Machine Learning-Driven Omics Approaches for Diagnostic Biomarker in Tuberculosis based on Differential Gene Expression data

Javed Aalam (javedalambioinformatics@gmail.com) 1*, Rafat Parveen (rparveen@jmi.ac.in)1

1Department of Computer Science, Jamia Millia Islamia, New Delhi, India – 110025

INTRODUCTION & AIM

Introduction: Tuberculosis (TB) remains a major global health threat with rising cases. The World Health Organization's 2023 Global Tuberculosis Report reveals a substantial increase in TB cases, with 8.2 million new diagnoses, up from 7.5 million in 2022 [1]. Traditional diagnostic methods like the Tuberculin Skin Test and IGRA have limitations. OMICS technologies— genomics, transcriptomics, and proteomics—offer the potential for identifying specific biomarkers, and enhancing the accuracy of TB diagnostics and biomarker discovery.

Aim: This study explores how OMICS technologies (genomics, transcriptomics, proteomics) combined with machine learning algorithms can improve TB biomarker identification, focusing on promising biomarkers.
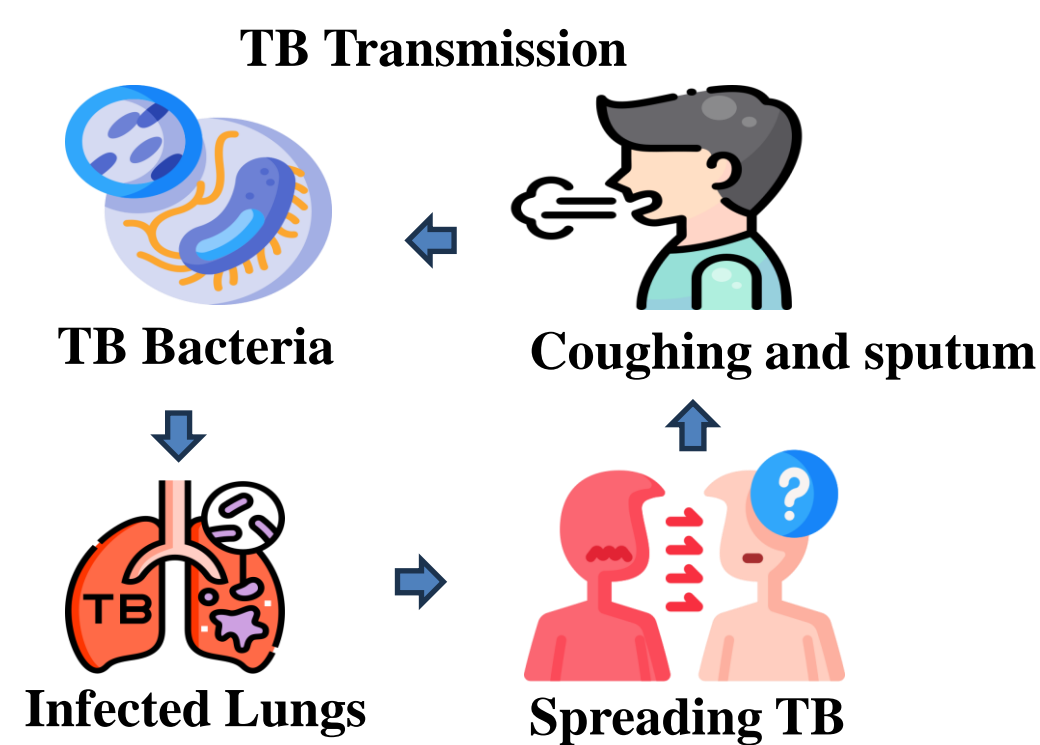

TB Transmission

TB Bacteria

Coughing and sputum

Infected Lungs

Spreading TB

Figure 1. Primary Modes of Tuberculosis Spread Worldwide

METHOD

Approach: A comprehensive analysis of over 100 studies from PubMed, Nature, Scopus, and IEEE Xplore was conducted, focusing on TB, ML, and OMICS technologies.

Focus: Emphasis was placed on machine learning models (KNN, SVM, deep learning) applied to OMICS data for TB biomarker discovery, prioritizing studies on genomic, transcriptomic analysis.

