB elimination is delayed or missed diagnosis[7]. Current tools rely heavily on sputum-based methods such as smear microscopy, culture, and nucleic acid amplification tests[8]. While effective in laboratories, these approaches are impractical for community screening and often less or inaccurate to children, PLH, and individuals with paucibacillary or extrapulmonary disease. Chest radiography, even when supported by computer-aided detection, requires infrastructure and

trained personnel. As a result, many cases remain undetected or are diagnosed too late, sustaining transmission and delaying treatment.

The World Health Organization (WHO) has identified rapid, non-sputum-based diagnostics as a global priority for TB control[9,10]. Among emerging strategies, cough analysis is particularly compelling. Cough is the most common symptom of pulmonary TB, requires no consumables or invasive sampling, and can be recorded on ubiquitous devices such as smartphones. Advances in acoustic signal processing and artificial intelligence now make it possible to transform cough from a subjective symptom into a quantifiable biomarker[11]. Early studies suggest that TB coughs carry distinctive acoustic signatures[11–14], and proof-of-concept models have reported promising sensitivity. However, only two of 14 recent studies met both WHO sensitivity and specificity benchmarks, and most datasets are geographically concentrated in a limited number of countries[15,16]. Together, these findings highlight the translational gap that must be closed for cough analysis to achieve clinical impact.

This review synthesizes evidence from 123 studies, including 40 that directly evaluated AI-based models for cough-based TB diagnostics. We situate cough analysis within the broader landscape of emerging diagnostics, discuss its biological and technical foundations, and evaluate opportunities and challenges for translation, equity, and integration into TB programs worldwide.

1.1 Current Diagnostic Landscape and Limitations across the TB Disease Spectrum

TB does not progress in a strict binary from latent infection to active disease. Instead, it spans a continuum that includes incipient and subclinical stages, marked by rising bacterial burden, evolving immune responses, and increasing infectiousness[17]. Detecting individuals in these early phases is especially difficult, since they may transmit disease despite mild or absent symptoms.

Subclinical TB exemplifies this challenge: individuals may carry transmissible infection while remaining asymptomatic, and conventional diagnostics often fail in this window[17]. These limitations have spurred interest in non-invasive, community-deployable biomarkers, with cough emerging as particularly attractive. Detecting TB-associated cough patterns before overt disease could enable earlier intervention and reduce transmission.

Traditional tools—including symptom screens, chest radiography, and sputum-based assays—remain central to practice[8], but they are slow, infrastructure-dependent, and limited in sensitivity for early or extrapulmonary disease. Alternative approaches under investigation include urine lipoarabinomannan assays[18], host-response biomarkers in blood and urine[19,20], breathomics[21–24], oral swabs[25], and AI-assisted imaging[26]. Each offers promise but none yet achieves the combination of scalability, robustness, and ease of integration required for population-level screening[27,28].

1.2 Cough as a Biomarker

Cough has long been recognized as the hallmark symptom of pulmonary TB. Pathological changes in the lungs alter airflow dynamics and sound production, generating distinctive acoustic features. With the spread of low-cost digital sensors, these patterns can now be captured and analyzed using AI algorithms[11].

Recent studies confirm that TB-associated coughs carry unique temporal and spectral signatures[11–14]. Diagnostic pipelines typically involve five stages: acquisition, preprocessing, feature extraction, AI-based classification, and validation against clinical standards (Figure 1). Current evidence, though geographically concentrated, demonstrates promising sensitivity but persistent limitations in specificity relative to WHO targets[15,16].

The context in which coughs are collected matters critically. Most datasets rely on active (solicited) coughs, elicited on instruction or via induction, whereas passive (unsolicited) coughs—spontaneous, activity-triggered, or nocturnal—better reflect community settings[29]. Models trained only on active coughs may fail to generalize to real-world deployment scenarios.

Beyond TB, cough has shown value as a surrogate biomarker in other respiratory conditions, correlating with lung function and disease burden[30]. Initiatives such as TBScreen and CODA[11,14] illustrate efforts to capture both explosive and full-sequence coughs, opening opportunities to extract richer acoustic and temporal features.

These advances position cough as a uniquely accessible biomarker with potential for community-based screening that reduces reliance on sputum and brings TB detection closer to the point of need.

2 Study Design

To provide a rigorous evidence base for this rapidly evolving field, this technical review was conducted in accordance with the PRISMA (Preferred Reporting Items for Technical Reviews and Meta-Analyses) framework[31] and synthesizes recent advances in AI-assisted TB testing using acoustic biomarkers, with particular emphasis on cough as a non-invasive and scalable triage tool.

2.1 Search Strategy

We conducted a comprehensive literature search on March 25, 2025 across Scopus and PubMed, restricted to English-language studies published from 2020 onward. The search combined three thematic categories: artificial intelligence and machine learning approaches; disease-specific terminology (including TB and broader respiratory diseases); and diagnostic modalities such

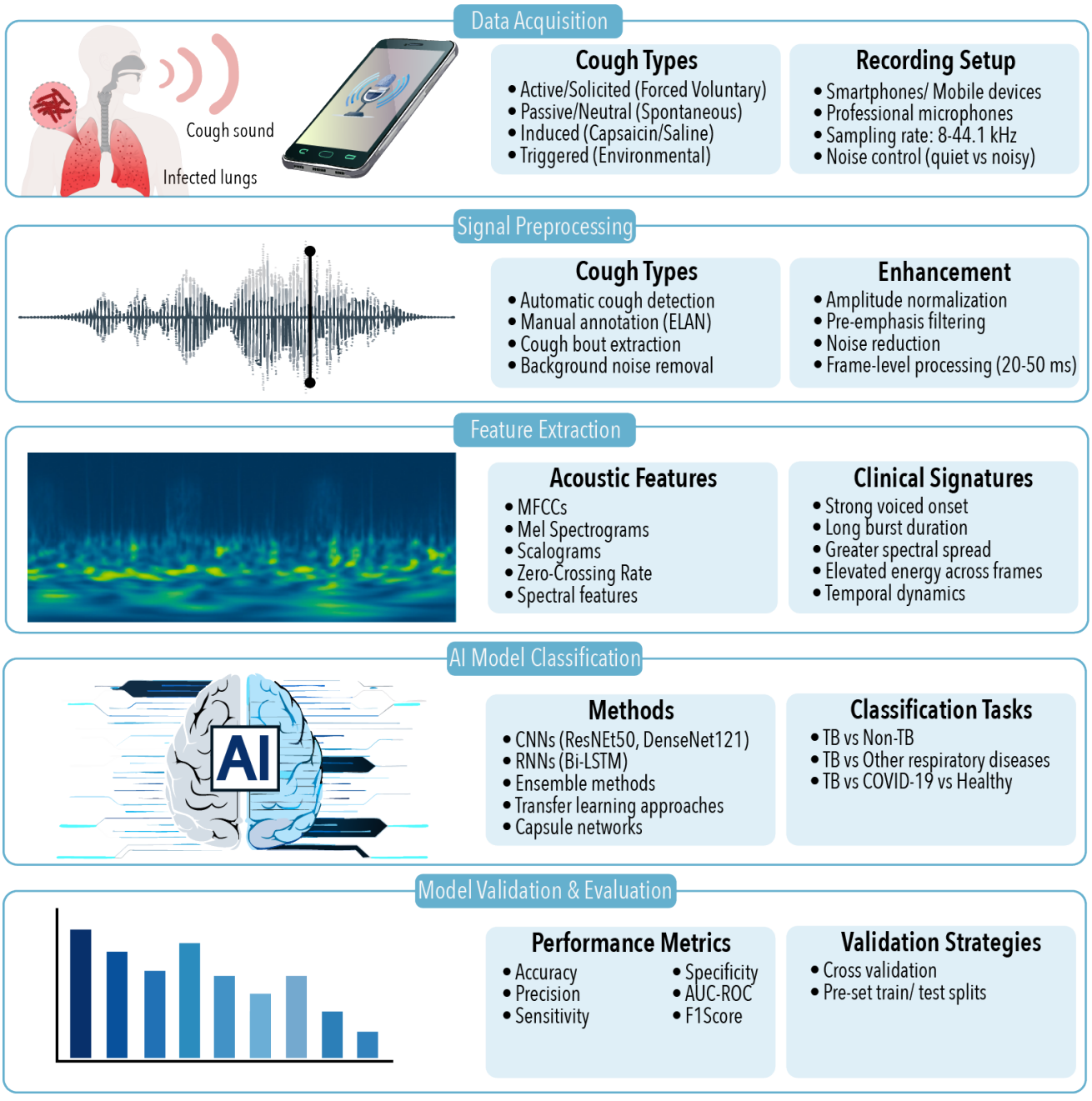

Figure 1. Five-stage cough-based TB diagnostic pipeline encompassing data acquisition, preprocessing, feature extraction, AI classification, and validation.

as cough, acoustic markers, point-of-care platforms, and breathomics. Use of the broader term "respiratory disease" allowed capture of studies situating TB within differential diagnostic frameworks. Breathomics and point-of-care terms were retained to reflect the translational potential of integrating complementary non-invasive modalities into cough-based screening. No restrictions were placed on geographic origin or study setting.

2.2 Article Selection and Categorization

The primary inclusion criterion encompassed studies employing AI for TB diagnosis, screening, or triage based on cough sound analysis. As this represents a relatively emerging research domain, secondary inclusion criteria were defined to ensure comprehensive coverage and contextual understanding. These included studies exploring AI-based or point-of-care triage approaches for TB through alternative modalities, such as medical imaging, breath volatile organic compound (VOC) profiling, electronic health records (EHR), auscultation signals, multimodal fusion frameworks, and deployment evaluations. Additionally, non–TB-specific AI-assisted cough diagnostic studies were retained when encountered, given their methodological relevance. The exclusion criteria eliminated publications that were non-diagnostic or non-computational in nature, including clinical case reports, biological or epidemiological investigations, and studies primarily addressing disease transmission. Commentaries, reviews not specific to TB, studies confined to extrapulmonary TB, and AI studies unrelated to TB (e.g., COVID-19, pneumonia, or other respiratory diseases) that did not involve cough-based analysis were also excluded.

