INSPIRE HEALTH

Inspire Health Journal
ISSN 3110-9438, Vol. 1, No. 1, 2026
doi.org/10.65718/inspireHealth.2026.2002

InspireQuill.org
Inspire Academic Press (Canada)

Research Paper OPEN ACCESS

Leveraging Non-Clinical Factors and Machine Learning for Improved Tuberculosis Prediction in Resource-Limited Areas

Prince Manzano [1, *], Said Baadel [2] and Faith-Michael Uzoka [3]

[1] Department of Mathematics and Computing, Mount Royal University, Calgary, Canada
[*] Corresponding Auhtor: pmanz282@mtroyal.ca

Abstract

Despite being curable and treatable in the majority of modern nations, tuberculosis (TB) remains a major public health concern and a top cause of infection-related deaths globally. Non-clinical factors like biological, socio-economic, and environmental factors are relevant to prediction and prevention, as relying only on clinical signs does not encompass other risk factors that impact tuberculosis transmission. This study's main goal is to examine the application of machine learning (ML) techniques while emphasizing the impact of non-clinical elements. Data for the study was collected from a variety of public and private healthcare facilities in a few chosen states in the Niger Delta region of Nigeria. The ability of three machine-learning classifiers is evaluated to predict and detect dengue fever for the application in TB. RIPPER was found to be a balanced classifier due to its high F-Measure and ROC-AUC scores. This study focuses on non-clinical factors that affect the spread of TB in addition to the significance of ML with its inclusion of non-clinical factors with its approach to prediction.

Keywords: Tuberculosis Prediction, Causative Data, Machine Learning, Tuberculosis, Resource-Limited Areas, RIPPER.

1. Introduction

With an estimated 10.6 million individuals in 2021 contracting the disease, 10.1 million in 2020, 1.6 million fatalities in 2021, and 1.5 million fatalities in 2020, TB remains a leading cause of infectious death globally. With an incidence rate of tuberculosis which increased by 3.6% in 2021, in comparison to 2020, this suggests a reversal in trend, decreasing nearly 2% yearly in the span of the past two decades [1]. Ranking first in Africa and sixth globally, Nigeria reports for about 4.6% of global burden resulting from TB, with approximately 15 Nigerian fatalities every hour due to TB, which is roughly 125,000 deaths yearly [2]. According to provisional data, over 361,000 TB cases were reported in Nigeria in 2023. Overall, this marked a 26% increase in cases compared with 2022 [3].

Academic Editor:
Jaafar Alghazo

Received: 12/07/2025
Revised: 18/08/2025
Accepted: 21/10/2025
Published: 04/01/2026

Citation
Manzano, P., Baadel, S. & Uzoka, F. M. (2026). Leveraging Non-Clinical Factors and Machine Learning for Improved Tuberculosis Prediction in Resource-Limited Areas. *Inspire Health Journal, 1*(1), 13-20.

Although clinical signs and symptoms of TB benefit presumptive diagnosis: cough, hemoptysis, dyspnoea, chest pain, night sweating, anaemia, tachycardia, lung-auscultation finding, fever, low body-mass index, low mid-upper arm circumference, expeditious diagnosis and treatment are advocated to reduce and prevent tuberculosis (TB) transmission, resulting in a reduction of the TB infection reservoir, which is essential for control and eradicate TB worldwide.

Depending solely on clinical symptoms, however, may not be enough, especially in areas with limited resources as TB symptoms frequently overlap with other febrile illnesses. To address this, utilizing non-clinical factors such as biological, environmental, and socioeconomic variables could provide fresh perspectives for enhancing TB prediction models. The prevalence of diseases and their dynamics of transmission are significantly influenced by non-clinical factors [15]. While environmental factors affect vector habitats and contribute to the spread of vector-borne diseases, biological factors can affect an individual's susceptibility to infectious diseases [16, 17]. Additionally, socioeconomic factors have a significant impact, especially in low-resourced communities. Machine learning techniques can capture intricate relationships and interactions that could improve the accuracy of TB predictions by incorporating these various factors into predictive models which can be particularly useful in areas with limited resources where laboratory-based diagnostics are scarce.

Through the contribution of machine learning (ML), healthcare has seen substantial developments and advancements, especially in the diagnostics of febrile diseases, and in identifying subtle patterns human experts may overlook [4, 5]. Although clinical symptoms have been used to diagnose febrile diseases using ML techniques, recent research has also looked at using non-clinical factors to predict infectious diseases [15, 16]. ML techniques have been applied in predicting malaria using environmental factors [6, 17]. Similarly, environmental factors on febrile-disease prediction, including long-lasting insecticide-treated nets, indoor residual spraying, intermittent preventive prophylaxis, malaria prevention strategies, and behavioural change education, significantly impacted malaria prediction [7].