Data: OMICS data, including gene expression profiles, protein biomarkers, and metabolomic data, were integrated for analysis. ML techniques were then used to identify potential TB biomarkers, with a primary focus on optimizing the accuracy, sensitivity, and specificity of the results [2].
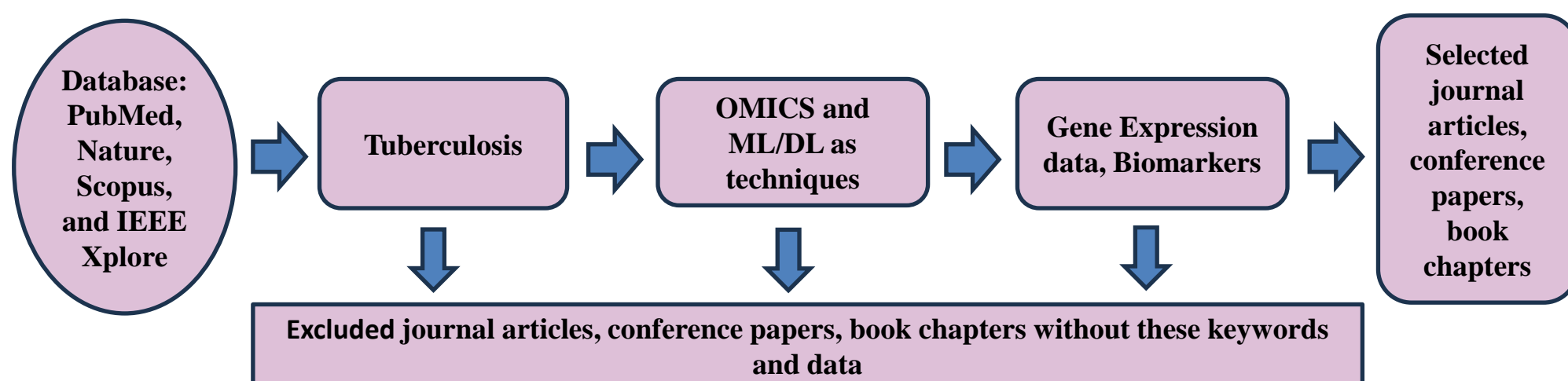

Database: PubMed, Nature, Scopus, and IEEE Xplore → Tuberculosis → OMICS and ML/DL as techniques → Gene Expression data, Biomarkers → Selected journal articles, conference papers, book chapters

Excluded journal articles, conference papers, book chapters without these keywords and data

Figure 2. Flowchart detailing the search criteria for including and excluding research publications

Literature Review and OMICS Data Collection

OMICS Data Preprocessing

Feature Selection and Extraction

Machine Learning Model Selection

Analysis and Final Result
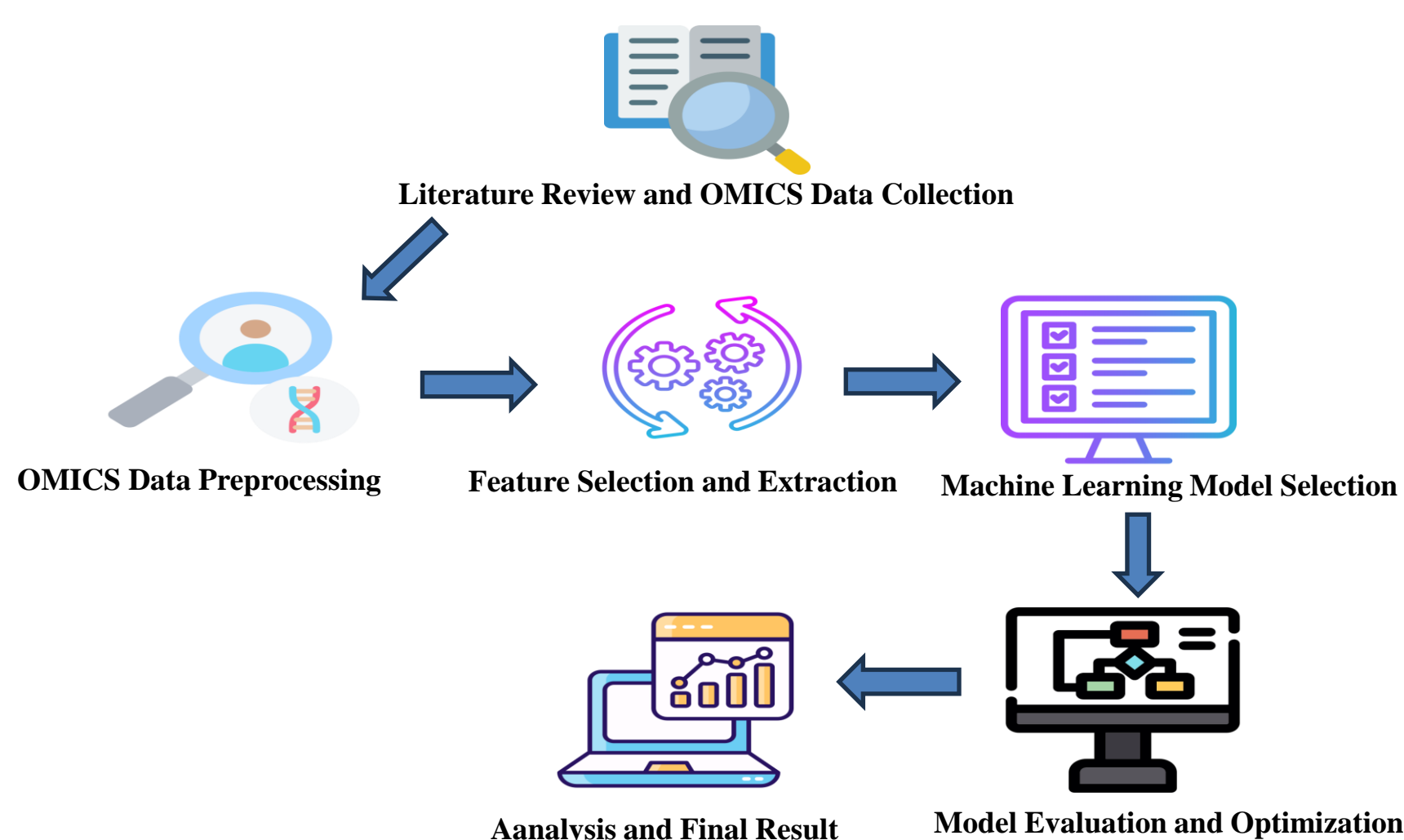
Model Evaluation and Optimization

