Section: Miscellaneous

Review Article

DETECTION FOR NEW BIOMARKERS OF TUBERCULOSIS INFECTION ACTIVITY USING MACHINE LEARNING METHODS

Rakesh Ranjan Swain[1], Bhanjan Kumar Meher[2], Saroj Ranjan Mahanty[3], Sasmita Mehe[4]

[1] *Associate Professor Department of General Surgery of Bhima Bhoi, Medical College and Hospital, Balangir, Odisha, India.*
[2] *Associate Professor Department of General Surgery of Bhima Bhoi, Medical College and Hospital, Balangir, Odisha, India.*
[3] *Associate Professor Department of Medicine Bhima Bhoi, Medical College and Hospital, Balangir, Odisha, India.*
[4] *Associate Professor Department of Pulmonary Medicine, Bhīma Bhoi, Medical College and Hospital, Balangir, Odisha, India.*

Received *: 05/04/2026*
Received in revised form *: 22/05/2026*
Accepted *: 09/06/2026*

Corresponding Author:
Dr. Sasmita Meher
Associate Professor Department of Pulmonary Medicine, Bhīma Bhoi, Medical College and Hospital, Balangir, Odisha, India.
Email: dr.sasmitameher@gmail.com

DOI: 10.70034/ijmedph.2026.2.593

Source of Support: Nil,
Conflict of Interest: None declared

Int J Med Pub Health
2026; 16 (2); 3596-3601

ABSTRACT

Background: This review has been prepared under the Department of Pulmonary Medicine, Bhima Bhoi Medical College and Hospital, with a focus on emerging diagnostic approaches for tuberculosis. As tuberculosis continues to pose a significant public health challenge in India, the Department is committed to promoting research and academic activities aimed at improving early diagnosis, treatment monitoring, and disease control. The present review explores the application of omics technologies and artificial intelligence-based methods for distinguishing latent tuberculosis infection (LTBI) from active tuberculosis (ATB), highlighting their potential clinical utility in pulmonary medicine. Latent tuberculosis infection (LTBI) is a hidden form of tuberculosis that can later develop into active tuberculosis (ATB). Current tests, such as interferon-gamma release assays (IGRAs), cannot reliably distinguish LTBI from ATB.

Materials and Methods: Researchers use transcriptomics (gene expression) and proteomics (protein analysis) to identify biomarkers linked to LTBI and ATB. Machine learning (ML) techniques, including feature selection, deep learning, and multimodal data integration, help analyze these complex datasets and build diagnostic models.

Results: ML-based models using gene and protein biomarkers show better accuracy than traditional tests for differentiating LTBI from ATB. Combining multiple data types, such as molecular markers, immune responses, and imaging data, further improves diagnostic performance. These findings support the development of simple qRT-PCR biomarker tests for clinical use.

Conclusion: Integrating omics technologies with AI and machine learning offers a promising approach for early and accurate tuberculosis diagnosis. This review compares current biomarkers, ML methods, and validation strategies, while highlighting challenges such as limited external validation and study heterogeneity. Future research should focus on improving the clinical applicability of these diagnostic tools.

Keywords: Latent Tuberculosis Infection, Active Tuberculosis, Machine Learning, Artificial Intelligence, Transcriptomics, Proteomics, Biomarkers, qRT-PCR, Pulmonary Medicine, Bhima Bhoi Medical College and Hospital.