The search retrieved 504 articles, of which 85 studies met the inclusion criteria following duplicate removal and title–abstract screening. Upon full-text assessment, the included studies were subsequently categorized into four thematic groups: (i) cough-based analysis ($n = 29$), (ii) other diagnostic modalities and deployment studies ($n = 41$), (iii) multimodal frameworks ($n = 3$), and (iv) review or survey articles ($n = 12$). The primary focus of this review is on the first category, encompassing AI-driven cough-based approaches for TB diagnosis and triage. This hierarchical structure allowed a focused evaluation of cough-based methods while contextualizing them within the broader AI-enabled TB diagnostic landscape. Screening and categorization were performed independently by two reviewers, with any discrepancies resolved through consensus discussion.

2.3 Quality Assessment

Although most studies clearly defined prediction tasks, architectures, and evaluation metrics, few addressed class imbalance rigorously, and external validation was rare. Reporting of hyperparameter optimization was particularly limited, reflecting broader challenges in reproducibility across AI-driven triage research. Nevertheless, nearly all studies specified model architectures, training procedures, and evaluation metrics, providing a basis for cross-study comparison, even if reproducibility remains limited.

2.4 Performance Metrics

Cough-based AI triage tests were assessed using standard classification metrics, each offering distinct clinical insights. Sensitivity (true positive rate) is critical for minimizing missed TB cases in high-burden settings, whereas specificity (true negative rate) helps avoid unnecessary treatment and patient anxiety. Accuracy provides an overall measure of performance but is often misleading in imbalanced datasets. Precision quantifies the reliability of positive predictions, and the F1 score balances precision and sensitivity when distributions are skewed. The area under the receiver operating characteristic curve (AUC-ROC) offers a threshold-independent measure of discriminative ability. Future evaluations should prioritize sensitivity and specificity in line with WHO target product profiles (TPP) ($\geq$65% sensitivity, $\geq$98% specificity for triage tests)[32], while also reporting threshold-independent measures such as AUC to enable fair benchmarking across studies. Formal definitions of these metrics are provided in Appendix A.

3 Datasets, Acoustic Biomarkers, and AI Approaches for TB Triage

Pulmonary tuberculosis (PTB) accounts for approximately 90% of cases worldwide[33]. Cough, a cardinal symptom, has long been recognized as a potential biomarker, and accumulating evidence shows that it carries disease-specific acoustic signatures[12,14,34]. This section synthesizes the landscape of cough-based studies, from datasets and acoustic signatures to AI methodologies and clinical performance benchmarks relative to WHO target product profiles (TPPs). Detailed dataset attributes are provided in Tables 1–2.

3.1 Study landscape and acoustic datasets

Of the 29 cough-based analysis studies identified, 14 were published in peer-reviewed journals, 8 in conference proceedings, 4 as preprints, 1 as a news article, 1 as a book chapter, and 1 as a doctoral dissertation. Most adopted retrospective or cross-sectional designs and focused on supervised classification tasks such as TB vs. non-TB, TB vs. healthy, TB vs. COVID-19, or multi-class problems (e.g., TB vs. healthy vs. COVID-19).

Notably, 19 of these 29 studies reported dataset creation or compilation, underscoring that the field remains in an early stage where building acoustic resources is

as central as methodological innovation. Collectively, these studies describe 17 unique acoustic datasets reported to date for TB diagnosis: sixteen cough-based and one respiratory-auscultation-based[34]. Of the 16 cough-based datasets, 13 are detailed in Tables 1–2, while three studies[35–37] mentioned TB cough data collection but could not be independently identified or validated.

Substantial heterogeneity is evident in elicitation mode (solicited vs. passive), study design (cross-sectional vs. longitudinal), control composition (healthy vs. TB-RD vs. mixed), recording protocols, and sample size. For instance, the Peru TB dataset[38] recorded serial coughs during treatment (days 0, 3, 7, 14, 30, 60), whereas TASK[39] captured natural coughs from hospitalized patients over therapy. CODA TB DREAM Challenge spanned seven international sites and included both active and passive coughs, while most other collections were single-center with standardized active elicitation. Recording devices ranged from smartphones to dedicated microphones; sampling rates spanned 8–44.1 kHz; and durations covered single-cough events to 10-second segments. These methodological differences are not trivial—elicitation mode, acoustic environment, and device type can meaningfully alter cough signatures, affecting diagnostic reliability.

Only Swaasa TB[16], CIDRZ TB[40,41], and CODA[11,42] enrolled more than 500 participants; most datasets included fewer than 200. Public access remains limited (three open datasets, one upon request). A smaller subset of studies investigated feasibility or treatment-monitoring applications. Overall, the field remains exploratory, characterized by limited longitudinal follow-up and few deployment-ready evaluations. A full dataset inventory is provided in Tables 1–2.

3.2 Limitations and representativeness

Despite the global burden of tuberculosis, available cough datasets remain geographically and structurally imbalanced (Fig. 2). Fig. 2a illustrates that most datasets originate from South Africa (Brooklyn[43], Wallacedene[15,44], TASK[39], one CODA[11,42] site) and China (Zhejiang V1[46], V2[47], Hangzhou[48,49]), with India contributing Swaasa TB[16] and one CODA[11,42] site. Additional single-site studies came from Pakistan, Kenya, Ethiopia, Peru, Vietnam, the Philippines, Zambia, Madagascar, Uganda, and Tanzania. Entire high-burden regions—Central and West Africa, Central Asia, and much of Southeast Asia—remain unrepresented. CODA[11,42] is the only multi-country dataset, providing the broadest geographic coverage among available cohorts. Epidemiological representativeness is similarly limited. Fig. 2b shows that dataset-level TB prevalence spans nearly an order of magnitude, with smaller, case-enriched cohorts exhibiting higher prevalence and wider uncertainty intervals, whereas larger datasets show systematically lower prevalence and narrower uncertainty intervals. Confidence-interval widths narrow systematically with increasing cohort size, reflecting reduced statistical uncertainty in larger studies. This imbalance underscores that evidentiary weight is concentrated in one large, low-prevalence dataset (CODA[11,42]), while smaller studies contribute higher uncertainty and prevalence estimates. Phase-space analysis (Fig. 2c) reveals that most datasets cluster within a 30–60 % prevalence band (IQR 28.6–59.3 %), spanning nearly two orders of magnitude in cohort size. CODA[11,42] uniquely occupies the high-N, low-prevalence quadrant, whereas Peru TB[38] and TASK[39] lie near the 100 % TB boundary. Such heterogeneity illustrates why models trained primarily on mid-size, mid-prevalence datasets may generalize poorly to population-level screening or high-burden outbreak contexts. In practice, specificity—rather than sensitivity—remains the limiting factor for deployment.

3.3 Clinical and acoustic signatures of TB cough

TB coughs exhibit distinctive features compared to non-TB coughs. Energy tends to concentrate below 500 Hz, with longer burst durations[14], oscillatory spectral patterns above 2.5 kHz[12], and highly informative initial voiced regions[12]. Frame-level feature extraction (20–50 ms) captures non-stationary dynamics, while amplitude normalization and pre-emphasis filters reduce variance due to patient effort and microphone distance. Common representations include Mel-frequency cepstral coefficients (MFCCs), scalograms, and spectral contrast[13]. These acoustic features are physiologically grounded: cavitary lesions, airway narrowing, and altered lung compliance in TB disrupt normal airflow dynamics, producing lower-frequency concentration and irregular high-frequency oscillations. Clinically, the value extends beyond binary discrimination: differentiating TB from other respiratory diseases (TB-RD, pneumonia, COVID-19) is more challenging yet more relevant, since it directly determines triage and referral pathways.

3.4 AI methodologies

Convolutional neural networks (CNNs) dominate. Xu et al.[48] achieved 94.3% accuracy and 97.7% sensitivity with a CNN on Mel spectrograms, while Yuan et al.[49] combined ResNet50 and DenseNet121 with MFCC and Mel spectrogram inputs, reporting F1 = 91.7% and AUC = 0.94. Pahar et al.[44] used ResNet50 transfer learning, achieving F1 = 92.6% for TB vs COVID-19 and 86.3% for TB vs COVID-19 vs healthy. Large-scale implementations further extend this success: Suda et al.[51] trained CNN + XGBoost models on over 700,000 coughs, demonstrating strong generalization across cohorts. Temporal architectures extend performance: Frost et al.[52] used a Bi-LSTM with attention (80.0% accu-

Table 1. Overview of cough sound datasets for TB triage (Part 1). "N/A" indicates information not reported. TB-RD refers to subjects with other respiratory diseases but confirmed TB-negative. Notes: (1) For the CODA TB DREAM Challenge dataset[11,42], the number of TB-positive and TB-negative respiratory disease recordings is publicly available and reported only for the training set. (2) One dataset originally published prior to 2020 (Brooklyn[43]) is included because it has been widely reused in successive studies and remains foundational to the field. (3) Confirmatory tests used to establish TB status include culture (Brooklyn[43]; CIDRZ TB[40,41]) and GeneXpert (Wallacedene[15,44]; CODA TB DREAM Challenge[11,42]; Swaasa TB[16]; Nairobi[14]; CIDRZ TB[40,41]). (4) Publication type abbreviations used in the Dataset Name column are as follows: (J) = journal article; (C) = conference proceeding; (P) = preprint.