Figure 3. Optimized workflow for ML-driven OMICS approaches in TB biomarker identification using DEGs Data

RESULTS & DISCUSSION

Authors/Ref.	Aim	Datasets	Algorithms	Findings	Future Work
[1] Juliana Jc Chen et al. (2022)	Screen lncRNAs for TB diagnosis	2,759 subjects, EHR metrics, and GEO database	SVM, Logistic regression, ML	Identified two lncRNAs as TB biomarkers (88% accuracy)	Integrate lncRNAs and EHR for early diagnosis
[2] Shuai Ma et al. (2023)	Identify inflammatory genes for TB progression	GEO database (GSE37250, GSE19439)	LASSO, RF, SVM-RFE	Key inflammatory genes found (90% accuracy)	Identify therapeutic targets and drugs
[3] A. F. Fernande et al. (2022)	Discover molecular biomarkers for TB diagnosis	SwissProt, MBP-GFP databases	Multi-omics with ML	Identified TB biomarkers (92% accuracy)	Develop point-of-care diagnostics
[4] Hooman H. Rashidi et al. (2021)	Predict active TB from serological data	Retrospective dataset (199 TB cases, 333 plasma samples)	MILO platform with ML	23-antigen model with 93% accuracy	Improve clinical interpretation for TB diagnosis
[5] Sang-Mok Lee et al. (2023)	Identify signature genes for Mycobacterium infection	16S rRNA sequencing of cattle gut microbiota	16S rRNA sequencing with ML	Achieved 96% accuracy in disease identification	Apply ML for microbiota-based diagnostics

Machine learning-driven OMICS approaches have identified TB biomarkers using algorithms like SVM, LASSO, and MILO, achieving 88% to 96% accuracy. These studies focused on lncRNAs, inflammatory genes, and microbiota signatures for diagnosis and progression. Future efforts aim to integrate omics data, improve clinical tools, and develop point-of-care diagnostics [3].

CONCLUSION AND FUTURE WORK

Conclusion: In conclusion, recent advancements in machine learning-driven OMICS approaches have significantly enhanced our ability to discover and identify biomarkers for TB. By leveraging algorithms like LASSO, RF, and SVM across various data types, including metabolomics, proteomics, transcriptomics, and microbiota profiles, researchers have achieved 96% high accuracy in TB diagnosis, progression monitoring, and drug resistance identification. Notably, studies have demonstrated the potential for early diagnosis, non-sputum-based diagnostic approaches, and point-of-care tools. These findings underscore the promising role of ML in advancing TB diagnostics and therapy, paving the way for more efficient, personalized healthcare solutions.

Future Work : Future research should focus on integrating multi-omics data with advanced machine learning to improve TB biomarker identification. Key areas include enhancing non-sputum-based diagnostics, developing point-of-care tools, and utilizing real-time electronic health record data and deep learning to boost diagnostic accuracy. These innovations will be vital for translating ML approaches into personalized, effective TB management.

REFERENCES

1. World Health Organization, "Global Tuberculosis Report 2023," WHO, Geneva, 2023. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports.

2. Ma, S., Peng, P., Duan, Z., Fan, Y., & Li, X. (2023). Predicting the Progress of Tuberculosis by Inflammatory Response-Related Genes Based on Multiple Machine Learning Comprehensive Analysis. *Journal of Immunology Research*, 2023, 1–22. https://doi.org/10.1155/2023/7829286.

3. Denkinger, C. M., Pai, M., & Menzies, D. (2014). New diagnostic tools for tuberculosis: How to improve their uptake and impact. *The Lancet Respiratory Medicine*, 2(7), 538-548. https://doi.org/10.1016/S2213-2600(14)70123-1.