INTRODUCTION

Tuberculosis (TB) remains a major global health challenge, with latent tuberculosis infection (LTBI) serving as a large reservoir for future active TB (ATB) cases. Although TB incidence and mortality have declined significantly in India and many other regions, LTBI continues to pose important clinical and epidemiological concerns. The COVID-19 pandemic disrupted TB control programs and may have increased the risk of LTBI reactivation through immune dysregulation.[1]

Since LTBI cannot be detected directly, diagnosis relies on host immune responses to Mycobacterium tuberculosis antigens. Both CD4+ and CD8+ T lymphocytes contribute to immune control, and differences in antigen-specific T-cell responses may help distinguish LTBI from ATB.[2]

Recent advances in artificial intelligence (AI) and machine learning (ML) have created new opportunities for identifying biomarkers that differentiate LTBI from active disease. ML algorithms can analyze large, complex transcriptomic and proteomic datasets to detect patterns associated with disease activity, supporting more accurate diagnosis and personalized prevention strategies.[3]

Successful ML-based biomarker discovery requires rigorous data preprocessing, including quality control, normalization, batch-effect correction, and robust feature selection. Commonly used algorithms include logistic regression, support vector machines, ensemble methods, and neural networks. Model performance is evaluated through cross-validation, external validation cohorts, and metrics such as AUC-ROC, sensitivity, and specificity.[4]

Interpretability is enhanced using explainable AI tools such as SHAP and LIME, while reproducibility depends on transparent reporting, public code availability, and adherence to guidelines such as TRIPOD and MIAME.

Transcriptomics

Transcriptomics is one of the most promising approaches for assessing TB infection activity because it captures host gene-expression responses to M. tuberculosis. Whole-transcriptome profiling is mainly performed using microarrays and RNA sequencing (RNA-seq). Combined with ML methods, transcriptomic data can improve differentiation between LTBI and ATB and may help identify individuals at high risk of disease progression.[5]

Recent studies have identified gene-expression signatures with strong diagnostic performance, including metabolic and immune-related gene panels capable of distinguishing LTBI from ATB with AUC values ranging from approximately 0.80 to 0.87. These findings highlight the potential of ML-driven transcriptomic biomarkers for improving TB diagnosis and supporting precision medicine approaches.

Table 1: Key transcriptomic signatures for the diagnosis of LTBI and active TB

Study	Data Type	Signature Composition	Performance	Application
Suliman, S. et al., 2016,[24]	RNA-seq, microarray	GBP1, IFITM3, P2RY14, ID3	AUC 0.82–0.89; sensitivity 73–85%; specificity 76–78%	LTBI/ATB differentiation; PoC potential
Vargas, R. et al., 2023,[25]	Whole-blood transcriptomics	GBP2, FCGR1B, SERPIN C1 inhibitor, TUBGCP6, TRMT2A, SDR39U1	AUC 0.93; sensitivity 90.9%; specificity 88.5%	ATB/LTBI differentiation; treatment monitoring
Gong et al., 2021,[12]	datasets + qRT-PCR	SERPING1, VAMP5	Sensitivity ~88%; specificity ~78%; sensitivity 100%(BATF2 + VAMP5)	ATB diagnosis; LTBI vs. ATB; therapy monitoring
Kwan, P.K.W. et al., 2020,[26]	Targeted transcript levels + ML	FCGR1B, GBP1, GBP5	Spec. ~72.7%	LTBI identification vs. uninfected
Qingqing, S. et al., 2024,[27]	GEO whole blood + ML	SLC26A8, ANKRD22, FCGR1B	AUC 0.801	LTBI/ATB discrimination(multi-cohort)

Critical Appraisal of Transcriptomic Approaches in LTBI and ATB Differentiation

Transcriptomic profiling has improved understanding of host immune responses to Mycobacterium tuberculosis and shows promise for differentiating latent tuberculosis infection (LTBI) from active tuberculosis (ATB). However, several challenges limit its clinical translation.

Many studies use small discovery cohorts, sometimes including fewer than 10 participants per group, reducing statistical power and increasing the risk of overfitting. Most investigations have also been conducted in single geographic regions, particularly China, limiting the generalisability of findings to diverse populations. Variability in sample types (whole blood vs. PBMCs), stimulation methods, and analytical platforms further contributes to inconsistent results and hinders standardisation.

Most transcriptomic studies are cross-sectional, providing limited insight into disease progression and the risk of LTBI developing into active disease. Additionally, many promising gene signatures have not been extensively validated in independent cohorts, raising concerns about reproducibility. Differences in sequencing platforms, data processing methods, and batch effects can also influence signature performance.