Dataset Name	Population	Location	#Subjects	#Recordings	Public	Solicited	Passive	Recording Software/ Hardware	Recording Condition	Cough Segmentation
Brooklyn [43](J)	TB+ patients, healthy subjects (with/without symptoms) who had contact with TB+ patient	South Africa	38 (17 TB+, 21 Healthy)	746 (501 TB+, 245 Healthy)	×	✓	×	Tascam DR-44WL + RØDE M3 mic	Noise-free	Automatic + manual inspection
Wallacedene [15](J), [44](C)	18Y+, TB+, TB presumptive but negative	Cape Town, South Africa	51 (16 TB+, 35 TB- RD)	1,358 (402 TB+, 956 TB- RD)	×	✓	×	RØDE M3 mic + ZOOM F8N field recorder	Noisy hospital (outside sputum booth)	Manual (ELAN software)
CODA TB DREAM Challenge [11](J), [42](P)	18Y+, TB presumptive	India, Philippines, South Africa, Uganda, Vietnam, Tanzania, Madagascar	2143 (552 TB+, 1,591 TB-RD)	733,756 (4,43,707 TB+, 2,80,987 TB- RD); Solicited: 18,834; Passive: 7,14,922	✓	✓	✓	Smartphones with Hyfe Research App	Clinic rooms (Solicited), Outpatient (Passive)	Automatic
Swaasa TB [16](J)	18Y+, TB presumptive and healthy	Andhra Medical College, India	567 (278 TB+, 289 TB-)	N/A	×	✓	×	Smartphones/ Tablets	Noise-free (filtering applied)	Automatic
CIDRZ TB [40](P),[41]	18Y+, TB presumptive, had contact with TB+ patients, newly diagnosed with HIV	Lusaka, Zambia	599 (92 TB+, 217 HIV+)	15423	✓	✓	×	Zoom H2N microphone, Samsung Galaxy (A22, A12), Pixel3a	Noise-free	N/A
Nairobi [14](J)	18Y+, TB confirmed & presumptive (patients with other RD)	Kenya Medical Research Institute, Nairobi, Kenya	149 (103 TB+, 46 TB-RD)	34,866; Passive: 33,641 (23,191 TB+, 10,450 TB-), Solicited (Forced): 1,225 (991 TB+, 234 TB-)	✓	✓	✓	Google Pixel 2 + boundary mic + Yeti condenser mic	Quiet room (minimal background noise)	Manual

Table 2. Overview of cough sound datasets for TB triage (Part 2). "N/A" indicates information not reported. TB-RD refers to subjects with other respiratory diseases but confirmed TB-negative. Notes: (1) Reported TB data collection/usage from[35–37] could not be independently verified during review. (2) Confirmatory diagnostic tests used include culture (Peru TB[38]; TASK[39]; Zhejiang Red Cross V1[46]; Zhejiang Red Cross V2[47]; Hangzhou TB[48,49]) and GeneXpert (CCD[45]; Mayo TB [13]; Hangzhou TB[48,49]). (3) Publication type abbreviations in the Dataset Name column are as follows: (J) = journal article; (C) = conference proceeding; (P) = preprint.

Dataset Name	Population	Location	#Subjects	#Recordings	Public	Solicited	Passive	Recording Software/ Hardware	Recording Condition	Cough Segmentation
Peru TB [38](J)	18Y+, TB + on screening day and after 3, 7, 14, 30, and 60 days of treatment	Two tertiary hospitals, Lima, Peru	69	358	×	×	✓	Modified CayCaMo monitor (piezoelectric vibrometer)	Vibrometer used to minimize speech or background	Automatic + manual inspection
Cough classification dataset (CCD)[45](C)	TB+ and TB presumptive (other RD)	Bahir Dar Felege Hiwot Hospital, Ethiopia	71 (15 TB+, 56 TB- RD)	3238 (1080 TB+, 2158 TB- RD cough events)	×	N/A	N/A	Philips DVT1200, HM1000 mic, Infinix Hot 8 smartphone	Noisy outpatient hospital	Manual (Audacity)
Mayo TB [13](J)	18Y+, TB+ & TB presumptive, TB- included healthy subjects and patients with other RD	Mayo Hospital, Lahore, Pakistan	145	435	× (Available on request)	✓	✓	Smartphones with AI4LYF DCT App	N/A	Automatic
TASK[39](J)	Patients undergoing TB treatment	Cape Town, South Africa	14 (all TB+, 6,000 cough + 68,000 non-cough)	N/A	×	×	✓	Smartphone with accelerometer + external mic	Noisy hospital ward	Manual
Zhejiang Red Cross V1 [46](J)	18Y+, TB confirmed and healthy	Hangzhou Red Cross Hospital, Zhejiang, China	144 (70 TB+, 74 healthy)	456 (230 TB+, 226 healthy)	×	✓	×	Smartphone	Noise-free (inpatient)	N/A
Zhejiang Red Cross V2 [47](C)	18Y+, TB confirmed and healthy	Hangzhou Red Cross Hospital, Zhejiang, China	356 (176 TB+, 180 healthy)	361 (500 segments/class)	×	✓	×	Smartphone	Noise-free (inpatient)	N/A
Hangzhou TB[48](J), [49](C)	18Y+, TB presumptive and healthy	Hangzhou Red Cross Hospital, Zhejiang, China	345 (141 TB+, 52 TB- RD, 152 healthy)	1,323 (441 TB+, 441 TB- RD, 441 healthy)	×	✓	×	Smartphone	Noise-free (inpatient)	N/A

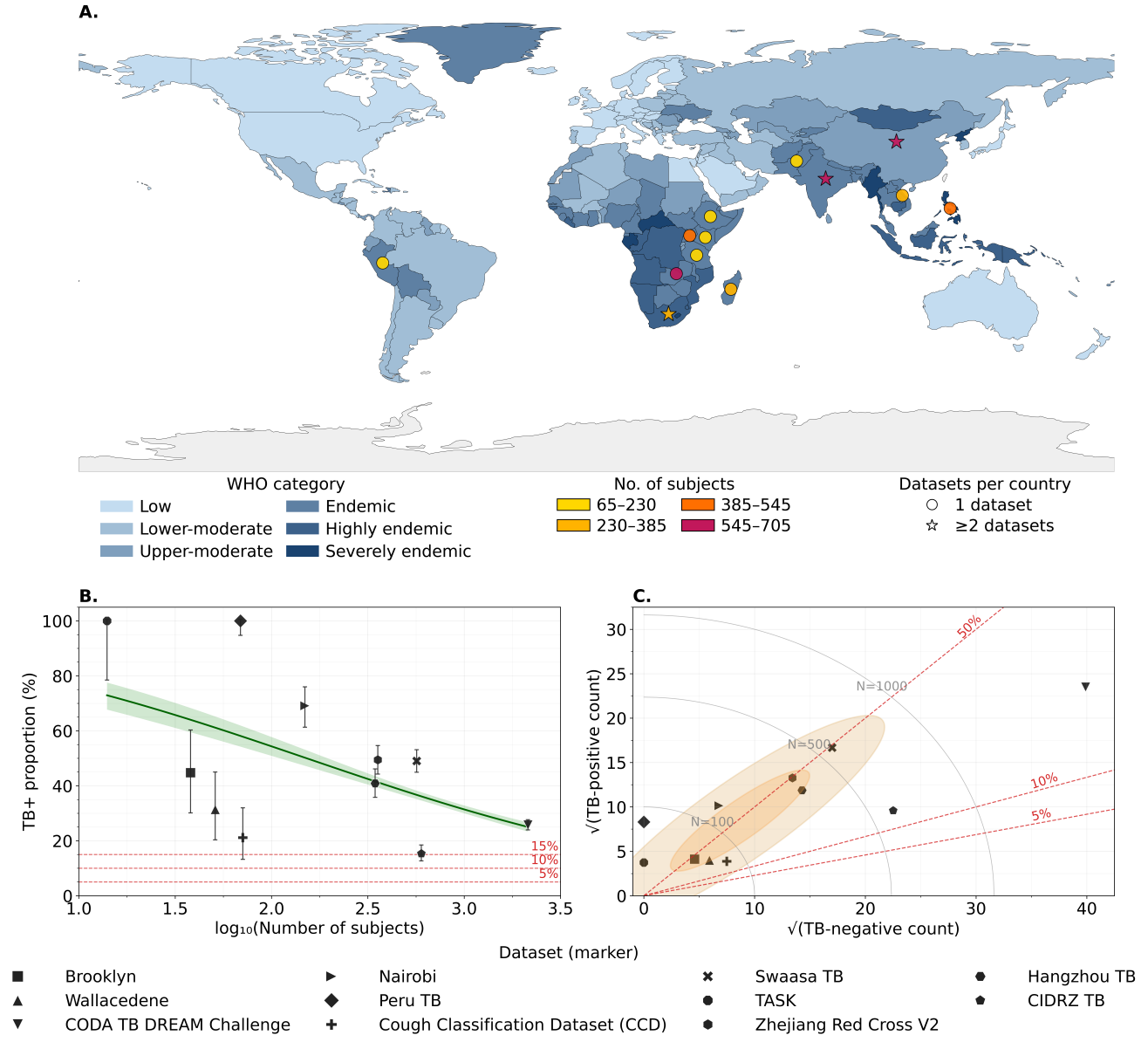

Figure 2. Geographic and epidemiological structure of tuberculosis cough datasets. A. Global distribution of available cough datasets overlaid on WHO TB incidence categories[50], shaded by incidence rate (per 100,000 population). Marker color encodes the number of enrolled participants, and marker shape distinguishes countries contributing a single dataset (circle) from those contributing two or more datasets (star). **B.** Dataset-level tuberculosis prevalence plotted against cohort size ($\log_{10}$ scale). Confidence-interval widths narrow systematically with increasing cohort size, reflecting reduced statistical uncertainty in larger studies. **C.** Class-count phase space showing square-root–scaled TB-positive and TB-negative subject counts. Orange ellipses (95 % robust covariance contours) highlight clustering within a 30–60 % prevalence band (IQR 28.6–59.3 %), spanning nearly two orders of magnitude in cohort size. Iso-prevalence rays (red dashed lines) and iso-sample-size arcs (gray curves) provide geometric reference. *Note:* Zhejiang Red Cross V2 supersedes V1 and is treated as the consolidated dataset (V2 only). Mayo TB lacked explicit TB-positive counts and is therefore excluded from **B** and **C** but retained in **A** for geographic completeness.

racy, 77.8% sensitivity, 81.3% specificity, AUC = 0.85). Pretraining improved specificity (95.8%) but reduced AUC (0.79)[12]. Xu et al.[46] combined CNN + Bi-LSTM, achieving 96.3% accuracy and 98.1% sensitivity. Ensemble approaches (ResNet50, GoogLeNet, Bi-LSTM) yielded 98.1% accuracy, 98.8% sensitivity, and 97.5% specificity[47]. Traditional classifiers remain relevant. Logistic regression on MFCC and handcrafted features reached AUC = 0.94 and 95% sensitivity[15]. ANN models achieved 92.3% accuracy[45]. Random Forests optimized for mobile deployment achieved 96.6% accuracy, 93.1% specificity, and 100% sensitivity[13]. Novel designs include capsule networks (97% accuracy, 98% sensitivity[53]) and NLP-inspired embeddings (AUC = 0.81, F1 = 79%[54]). Lightweight CNNs also show promise for low-resource deployment (91% accuracy[55]). Real-world evaluations highlight challenges. Sharma et al.[14] trained ResNet18 on passive coughs, achieving AUC = 0.79 $\pm$ 0.06 (balanced) and 0.82 $\pm$ 0.03 (unbalanced), but only 0.64 $\pm$ 0.05 on forced coughs, suggesting elicitation-dependent variability.