Practical implementation remains challenging because many transcriptomic assays require large gene panels, specialised equipment, and significant technical expertise, making them costly for resource-limited, high TB-burden settings. Despite these limitations, advances in artificial intelligence, single-cell transcriptomics, and multi-cohort validation studies have led to simpler gene signatures, such as FCGR1B, SLC26A8, and

ANKRD22, with improved diagnostic accuracy and potential for point-of-care applications.

Microarray-Based Approaches

Machine learning applied to microarray datasets has identified several biomarkers for LTBI–ATB differentiation. One study developed a naïve Bayes classifier based on spermatogenesis regulator, DEAH-box protein 29, and PTPRC, achieving sensitivity and accuracy above 97%, although external validation was lacking. Another study identified CXCL10, ATP10A, and TLR6 and used a decision-tree model that achieved 71% sensitivity and 89% specificity in an independent validation cohort. However, the small discovery sample size (four participants per group) highlights the need for larger and more diverse cohorts to ensure reliable and generalisable results.

RNA Sequencing Approaches

RNA sequencing (RNA-seq) provides greater sensitivity and transcript coverage than microarrays and has enabled the identification of novel LTBI and ATB biomarkers. Wang et al. identified a three-gene signature (TNFRSF10C, EBF3, and A2ML1) that achieved 91.5% classification accuracy, with 86.2% sensitivity and 94.9% specificity. Validation in individuals with suspected ATB showed 82.4% sensitivity and 92.4% specificity. While these findings demonstrate the potential of ML-assisted RNA-seq diagnostics, all studies were conducted in China, and larger multi-centre studies are needed to confirm their applicability across diverse populations.

Overall, transcriptomic and ML-based approaches show considerable promise for improving LTBI diagnosis, but broader validation, standardisation, and cost reduction are essential before routine clinical implementation. [Figure 1]

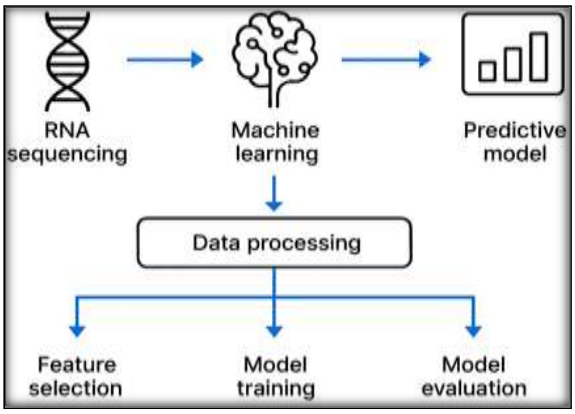

Figure 1: ML-driven RNA-sequencing process.

Real-Time Polymerase Chain Reaction (RT-PCR)

Developing LTBI diagnostic models using microarray or RNA-seq technologies can be challenging due to their cost, complexity, and limited feasibility in resource-constrained settings. Recent studies, however, demonstrate that real-time PCR (RT-PCR) combined with ML approaches offers an affordable and practical alternative for improving LTBI diagnostics. [Figure 2]

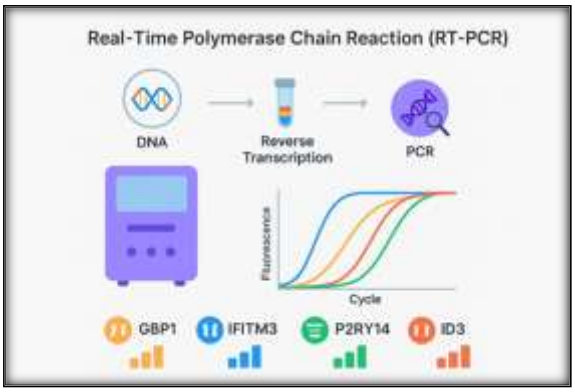

Figure 2: PCR (RT-PCR) combined with ML approaches.

A landmark study from Bangalore, India, combined RT-PCR with machine-learning analysis of gene-expression data to identify a four-gene diagnostic signature (GBP1, IFITM3, P2RY14, and ID3). The model demonstrated strong external validation, achieving an AUC of 0.89 (85% sensitivity, 76% specificity) in The Gambia and 0.82 (73% sensitivity, 78% specificity) in Uganda. Its large and diverse cohort, including ATB, LTBI, and healthy controls, strengthened biomarker validation and highlighted the potential of RT-PCR–based ML diagnostics across different geographical settings.

Importantly, high diagnostic accuracy has also been achieved without machine learning. A six-gene whole-blood signature (GBP2, FCGR1B, SERPINC1, TUBB6, TRMT2A, and SDR39U1) distinguished ATB from LTBI with an AUC of 0.93, 90.9% sensitivity, and 88.5% specificity. Another transcriptomic study identified four biomarkers (UBE2L6, BATF2, SERPING1, and VAMP5) that achieved 88% sensitivity and 78% specificity, outperforming the T-SPOT assay. Together, these findings demonstrate the strong potential of transcriptomic signatures, with or without ML integration, for improving LTBI and TB diagnosis.

Table 2: Diagnostic potential of scRNA-seq for LTBI

scRNA-seq Task	Description	Diagnostic Relevance
Identification of LTBI-related genes	Differential expression between LTBI, ATB and controls	Molecular signature discovery
Characterisation of immune subpopulations	Profiling of T cells, B cells, macrophages, (DCs)	Detection of LTBI-specific cellular clusters
Assessment of immune activation	IFN-I/II, inflammatory, and metabolic pathways	Identification of high-risk LTBI phenotypes
Predictive modelling	Machine-learning analysis of expression matrices	Early stratification of individuals at risk of progression

External validation is typically performed using independent cohorts to assess model generalisability, while translational validation often employs qRT-PCR with primer-efficiency testing and ΔCt normalisation to ensure reproducibility. Single-cell RNA sequencing (scRNA-seq) provides high-resolution profiling of individual cells, enabling identification of immune-cell subsets, transcriptional states, and molecular signatures associated with latent tuberculosis infection (LTBI). By comparing LTBI, active TB (ATB), and uninfected individuals, scRNA-seq can reveal LTBI-specific phenotypes and support machine-learning models for early risk stratification.

Proteomic Approaches for LTBI vs.

Proteomics combined with machine learning has emerged as a promising strategy for distinguishing LTBI from ATB. Research has focused on MTB-specific proteins, host antibodies, and serum/plasma protein signatures. Although current tests such as the tuberculin skin test and interferon-gamma release assays (IGRAs) improve TB detection, they cannot reliably differentiate LTBI from ATB. High-throughput proteomic studies have identified antigen and inflammatory protein signatures with strong diagnostic potential. Future work should prioritise external validation, assay standardisation, and cost-effective translation into clinical practice, with particular emphasis on LTBI-specific antigens and biomarkers. [Table 3]

Table 3: Proteomic biomarkers and machine-learning analysis

Biomarker Category	Technology	Advantages	Limitations	ML Application
Cytokines/Chemokines	Luminex, ELISA	High biological relevance	Wide dynamic range; variability	Boosting, random forest for multiplex panels
Acute-Phase Proteins	Mass Spectrometry	High information content; relevance	High cost; pre-analytical complexity	Feature selection, clustering
Complement and Innate Immunity Markers	Targeted Proteomics	Pathogenesis	Platform-dependent batch effects	Integration with transcriptomic data

Machine Learning and Biomarkers for LTBI vs. ATB Diagnosis

Proteomic, transcriptomic, and cellular biomarkers combined with machine-learning (ML) approaches show strong potential for distinguishing latent tuberculosis infection (LTBI) from active tuberculosis (ATB). Several studies have identified antigen-based signatures, cytokine panels, gene-expression profiles, and immune-cell subsets that achieve sensitivities and specificities frequently exceeding 85–95%.