Tables 3, 4 provide a comprehensive overview of tuberculosis classification methods based on cough sound features, systematically organizing the datasets, feature extraction approaches, classifiers, and performance metrics discussed above.

3.5 Performance Landscape and Target Product Profile Alignment

The performance of AI-based TB cough classifiers varies widely across 13 studies using heterogeneous datasets. Because dataset composition, data extraction methods, and evaluation protocols differ, we avoid head-to-head comparisons and instead assess each report against WHO TPPs. To ensure consistent reporting, multi-class studies are reported using TB-class sensitivity, specificity, and AUC, for contextual placement alongside binary results.

Achievement rates against WHO benchmarks indicate a shift in the limiting factor (Fig. 3A). Under the 2014 TPP for triage (Sen $\geq$ 90%, Spec $\geq$ 70%)[56], 8/13 studies met the sensitivity threshold and 13/13 met specificity. Under the 2024 TPP for non-sputum diagnostics (Sen $\geq$ 65%, Spec $\geq$ 98%)[32], 13/13 exceeded the sensitivity minimum but only 2/13 (i.e, Xu et al.[48](journal article), and Yuan et al.[49](conference proceeding)) met the elevated specificity requirement. Thus, specificity, not sensitivity, is the primary constraint on deployment. Clinically, high specificity governs throughput and cost: false positives trigger confirmatory workflows, burden laboratories and patients, and, especially in low-prevalence settings, reduce positive predictive value.

Youden's J (percentage points) summarizes the sensitivity–specificity balance (Fig. 3B). Among binary classifiers, J spans 41.0–96.3, reflecting variation in datasets, prevalence, and calibration. For the two multi-class studies, TB-class J is mixed; with $n = 2$, no inference of superiority is warranted.

The specificity bottleneck is clearer in the false-positive-rate (FPR) versus AUC relationship (Fig. 3C). Although 9/11 systems achieve AUC $>$ 85%, only 1/11 operates at FPR $\leq$ 2% (equivalently, Spec $\geq$ 98%). Because AUC is threshold-invariant, high AUC does not guarantee performance at the high-specificity operating point required for large-scale triage; the practical challenge is choosing sufficiently selective decision thresholds.

The uncertainty in balanced performance decreased systematically with increasing cohort size (Fig. 3D). The 95% confidence-interval width closely followed an approximate inverse–square-root relationship with effective sample size, consistent with binomial sampling theory under sufficiently large and representative cohorts. This trend suggests that much of the apparent variability among studies reflects finite-sample statistical effects rather than intrinsic model instability. Smaller datasets exhibit broader confidence intervals, highlighting the dominant role of sample size in determining measurement reliability. Nonetheless, methodological heterogeneity across studies likely contributes additional variance beyond sampling error and should be considered when interpreting cross-study differences.

A geometric analysis (Fig. 3E) using Euclidean distance in (sensitivity, specificity) space shows that only 2 of 13 studies lie above the line of equal distance, i.e., closer to the 2024 TPP corner (65%, 98%) than to the 2014 corner (90%, 70%). Most studies remain closer to the 2014 target, underscoring that the last few percentage points of specificity are the hardest step toward 2024 compliance. Distances to the 2024 target fall within the range 19.2–35.3 (median 28.3), suggesting the field is approaching, but has not yet overcome, the specificity barrier.

Overall, heterogeneity in recording hardware, acoustic environments, and population characteristics likely contributes to dispersion, while partial reuse of public corpora introduces limited cross-study correlation yet enables some benchmarking. To translate AI-based cough triage into scalable practice, prospective multi-site validation, standardized TB-class reporting for multi-class models, and harmonized evaluation across shared benchmarks and new cohorts are needed to establish generalizability, and real-world utility.

4 Implementation Requirements and Stakeholder Framework for AI-Based TB Cough Analysis

Translating AI-based cough analysis from research prototypes to deployed triage tools requires more than al-

Table 3. Summary of TB classification methods based on cough sound features (Part 1). Notes: (1) For multi-class studies, only TB-specific metrics (sensitivity, specificity) are reported. (2) Publication type abbreviations used in the Method column are as follows: (J) = journal article; (C) = conference proceeding; (M) = master's thesis.

Method	Dataset Name	Features	Classifier	Classification Type	Acc	Sen	Spec	AUC
Pahar et al.[15](J)	Wallacedene	Log-filterbank energies, MFCC, ZCR, Kurtosis	Logistic Regression	Binary	–	93.0	95.0	0.9421
Pahar et al.[44](C)	TASK, Sarcos, Brooklyn, Wallacedene, Coswara, ComParE. *Pretraining:* Google Audio Set, Freesound, LibriSpeech, TASK	MFCC	Transfer learning of ResNet-50	Binary Multi-class	– 86.89	93 –	80 –	0.9245 –
Frost et al.[52](C)	Brooklyn, Wallacedene	Mel spectrograms	Bi-LSTM with attention	Binary	80.0	77.8	81.3	0.85
Frost et al.[12](M)	Brooklyn, Wallacedene; *Pretraining:* TASK, COUGHVID, AudioSet, FreeSound	Mel spectrograms	Pre-training Bi-LSTM	Binary	–	77.78	75.0	0.8632
Xu et al.[48](J)	Hangzhou TB	Mel spectrogram	CNN	Multi-class	94.32	97.73	99.43	–
Sharma et al.[14](J)	Nairobi	Scalograms	Pre-trained ResNet18	Binary	–	70±11	71±10	0.79±0.06
Bao et al.[47](C)	Zhejiang Red Cross V2	MFCC, ZCR, TSE, RMS, chroma_cens, Mel spectrogram	Ensemble: ResNet50, GoogLeNet, Bi-LSTM+GoogLeNet	Binary	98.05	98.82	97.49	0.99
Xu et al.[46](J)	Zhejiang Red Cross V1	MFCC, ZCR, Chroma CENS, RMS, short-time energy	RNN + CNN (Feature fusion)	Binary	96.33	98.13	94.99	0.95

Table 4. Summary of TB classification methods based on cough sound features (Part 2). Notes: (1) Certain evaluation metrics for Mahmood et al.[13], Yuan et al.[49], Yadav et al.[55], and Suda et al.[51] were computed using confusion matrices provided in the original works. (2) For multi-class studies, only TB-specific metrics (sensitivity, specificity, AUC) are reported. (3) The study by Pahar et al.[54] uses the same data collection protocol as the Wallacedene dataset[15,44] but reports 48 subjects (15 TB+, 33 TB-RD) instead of 51, indicating a likely subset. (4) Rajasekar et al.[53] use only the solicited cough subset of the CODA TB DREAM Challenge dataset[11,42], whereas Yadav et al.[55] incorporate both solicited and longitudinal coughs while limiting the maximum number of recordings per participant. (5) Publication type abbreviations used in the Method column are as follows: (J) = journal article; (C) = conference proceeding; (P) = preprint.

Method	Dataset Name	Features	Classifier	Classification Type	Acc	Sen	Spec	AUC
Mahmood et al.[13](J)	Mayo TB	MFCC, RMSE, Spectral features, ZCR, Mel spectrogram, Chroma variants, Tonnetz	Random Forest	Binary	96.55	100	93.1	–
Rajasekar et al.[53](J)	CODA TB DREAM Challenge	Histogram of Oriented Gradients	Capsule Network + FCNN	Binary	97.0	98.0	96.0	0.97
Yuan et al.[49](C)	Hangzhou TB	MFCC, Mel spectrogram	CNN	Multi-class	91.67	93.2	100	0.9659
Yadav et al.[55](C)	CODA TB DREAM Challenge	MFCC	CNN	Binary	91±1	84±3	95±2	0.91±0.005
Jember et al.[45](C)	Cough classification dataset (CCD)	MFCC	ANN	Binary	92.3	–	–	–
Suda et al.[51](P)	CODA TB Challenge	Mel spectrogram, MFCC, Spectral contrast	2D-CNN	Binary	87.87	89.17	85.72	0.88
Pahar et al.[54](C)	Wallacedene (Subset)	NLP-style cough embeddings, MFCC, ZCR, Kurtosis	LSTM	Binary	–	–	–	0.81