Proteomic studies demonstrated that ML algorithms such as random forests, logistic regression, support vector machines (SVMs), and clustering methods significantly improve diagnostic performance by selecting optimal biomarker combinations. For example, four-antigen and seven-antigen signatures achieved sensitivities and specificities above 90%, while cytokine-based panels including CXCL10 (IP-10), VEGF, ADA2, and MCP-1 also showed excellent discriminatory power.

Transcriptomic analyses identified several promising gene signatures, including CXCL10–ATP10A–TLR6, TNFRSF10C–EBF3–A2ML1, and UBE2L6–BATF2–SERPING1–VAMP5, with sensitivities ranging from 71–100% and specificities from 78–95%. These signatures often outperformed conventional interferon-gamma release assays (IGRAs) and may additionally monitor treatment response. Cellular biomarkers, including monocyte-to-lymphocyte ratio (MLR), neutrophil-to-lymphocyte ratio (NLR), MAIT cells, CD161+ T cells, and MTB-specific activated CD4+ T cells, also demonstrated strong diagnostic utility. Advanced flow-cytometry approaches such as the TB-Flow Assay achieved sensitivity of 93.6% and specificity of 97.1%.

Despite promising results, several limitations remain. Most studies involve small, geographically restricted cohorts, limiting generalisability. Variability in biomarker panels, analytical platforms, and ML methods complicates comparison and standardisation. Multi-omics integration offers improved accuracy but increases computational complexity and risk of overfitting.

Recent advances in deep learning further support LTBI diagnostics. Convolutional neural networks (CNNs) have improved interpretation of T-SPOT.TB assays, while hybrid deep-learning models applied to chest X-rays achieved diagnostic accuracies exceeding 95%. Autoencoders and multimodal neural networks can integrate transcriptomic, proteomic, immunological, and radiological data, potentially enabling more accurate and personalised TB diagnostics.

Overall, ML-driven biomarker signatures consistently outperform single biomarkers and show considerable promise for LTBI and ATB differentiation. However, large multicentre validation studies, standardised protocols, and clinically practical assays are required before routine implementation.

Table 4: Neural-network approaches in LTBI diagnostics

Model / Approach	Data Source	Architecture	Results	Application
for T-SPOT.TB	Immunological imaging	Two-stage CNN + Logistic Regression	Improved LTBI/ATB differentiation	Enhancement of IGRA interpretation
Efficient Net + MLP-	Chest radiography	Hybrid Deep-Learning	Accuracy: 96.3%;	Screening and disease-

Mixer		Model	Sensitivity: 95.9%; Specificity: 96.6%	activity assessment
Auto encoders	Transcriptomic datasets	Neural Representation Learning	Extraction of latent molecular subtypes	Predictive modelling
Multimodal Transformers	Radiology, Omics, and Immunology	Attention-based transformer	Superior integrated diagnostic performance	Multimodal LTBI diagnostic

Predominant Focus on Diagnosis Rather than Prognosis

Most ML studies focus on distinguishing ATB from LTBI at a single time point, with relatively few addressing the prediction of progression from LTBI to active disease. Developing clinically useful prognostic models requires large longitudinal cohorts and the use of time-dependent evaluation metrics, such as time-dependent AUC, to assess predictive performance over time. Future research should prioritise long-term follow-up studies in high-risk populations to enable the development and validation of robust prognostic tools.