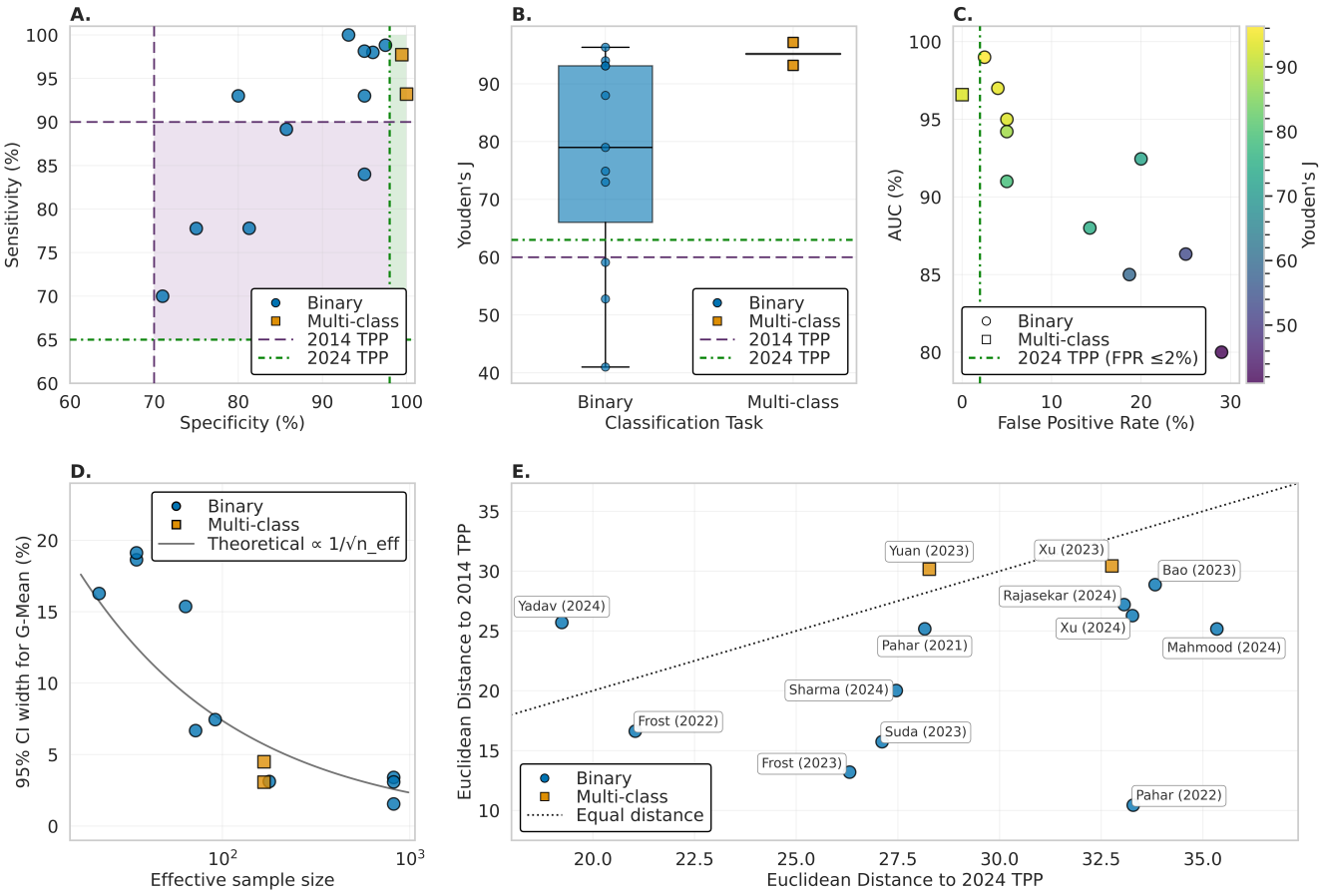

Figure 3. Performance landscape of AI-based tuberculosis cough classifiers relative to WHO Target Product Profiles (TPPs). A. Sensitivity versus specificity for binary (blue circles) and multi-class (orange squares) models, overlaid with WHO TPP thresholds for triage (2014; purple dashed) and non-sputum diagnostics (2024; green dash–dot). Most systems achieve high sensitivity but fall short of the 98 % specificity target. **B.** Youden's J statistic ($J = \mathrm{Sen} + \mathrm{Spec} - 100$) summarizes the sensitivity–specificity balance across task types. Binary classifiers show broad variability, whereas the two multi-class studies exhibit mixed TB-class performance. **C.** False-positive rate (FPR) versus area under the receiver operating characteristic curve (AUC), defined as $\mathrm{FPR} = 100 - \mathrm{Spec}\,(\%)$. Although most models achieve $\mathrm{AUC} > 85\,\%$, few operate at $\mathrm{FPR} \leq 2\,\%$ (equivalently, $\mathrm{Spec} \geq 98\,\%$), underscoring the specificity bottleneck at deployment-relevant thresholds. **D.** Uncertainty in balanced performance, expressed as the 95 % confidence-interval width for the geometric mean (G-Mean $= \sqrt{\mathrm{Sen} \times \mathrm{Spec}}$), plotted against effective sample size (N_{eff}, log scale). The observed inverse–square-root scaling ($\propto 1/\sqrt{N_{\mathrm{eff}}}$) follows binomial expectations, indicating that much of the apparent variability among studies reflects finite-sample effects rather than model instability. **E.** Geometric distance analysis in sensitivity–specificity space showing the proximity of each study to the WHO 2014 and 2024 TPP targets.

gorithmic accuracy. Experiences from digital health interventions show that impact at scale depends not only on model performance but also on rigorous external validation, seamless integration into clinical workflows, and sustainable governance and financing. This section synthesizes lessons from comparable tools, outlines stakeholder requirements and system-level challenges, highlights research gaps, and proposes a phased framework for translation.

4.1 Insights from Related Digital Health Interventions

Among digital health tools, AI-assisted chest X-ray interpretation provides the clearest precedent for scaling diagnostic AI. Following WHO's 2021 policy guidance recommending computer-aided detection (CAD) software[57], several countries have piloted integration into national programs with mixed results. A 2024 external validation of twelve commercial CAD products using South Africa's national TB prevalence survey revealed substantial variation across products, with performance consistently lower in older adults, people with previous TB, and individuals living with HIV[58]. Community-based studies confirmed that in real-world conditions, CAD often fails to meet the 2024 WHO's target product profile for TB screening ($\geq$90% sensitivity and $\geq$70% specificity)[59].

Beyond imaging tools, mobile applications that estimate TB risk from simple symptom and demographic inputs represent another class of lightweight screening solutions. The mTBScreen project harmonized over 900,000 records from national TB prevalence surveys to train machine-learning and Bayesian models, achieving a balanced accuracy of roughly 60% on secondary validation datasets[60]. A proof-of-concept mobile app was piloted with 30 health workers in Zambia using clinical vignettes, where it was generally well-received and considered an improvement over paper-based workflows, though usability differed across cadres[60]. Subsequent implementation analyses showed that scaling such tools requires resolving interoperability barriers, particularly the inconsistent structures and APIs of national electronic TB databases[60].

Together, these experiences demonstrate that technical promise does not guarantee public health impact. They highlight the importance of system-level integration and continuous validation across diverse populations.

4.2 Stakeholder Ecosystem and Requirements

Successful development and deployment of AI-based cough analysis depend on coordinated engagement across multiple stakeholders[61,62]. Effective implementation requires alignment among technology developers, public health authorities, healthcare providers, and community organizations to ensure clinical utility, ethical deployment, and scalability. Fig. 4 illustrates the envisioned patient journey, from symptom onset through smartphone-based AI triage to clinical confirmation and treatment initiation.

Health system integration. For AI-based cough analysis to be useful in practice, it must fit cleanly into existing clinical workflows. Community health workers need training not only on how to elicit and record coughs consistently, but also on how to interpret AI outputs and act on them within referral pathways. Patient trust depends on clear consent processes and strong safeguards against stigma or misuse of data. Importantly, AI screening systems must be linked to confirmatory diagnostics in a way that does not overload already-constrained laboratory services; otherwise, increased case-finding may inadvertently create downstream bottlenecks.

Technology platforms. Heterogeneity in CAD performance underscores the risks associated with device variability. Ensuring consistent deployment requires robust offline-first functionality, secure data pipelines, and standardized device specifications across platforms. In parallel, data governance frameworks must ensure compliance with national regulations on informed consent, data retention, and cross-border transfer. Recognizing these interconnected technical and ethical challenges, major technology organizations have begun directing their AI and mobile health Research and Development (R&D) efforts toward solutions that enhance cross-device consistency, on-device intelligence, and responsible data stewardship, thereby supporting scalable, equitable, and regulation-aligned deployment of diagnostic tools in real-world healthcare settings[40,63].

Regulators and policymakers. Current regulatory frameworks for digital health—such as FDA's Software as a Medical Device (SaMD)[64] and WHO's AI ethics guidance[65]—provide general principles but are not yet tailored to acoustic biomarkers or AI models that update over time. National TB Programs will need to determine how cough-based AI fits into their case-finding algorithms, procurement policies, and reporting systems to prevent fragmented deployment. Clear standards for validation, model updates, and integration with national surveillance systems will be essential for safe and consistent use.

Funders and governance bodies. Beyond pilot projects, sustainable financing, transparent data governance, and equitable benefit-sharing mechanisms are essential to ensure long-term scalability.

4.3 Cross-Cutting Challenges

Even with stakeholder alignment, recurring barriers remain.

Trust and acceptance. Health workers express greater confidence in interpretable AI systems that ex-

plain outputs rather than black-box classifications[66]. Patient concerns about privacy and stigma require transparent consent processes and community engagement.

Sustainability. Continuous validation and lifecycle management are essential. Programs must plan for model drift monitoring, device maintenance, and recurrent workforce training, supported by financing beyond initial donor funding.

Equity. Without deliberate design, digital tools risk exacerbating disparities. Vulnerable groups—including the elderly, PLH, and those without smartphone access—may be systematically excluded. Subgroup performance audits and inclusive design practices are needed to safeguard equity.

4.4 Critical Implementation Research Gaps

Although several methods have reported promising performance for AI-based cough analysis in controlled settings, no system has yet demonstrated consistent performance across diverse devices, and real-world recording conditions. A handful of models have met WHO screening thresholds, but these studies often relied on curated datasets, lacked external validation. As a result, further model refinement and rigorous field testing are still required before cough-AI can be considered clinically reliable. At present, no country has deployed cough-based AI for TB screening at scale, and all existing tools remain in research or pilot phases rather than routine clinical use. Against this backdrop, four implementation research gaps remain central to translation:

Prospective, multi-site effectiveness. No large-scale prospective studies have evaluated AI cough analysis in routine practice with subgroup analyses and predictive values at observed prevalence.