Table 5: Limitations of current studies and ML models

Limitation	Manifestation	Consequences	Required Actions
Insufficient validation	Single-centre cohorts	Poor generalisability	Multi-centre cohorts; TRIPOD compliance
Batch effects	Platform and SOP variability	Spurious signals	Data harmonisation
Low interpretability	Complex ML/DL architectures	Limited clinical acceptance	SHAP/LIME; simplified models
Lack of prognostic evidence	Cross-sectional design	No prediction of progression	Longitudinal cohort studies

DISCUSSION

Recent advances in transcriptomics, proteomics, and machine learning (ML) have improved the ability to distinguish latent tuberculosis infection (LTBI) from active tuberculosis (ATB). Multi-omics approaches integrating molecular and immune biomarkers have shown promising diagnostic performance, often surpassing traditional immunological assays.[6]

However, several challenges limit clinical translation. Many studies use small, geographically restricted cohorts, lack external validation, and are affected by technical variability, including batch effects and differences in analytical platforms. Complex ML models also often suffer from limited interpretability and inadequate assessment in real-world settings.[7]

Bhimabhoi Medical College and Hospital, Balangir, Odisha, India, represents an emerging tertiary care and teaching institution in a high tuberculosis burden region of Eastern India. The hospital caters to a largely rural and socioeconomically diverse population, where tuberculosis remains a major public health challenge with significant overlap of other chronic respiratory and inflammatory conditions. Such a setting provides a valuable real-world clinical environment for evaluating novel diagnostic approaches, including multi-omics and machine learning–based tools for distinguishing latent tuberculosis infection (LTBI) from active tuberculosis (ATB). Integration of advanced molecular diagnostics within this context may help bridge the gap between experimental models and practical clinical application in resource-limited settings.[8]

A major limitation is the insufficient inclusion of clinically relevant confounders such as viral respiratory infections, sarcoidosis, non-tuberculous mycobacterial infections, COPD, and other inflammatory diseases, which can share similar immune and transcriptomic signatures with TB. This may lead to false-positive or false-negative classifications and reduce diagnostic specificity.[9]

Future research should focus on large, multi-centre longitudinal cohorts, standardised methodologies, multi-omics integration, and explainable AI approaches. Evaluation should extend beyond accuracy metrics to include clinical utility, predictive values, decision-curve analysis, and cost-effectiveness.[10]

CONCLUSION

Omics-based biomarkers combined with AI and ML offer significant potential for improving LTBI diagnosis and risk stratification. Nevertheless, widespread clinical implementation remains limited by methodological heterogeneity, insufficient validation, and challenges in model interpretability. Progress will require harmonised datasets, rigorous external validation, inclusion of diverse patient populations—including high-burden and resource-limited settings such as Balangir—and development of affordable point-of-care diagnostic platforms. With continued advances in multi-omics and explainable ML, these technologies could substantially improve TB diagnosis and global tuberculosis control.

REFERENCES

1. World Health Organization. *Global Tuberculosis Report 2022*. Geneva: WHO; 2022.
2. World Health Organization. Global Tuberculosis Report 2024. Geneva: WHO; 2024.
3. Starshinova A, et al. Tuberculosis and COVID-19 dually affect human Th17 cell immune response. Biomedicines. 2023;11:2123.

4. Kudryavtsev I, et al. Immune response in tuberculosis: From latent infection to active disease. Front Tuberc. 2024;2:1438406.
5. Migliori GB, et al. Worldwide effects of the COVID-19 pandemic on tuberculosis services. Emerg Infect Dis. 2020;26:2709–2712.
6. Li Z, et al. Serum biomarker panel for distinguishing latent and active pulmonary tuberculosis. Sci Rep. 2021;11:14516.
7. Cohen A, et al. Global prevalence of latent tuberculosis: A systematic review and meta-analysis. Eur Respir J. 2019;54:1900655.
8. Gong W, Wu X. Differential diagnosis of latent and active tuberculosis. Front Microbiol. 2021;12:745592.
9. Suliman S, et al. Four-gene blood signature predicts progression to tuberculosis. Am J Respir Crit Care Med. 2018;197:1198–1208.
10. Luo Y, et al. Machine learning algorithm for distinguishing active and latent tuberculosis infection. BMC Infect Dis. 2022;22:965..