Economic evaluation. Rigorous cost-effectiveness studies that incorporate confirmatory testing loads, patient costs, and system throughput are needed to support policy adoption.

Health system impact. Effects on laboratory workflows, provider roles, and patient pathways must be systematically assessed before wide deployment.

Regulatory science. Demonstration projects with shared protocols and audit trails are needed to establish precedents for regulatory approval and oversight of acoustic biomarkers and adaptive models.

4.5 A Phased Implementation Framework

To address these challenges, we propose a phased, gate-based framework emphasizing explicit milestones rather than fixed timelines.

Phase 0: Preparation. Map clinical pathways, secure ethics and data agreements, define device specifications, and assess site readiness.

Phase 1: Controlled pilots. Deploy in select facilities to evaluate usability, workflow integration, and diagnostic performance under local conditions.

Phase 2: Adaptive scale-up. Expand using pragmatic trial designs; monitor model drift; recalibrate with local data; integrate with surveillance systems; conduct economic evaluations.

Phase 3: System integration. Embed within national TB strategies, institutionalize quality assurance, establish transparent reporting of model versions and thresholds, and secure sustainable financing.

At every phase, safeguards must ensure fairness, interpretability, privacy, and community trust. In summary, continuous validation, actionable workflows, and program stewardship—not algorithms alone—will determine whether AI-based cough analysis delivers equitable gains in TB case finding. As the experience with CAD and other digital interventions shows, technical innovation delivers public health impact only when embedded within robust systems that prioritize trust, sustainability, and equity.

5 Limitations and Future Directions

5.1 Limitations

5.1.1 Data scarcity and variability

The most fundamental barrier to cough-based AI for TB triage is the paucity of large, standardized datasets. Existing collections are fragmented across geographies, devices, and protocols, with no universally accepted acquisition guidelines. As a result, studies report heterogeneous and often non-comparable results[22,68]. Variability in recording environments, microphone types, sampling rates, and patient behaviors introduces distribution shifts that degrade model performance on unseen data. In addition, most datasets are skewed toward adults from a handful of TB-endemic countries, with limited representation of children, women, and high-burden but under-studied regions. This scarcity and imbalance fundamentally constrain both the robustness and the generalizability of current classifiers.

5.1.2 Lack of generalization

Models that achieve high performance on internal validation consistently falter when deployed in new contexts. Cough recordings differ substantially across devices, environments, and patient groups, and models trained in one setting frequently overfit to local artifacts rather than genuine disease signatures. Such domain shifts result in marked accuracy losses when models are transferred across populations. Two complementary strategies are required: (i) development of generalizable models through domain adaptation, transfer learning, and systematic augmentation; and (ii) context-specific fine-tuning for particular populations, workflows, or hardware. Generalizable models support scalability, while targeted adaptation provides a practical route for near-term deployment.

Figure 4. An individual uses an AI-powered tool to analyze her cough and receives triage advice, ranging from reassurance to scheduling a clinic appointment. Such systems could enable timely tuberculosis (TB) triage in community and home settings. Inspired by[67] and generated with the assistance of ChatGPT (OpenAI, 2025). All scenes are simulated for illustrative purposes; the positions of medical instruments and actions are not intended to represent actual clinical practice.

5.1.3 Interpretability challenges

Deep learning approaches to cough analysis remain opaque, limiting their clinical credibility[69]. Without interpretability, it is impossible to verify that classifiers rely on physiologically meaningful acoustic signatures (e.g., burst duration or spectral profiles) rather than spurious noise. This opacity undermines clinical trust, obstructs regulatory approval, and weakens scientific insight. Transparent models are therefore not optional but essential for clinical translation.

5.1.4 Clinical validation and deployment

The most consequential gap is determining whether existing models are clinically reliable—specifically, whether they achieve sufficiently robust sensitivity, specificity, and predictive values under real-world conditions. To date, nearly all algorithms have been evaluated on small, retrospective datasets, leaving their true performance in routine practice unknown. Differences in disease prevalence, comorbidities, device hardware, posture, and user behavior can alter acoustic signatures and significantly affect accuracy outside controlled settings. Although a few studies report sensitivities approaching 96% under ideal conditions—meeting or exceeding WHO thresholds for triage[44]—no large-scale prospective or longitudinal field trials have been conducted. Until such studies demonstrate consistent performance across diverse populations and deployment environments, the clinical utility, safety, and cost-effectiveness of cough-based triage remain uncertain.

5.2 Future Directions

5.2.1 Standardized large-scale datasets

Progress will require the creation of large, standardized, and globally representative repositories of cough sounds. Just as the Foundation for Innovative New Diagnostics (FIND) biobanks transformed molecular diagnostic evaluation[70], a global acoustic data bank is essential to advance non-sputum diagnostics. Such an initiative must be coordinated by WHO, national TB programs, and clinical consortia, with universal acquisition protocols specifying device types, recording duration, segmentation, and noise control. Equitable inclusion across age, sex, geography, and comorbidities is mandatory to prevent biased models. Open-access, well-annotated repositories would accelerate collaborative benchmarking and establish fair comparative standards.

5.2.2 Domain adaptation and robustness

Robustness across settings cannot be assumed—it must be engineered. Domain adaptation methods, including adversarial alignment, transfer learning, and unsupervised distribution matching, are indispensable for enabling models to generalize across populations and devices. Pre-training on large respiratory sound corpora followed by TB-specific fine-tuning has already demonstrated benefit. Data manipulation strategies that simulate noise, microphone variability, and recording artifacts should be standard practice. Only models hardened against such variability will prove reliable in real-world deployment.

5.2.3 Explainable AI and model interpretability

Translation into practice demands explainability at the core of system design. Techniques such as attention mechanisms, saliency mapping, and feature attribution clarify which spectral or temporal components of a cough drive predictions[71]. These methods ensure that models rely on clinically meaningful signals, facilitate regulatory approval, and strengthen clinician trust. Moreover, interpretable frameworks may uncover novel acoustic biomarkers, transforming cough analysis from a classification tool into a discovery platform.

5.2.4 Rigorous clinical validation

Clinical translation will not occur without large-scale, prospective trials. Studies must span community screening, HIV/TB co-management clinics, correctional facilities, and urban slums, while explicitly evaluating performance in children, people with HIV, and those with COPD, diabetes, or undernutrition. Endpoints should extend beyond sensitivity and specificity to include patient outcomes, program throughput, and cost-effectiveness. Regulatory engagement and workflow integration must begin in parallel with trial design. Beyond triage, cough acoustics may offer insight into disease severity and treatment monitoring, but such applications remain speculative until rigorously tested.

5.3 Beyond sputum: other-modal strategies for TB

Cough represents one of several promising non-sputum signals. Breathomics, lung sound analysis, symptom-based models, skin-based sensing, and EHR-derived algorithms all offer complementary pathways. Individually, these modalities remain constrained by reproducibility or scalability, but their integration may deliver robust solutions. Breathomics and e-nose devices have shown variable diagnostic performance, with some studies reporting sensitivities above 90% and specificities above 85% (e.g.,[21,22]), while others report more modest results due to heterogeneous sensors and limited clinical trials[23,24]. Lung sound analysis aligns conceptually with cough but requires contact-based sensors and is noise-sensitive[34]. Symptom-based algorithms[72,73] and EHR-driven models[74] are under exploration, while skin-based approaches remain nascent[75].

Integration of acoustic data and imaging modalities, supported by advances in multimodal AI[76,77], is demonstrating the potential to offer a more credible path to robust non-sputum testing. Recent multimodal efforts

exemplify this direction: the Swaasa AI platform integrates cough features and symptom metadata for clinical screening[16]; deep learning frameworks combining cough-derived spectrograms with chest X-ray and CT imaging have been applied for multi-disease classification[76,77]; and other real-world systems leverage cough and patient metadata to enhance TB prediction[78]. Together, these studies illustrate how cough-centered models can evolve into multimodal, non-sputum triage platforms.

5.4 Ethical, regulatory, and equity considerations

As AI-based cough triage advances, ethical safeguards, regulatory oversight, and equity must be embedded from inception. Cough audio constitutes sensitive biometric data and demands rigorous governance for consent, retention, and sharing. Compliance with General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), and national laws is mandatory. Real-time systems that discard audio after analysis reduce risk, but continuous-learning pipelines require ongoing consent and transparent governance. Regulatory approval will hinge on frameworks such as FDA's Software as a Medical Device (SaMD)[79,80] and WHO's in vitro diagnostic prequalification[81,82]. Model fairness must be systematically evaluated: biases by age, sex, geography, or comorbidities cannot be tolerated in deployment[83–86]. Community engagement is equally essential. Participatory design and local co-development strengthen trust, cultural fit, and adoption in TB-endemic communities[87–89]. Equitable deployment must remain central. Without deliberate targeting, digital health tools risk widening disparities[90–93]. Ministries of health, NGOs, and funders must prioritize deployment in high-burden, resource-limited settings to ensure maximal impact.

5.5 Integration into TB programs

The ultimate test of AI-based cough analysis lies in its integration into TB control strategies. These systems are best positioned as triage tools to flag individuals for confirmatory testing by GeneXpert or chest radiography. Deployment research must therefore evaluate usability, acceptability, and cost-effectiveness, alongside diagnostic accuracy. Health-economic studies should account for both direct implementation costs and indirect benefits such as earlier detection, reduced transmission, and improved treatment outcomes. Integration requires interoperability with digital health infrastructures, following standards such as Health Level Seven – Fast Healthcare Interoperability Resources (HL7 FHIR), while ensuring offline functionality and local language support for low-resource settings. Pilot studies should inform workflow design, training needs, and strategies for continuous model monitoring[62]. Privacy safeguards—including de-identification, encrypted transfer, and role-based access—must be instituted before scale-up. Equally important are clear referral pathways for individuals flagged positive, ensuring timely confirmatory testing and treatment. Translation into national programs will succeed only if models are continuously monitored, periodically updated, and aligned with emerging global guidance on digital diagnostics. If achieved, AI-based cough testing has the potential not to replace sputum assays, but to transform who is tested and when—shifting TB programs toward earlier, broader, and more equitable detection.

6 Conclusion

Tuberculosis continues to challenge global health, and reliance on sputum-based diagnostics constrains timely case detection in community and primary care settings. The potential value of cough-based triage lies in its accessibility, low cost, and scalability—smartphone-based tools could shift initial screening from centralized laboratories to community settings, extending reach to populations that face the greatest diagnostic delays. Advances in artificial intelligence and mobile health technologies have positioned cough analysis as a promising non-invasive approach.

Proof-of-concept studies demonstrate that TB-associated acoustic patterns can often be detected with high sensitivity; however, specificity and overall clinical reliability remain key limitations, and existing systems cannot yet be considered ready for routine use. Even the most rigorous evaluations to date have been conducted under controlled clinical conditions with selected patient groups, offering only partial insight into real-world performance. Evidence remains limited for children, people living with HIV, and individuals with coexisting respiratory conditions—precisely the groups where TB burden is highest and diagnostic gaps are most pronounced.

Field studies show that smartphone-based cough collection is technically feasible and generally acceptable to patients in primary care settings. These early implementations indicate that the basic workflow can be integrated with minimal infrastructure and without specialized personnel. However, important questions remain about health worker training needs, durability of diagnostic performance outside supervised research environments, and fit within existing triage pathways in diverse low-resource settings. Early deployments suggest operational viability, but uncertainties persist regarding scalability, quality assurance, and long-term usability within community health programs.

Realizing the promise of cough-based TB screening will require progress on two fronts: developing reliable and affordable diagnostic models that meet WHO performance criteria across diverse populations and contexts, and demonstrating that these tools can be deployed effectively, acceptably, and sustainably in resource-

limited health systems. Achieving this vision will depend on building standardized and globally representative datasets, conducting rigorous multi-site prospective validation studies, and establishing clear regulatory and ethical frameworks tailored to audio-based diagnostics.

If these challenges are addressed, AI-assisted cough analysis could transform TB case finding and establish a scalable paradigm for symptom-based diagnostics across respiratory diseases—shifting screening from the laboratory to the community, where it is needed most.

7 Funding

Gutta J. Chowdary was funded by the 2025 Cockrell School of Engineering Academic Development Fund (PI: Panpradist). Nuttada Panpradist and Varunya Sakpuntoon were supported by the University of Texas at Austin Cockrell School of Engineering/Biomedical Engineering Startup Fund. The funding sources had no role in the study design, analysis, or interpretation; manuscript preparation; or the decision to submit this work for publication.

A Performance Metric Formulae

The following formulas define the standard classification metrics used to evaluate TB testing models based on cough sound analysis. To calculate these metrics, the following terms from the confusion matrix are used: **True Positives (TP)** are TB-positive cases correctly identified by the model, while **True Negatives (TN)** are TB-negative cases correctly identified. **False Positives (FP)** refer to TB-negative cases incorrectly classified as TB-positive, and **False Negatives (FN)** are TB-positive cases incorrectly classified as TB-negative.

- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$
- **Sensitivity (True Positive Rate)** = $\frac{TP}{TP+FN}$
- **Specificity (True Negative Rate)** = $\frac{TN}{TN+FP}$
- **Precision** = $\frac{TP}{TP+FP}$
- **F1 Score** = $2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = \frac{2TP}{2TP+FP+FN}$
- **AUC-ROC**: The Area Under the ROC Curve is computed by plotting the True Positive Rate (Sensitivity) against the False Positive Rate (FPR = $\frac{FP}{FP+TN}$) at various thresholds. It does not have a single closed-form formula but is computed numerically.

References

1. World Health Organization. Global tuberculosis report 2024 (2024). Licence: CC BY-NC-SA 3.0 IGO.
2. World Health Organization. Tuberculosis. WHO Fact Sheets (2024). Accessed: 17 July 2025.
3. World Health Organization. Tuberculosis deaths and disease increase during the COVID-19 pandemic (2022).
4. Williams, P. M. Tuberculosis—united states, 2023. *MMWR. Morb. mortality weekly report* **73** (2024).
5. for Disease Control, C., Prevention *et al.* Provisional 2024 tuberculosis data, united states (2024).
6. The Guardian. Kansas faces largest U.S. tuberculosis outbreak in decades amid public health challenges (2025). Accessed: 2025-07-22.
7. Hanson, C., Osberg, M., Brown, J., Durham, G. & Chin, D. P. Finding the missing patients with tuberculosis: lessons learned from patient-pathway analyses in 5 countries. *The J. infectious diseases* **216**, S686–S695 (2017).
8. Centers for Disease Control and Prevention. Clinical and laboratory diagnosis | tb | cdc (2024). Accessed: 2025-07-22.
9. Lawn, S. D. *et al.* Determine tb-lam lateral flow urine antigen assay for hiv-associated tuberculosis: recommendations on the design and reporting of clinical studies. *BMC infectious diseases* **13**, 407 (2013).
10. FIND. Diagnostic accuracy study published on next-generation urine test to detect tuberculosis in hiv-positive people (2023). Accessed: 2025-07-22.
11. Huddart, S. *et al.* A dataset of solicited cough sound for tuberculosis triage testing. *Sci. Data* **11**, 1149 (2024).
12. Frost, G. *Deep learning based methods for tuberculosis cough classification*. Ph.D. thesis, Stellenbosch: Stellenbosch University (2023).
13. Mahmood, H., Iftikhar, M., Wali, A., Ali, A. & Gulzar, M. A novel cascaded approach for classification of tuberculosis using cough audio in real-time environment. *IEEE Access* **12**, 191980–191993 (2024).
14. Sharma, M. *et al.* TBscreen: A passive cough classifier for tuberculosis screening with a controlled dataset. *Sci. Adv.* **10**, eadi0282 (2024).
15. Pahar, M. *et al.* Automatic cough classification for tuberculosis screening in a real-world environment. *Physiol. Meas.* **42**, 105014 (2021).
16. Yellapu, G. D. *et al.* Development and clinical validation of swaasa ai platform for screening and prioritization of pulmonary tb. *Sci. Reports* **13**, 4740 (2023).

17. Kendall, E. A., Shrestha, S. & Dowdy, D. W. The epidemiological importance of subclinical tuberculosis. a critical reappraisal. *Am. journal respiratory critical care medicine* **203**, 168–174 (2021).

18. Bulterys, M. A. *et al.* Point-of-care urine lam tests for tuberculosis diagnosis: a status update. *J. clinical medicine* **9**, 111 (2019).

19. Chang, A. *et al.* Circulating cell-free rna in blood as a host response biomarker for detection of tuberculosis. *Nat. Commun.* **15**, 4949 (2024).

20. Oreskovic, A. *et al.* Diagnosing pulmonary tuberculosis by using sequence-specific purification of urine cell-free dna. *J. clinical microbiology* **59**, 10–1128 (2021).

21. Fu, L. *et al.* A cross-sectional study: a breathomics based pulmonary tuberculosis detection method. *BMC Infect. Dis.* **23**, 148 (2023).

22. Ketchanji Mougang, Y. C. *et al.* On-field test of tuberculosis diagnosis through exhaled breath analysis with a gas sensor array. *Biosensors* **13**, 570 (2023).

23. Xu, R. *et al.* Breathomics for diagnosing tuberculosis in diabetes mellitus patients. *Front. Mol. Biosci.* **11**, 1436135 (2024).

24. Bijker, E. M. *et al.* Exhaled breath analysis: A promising triage test for tuberculosis in young children. *Tuberculosis* **149**, 102566 (2024).

25. Church, E. C. *et al.* Oral swabs with a rapid molecular diagnostic test for pulmonary tuberculosis in adults and children: a systematic review. *The Lancet Glob. Heal.* **12**, e45–e54 (2024).

26. Shastry, P. *et al.* Advancing chronic tuberculosis diagnostics using vision-language models: A multi modal framework for precision analysis. *arXiv preprint arXiv:2503.14536* (2025).

27. Organization, W. H. *et al.* Target product profile for tuberculosis diagnosis and detection of drug resistance. In *Target product profile for tuberculosis diagnosis and detection of drug resistance* (2024).

28. Organization, W. H. *Target product profile for tuberculosis screening tests* (World Health Organization, 2025).

29. Mai, Y. *et al.* Methods for assessing cough sensitivity. *J. thoracic disease* **12**, 5224 (2020).

30. Rudraraju, G. *et al.* Cough sound analysis and objective correlation with spirometry and clinical diagnosis. *Informatics Medicine Unlocked* **19**, 100319 (2020).

31. Mother, D., Liberati, A., Tetzlaff, J. & Altman, D. G. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS Med* **6**, e1000097 (2009).

32. World Health Organization. Target product profile for tuberculosis diagnosis and detection of drug resistance. Tech. Rep., World Health Organization, Geneva, Switzerland (2024).

33. Behera, D. Text book of Pulmonary Medicine. *Indian J. Chest Dis. Allied Sci* **52**, 173 (2010).

34. Shakeel, M., Mushtaq, Z., Gretschmann, P., Aziz, S. & Khan, M. U. Support vector machine-based diagnosis of tuberculosis. In *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, 1–6 (IEEE, 2021).

35. Tahir, A., Malik, H. & Chaudhry, M. U. Multi-classification deep learning models for detecting multiple chest infection using cough and breath sounds. In *Deep Learning for Multimedia Processing Applications*, 216–249 (CRC Press, 2024).

36. Kumar, A., Abhishek, K., Chakraborty, C. & Kryvinska, N. Deep learning and internet of things based lung ailment recognition through coughing spectrograms. *IEEE Access* **9**, 95938–95948 (2021).

37. Kumar, A. *et al.* Towards cough sound analysis using the internet of things and deep learning for pulmonary disease prediction. *Transactions on emerging telecommunications technologies* **33**, e4184 (2022).

38. Lee, G. O. *et al.* Cough dynamics in adults receiving tuberculosis treatment. *PloS one* **15**, e0231167 (2020).

39. Pahar, M., Miranda, I., Diacon, A. & Niesler, T. Automatic non-invasive cough detection based on accelerometer and audio signals. *J. Signal Process. Syst.* **94**, 821–835 (2022).

40. Baur, S. *et al.* Hear–health acoustic representations. *arXiv preprint arXiv:2403.02522* (2024).

41. Google Health AI. Google + CIDRZ Health AI Evaluation Zambia. https://www.kaggle.com/datasets/googlehealthai/google-health-ai (2025). Kaggle dataset.

42. Huddart, S. *et al.* Solicited cough sound analysis for tuberculosis triage testing: the coda tb dream challenge dataset. *MedRxiv* (2024).

43. Botha, G. *et al.* Detection of tuberculosis by automatic cough sound analysis. *Physiol. measurement* **39**, 045005 (2018).

44. Pahar, M. *et al.* Automatic tuberculosis and covid-19 cough classification using deep learning. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–9 (IEEE, 2022).

45. Jember, A. F., Ayano, Y. M. & Debelee, T. G. Robust cough analysis system for diagnosis of tuberculosis using. In *Pan-African Conference on Artificial*

Intelligence: First Conference, PanAfriCon AI 2022, Addis Ababa, Ethiopia, October 4–5, 2022, Revised Selected Papers, 3 (Springer Nature, 2023).

46. Xu, W. *et al.* Feature fusion method for pulmonary tuberculosis patient detection based on cough sound. *Plos one* **19**, e0302651 (2024).

47. Bao, X. *et al.* Tuberculosis detection based on cough sounds: a multi-model voting mechanism. In *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–6 (IEEE, 2023).

48. Xu, W., Yuan, H., Lou, X., Chen, Y. & Liu, F. Dmr-net based tuberculosis screening with cough sound. *IEEE Access* **12**, 3960–3968 (2023).

49. Yuan, H. *et al.* Tuberculosis screening with cough sounds using the deep learning models. In *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 444–448 (IEEE, 2023).

50. Organization, W. H. *et al.* Who global lists of high burden countries for tuberculosis (tb), tb/hiv and multidrug/rifampicin-resistant tb (mdr/rr-tb), 2021–2025: background document. In *WHO global lists of high burden countries for tuberculosis (TB), TB/HIV and multidrug/rifampicin-resistant TB (MDR/RR-TB), 2021–2025: background document* (2021).

51. Suda, C. Early detection of tuberculosis with machine learning cough audio analysis: Towards more accessible global triaging usage. *arXiv preprint arXiv:2310.17675* (2023).

52. Geoffrey T. Frost and Grant Theron and Thomas Niesler. TB or not TB? Acoustic cough analysis for tuberculosis classification. In *Interspeech 2022*, 2448–2452, DOI: 10.21437/Interspeech.2022-383 (2022).

53. Rajasekar, S. J. S. *et al.* Detection of tuberculosis using cough audio analysis: a deep learning approach with capsule networks. *Discov. Artif. Intell.* **4**, 77 (2024).

54. Pahar, M., Theron, G. & Niesler, T. Automatic tuberculosis detection in cough patterns using nlp-style cough embeddings. In *2022 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–6 (IEEE, 2022).

55. Yadav, J., Varde, A. S., Liu, H., Antoniou, G. & Xie, L. Audiovisual multimodal cough data analysis for tuberculosis detection. In *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1–8 (IEEE, 2024).

56. World Health Organization. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting (2014). WHO/HTM/TB/2014.18.

57. World Health Organization. Who consolidated guidelines on tuberculosis: Module 2: screening: systematic screening for tuberculosis disease (2021).

58. Qin, Z. Z. *et al.* Computer-aided detection of tuberculosis from chest radiographs in a tuberculosis prevalence survey in south africa: external validation and modelled impacts of commercially available artificial intelligence software. *The Lancet Digit. Heal.* **6**, e605–e613 (2024).

59. Scott, A. J. *et al.* Clinical evaluation of computer-aided digital x-ray detection of pulmonary tuberculosis during community-based screening or active case-finding: a case–control study. *The Lancet Glob. Heal.* **13**, e517–e527 (2025).

60. German Alliance for Global Health Research (GLOHRA). Ai tb screening tool: Development of a novel, easy-to use digital tuberculosis screening tool informed by machine learning approaches. https://www.globalhealth.de/fileadmin/user_upload/Documents/241203_case_study_AI_TB_Screening_Tool.pdf. Accessed: 21 September 2025.

61. Jaganath, D. *et al.* Accelerating cough-based algorithms for pulmonary tuberculosis screening: Results from the CODA TB DREAM Challenge. *medRxiv* 2024–05 (2024).

62. Isangula, K. G. & Haule, R. J. Leveraging ai and machine learning to develop and evaluate a contextualized user-friendly cough audio classifier for detecting respiratory diseases: protocol for a diagnostic study in rural tanzania. *JMIR research protocols* **13**, e54388 (2024).

63. Lenharo, M. Google ai could soon use a person's cough to diagnose disease. *Nature* **628**, 19–20 (2024).

64. U.S. Food and Drug Administration. Software as a medical device (samd): Clinical evaluation. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation (2025). Accessed: August 2025.

65. World Health Organization. Ethics and governance of artificial intelligence for health: Who guidance. https://www.who.int/publications/i/item/9789240029200 (2021).

66. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. Ai in health and medicine. *Nat. medicine* **28**, 31–38 (2022).

67. Zimmer, A. J. *et al.* Making cough count in tuberculosis care. *Commun. medicine* **2**, 83 (2022).

68. Hansun, S. *et al.* Diagnostic performance of artificial intelligence–based methods for tuberculosis detection: Systematic review. *J. Med. Internet Res.* **27**, e69068 (2025).

69. Ijaz, A. *et al.* Towards using cough for respiratory disease diagnosis by leveraging artificial intelligence: A survey. *Informatics Medicine Unlocked* **29**, 100832 (2022).

70. Foundation for Innovative New Diagnostics (FIND). Tuberculosis specimen bank. https://www.finddx.org/what-we-do/cross-cutting-workstreams/biobank-services/find-specimen-bank/tuberculosis-samples/ (2025). Accessed: 19 August 2025.

71. Chen, Z. *et al.* Exploring explainable ai features in the vocal biomarkers of lung disease. *Comput. Biol. Medicine* **179**, 108844 (2024).

72. Dolker, T. & Ramakrishnudu, T. Hybrid cnn and lstm network for communicable disease prediction. In *2023 IEEE 7th Conference on Information and Communication Technology (CICT)*, 1–6 (IEEE, 2023).

73. Senthilmurugan, M., Latha, M. & Chinnaiyan, R. Analysis and prediction of tuberculosis using machine learning classifiers. In *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 1–4 (IEEE, 2021).

74. Wang, M. *et al.* Clinical assistant decision-making model of tuberculosis based on electronic health records. *BioData Min.* **16**, 11 (2023).

75. Makhubela, P. C., Rohwer, E. R. & Naudé, Y. Detection of tuberculosis-associated compounds from human skin by gcxgc-tofms. *J. Chromatogr. B* **1231**, 123937 (2023).

76. Malik, H. *et al.* Deep learning-based classification of chest diseases using x-rays, ct scans, and cough sound images. *Diagnostics* **13**, 2772 (2023).

77. Malik, H. & Anees, T. Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds. *Plos one* **19**, e0296352 (2024).

78. Kafentzis, G. P. *et al.* Predicting tuberculosis from real-world cough audio recordings and metadata. *arXiv preprint arXiv:2307.04842* (2023).

79. U.S. Food and Drug Administration, Digital Health Center of Excellence. Software as a medical device (samd). Web page (2025). Accessed August 3, 2025; published online by FDA.

80. U.S. Food and Drug Administration, Digital Health Center of Excellence. Artificial intelligence in software as a medical device (samd). Web page (2025). Accessed August 3, 2025; published online by FDA.

81. World Health Organization. Who announces first prequalification of a tuberculosis diagnostic test (2024). Accessed: 2025-08-03.

82. World Health Organization. Public announcement to tb in vitro diagnostics manufacturers, procurement agencies and national tb programmes on inclusion of who prequalification for tb in vitro diagnostics (2021). Accessed: 2025-08-03.

83. Cross, J. L., Choma, M. A. & Onofrey, J. A. Bias in medical ai: Implications for clinical decision-making. *PLOS Digit. Heal.* **3**, e0000651 (2024).

84. Haider, S. A. *et al.* The algorithmic divide: a systematic review on ai-driven racial disparities in healthcare. *J. racial ethnic health disparities* 1–30 (2024).

85. Mittermaier, M., Raza, M. M. & Kvedar, J. C. Bias in ai-based models for medical applications: challenges and mitigation strategies. *NPJ Digit. Medicine* **6**, 113 (2023).

86. Hasanzadeh, F. *et al.* Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit. Medicine* **8**, 154 (2025).

87. Ashesh, A., Mehra, C., Madan, V. & Furin, J. Reimagining community engagement in tb elimination: a perspective from the field. *IJTLD open* **2**, 3–5 (2025).

88. Chavez-Rimache, L., Ugarte-Gil, C. & Brunette, M. J. The community as an active part in the implementation of interventions for the prevention and control of tuberculosis: a scoping review. *medRxiv* (2023).

89. Boulanger, R. F. *et al.* Engaging communities in tuberculosis research. *The Lancet infectious diseases* **13**, 540–545 (2013).

90. Abràmoff, M. D. *et al.* Considerations for addressing bias in artificial intelligence for health equity. *NPJ digital medicine* **6**, 170 (2023).

91. Marko, J. G. O., Neagu, C. D. & Anand, P. B. Examining inclusivity: the use of ai and diverse populations in health and social care: a systematic review. *BMC Med. Informatics Decis. Mak.* **25**, 57 (2025).

92. Celi, L. A. *et al.* Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS digital health* **1**, e0000022 (2022).

93. Green, B. L., Murphy, A. & Robinson, E. Accelerating health disparities research with artificial intelligence. *Front. Digit. Heal.* **6**, 1330160 (2024).