This study aims to assess the predictive power of non-clinical factors for tuberculosis, focusing on biological, environmental, and socioeconomic variables over traditional clinical symptoms. By evaluating various machine learning classifiers, the research will identify the model best suited to capture the complex interactions of non-clinical determinants that influence yellow fever risk. This approach will enhance predictive modelling capabilities, offering more precise tools for early risk assessment and supporting timely intervention strategies for TB management in high-risk regions.

The remainder of the paper is structured as follows: Section 2 has some literature review. Section 3 presents the methodology used in the study. The experimental settings are provided in section 4. This is followed by the results and analysis in section 5. Finally, the conclusions and future works are presented in section 6.

2. Literature Review

Recent advancements in healthcare highlight the potential benefits of Artificial Intelligence (AI). While human decision-making remains essential, human error is unavoidable. A collaborative relationship between AI-powered technologies and human judgment holds significant promises for strengthening healthcare systems and improving patient outcomes.

In the healthcare industry, pathologists and other professionals are notably affected by AI developments. Although current AI technologies have not yet reached full autonomy in diagnostic accuracy, confirmation from a pathologist is still required. Nevertheless, AI offers substantial advantages, such as reducing researcher workload and decreasing the likelihood of missed diagnoses. Designed to manage high workloads and address human error and bias, AI remains a desirable and effective tool [8].

In many developing countries, TB diagnosis continues to rely heavily on traditional methods, including blood tests, cultures, sputum analysis, and biopsies. These procedures often require one to two weeks or more to yield

diagnostic results and are sometimes affected by accuracy limitations. The integration of machine learning algorithms presents a viable solution to overcome these challenges [9].

Environmental factors also play an important role in the transmission of TB. In Sub-Saharan Africa, the mining industry has been strongly linked to higher TB incidence and mortality. Contributing factors include confined spaces, poor ventilation within mining facilities, and constant exposure to silica dust. Additionally, the proximity of mining hostels to high-risk areas, such as those with widespread sex work, increases transmission risks [10]. Poverty-related housing conditions further exacerbate the issue. Inadequate housing often leads to overcrowding and poor sanitation, both of which are recognized as major contributors to TB spread [10].

Biological factors significantly influence the prevalence of TB. TB is more common in the elderly population due to weakened immune systems. Since TB is airborne, it is more easily transmitted among individuals with compromised immunity [12]. Individuals who have a unhealthier lifestyle are also susceptible, specifically Individuals who are exposed to smoking, indoor air pollution, alcohol use, and under-nutrition face a considerable portion of mortality [13].

Socioeconomic factors also have a strong impact on TB transmission. In developed countries, TB is nearly eradicated. However, in many low-income or developing regions, the disease remains widespread. According to one study, regions with a high proportion of elderly individuals, particularly where access to healthcare services and insurance is limited, experience greater TB incidence [12]. In wealthier areas, access to healthcare services contributes to low TB rates, while in financially disadvantaged areas, the inability to afford adequate care results in TB becoming endemic [14].

3. Methodology

3.1 Data Collection

Data for this study were collected from both public and private secondary and tertiary health facilities in selected states within the Niger Delta region of Nigeria. The dataset, derived from patient consultation records, includes 4,868 patient instances. This dataset was used to train and test the classifiers, allowing for an evaluation of their predictive performance in identifying risk factors associated with dengue fever. Table 1 below provides the descriptive statistics of the study participants.

A total of 62 physicians, all experienced in diagnosing febrile illnesses, participated in the study. Among them, 43 were male and 19 were female. The majority (58) were aged 30 years or older, with 32 having more than 10 years of experience diagnosing and treating febrile diseases.

The reliability of the study instrument was measured using Cronbach's Alpha, which yielded a value of 0.740. Since this exceeds the 0.7 threshold [18], the instrument was deemed reliable. The validity of the research tool was ensured through a pilot study, and content validity was confirmed by an independent review conducted by experienced colleagues who were not involved in the study.

3.2 Ethical Considerations

Informed consent was obtained from all participants who opted to take part in the study. They were thoroughly informed about the study's purpose and procedures prior to providing their consent. Participant confidentiality was maintained through data anonymization to protect their identities, and the data was stored securely. Ethical approval for the study was sought and granted by the Mount Royal University Ethics Committee (Human Research Ethics Board # 102232) and the University of Uyo Teaching Hospital health research ethical committee (Ref # UUTH/AD/S/96/VOL.XX1/450).

Table 1. Descriptive Statistics of Study Participants

State	**Frequency**	**%**
Akwa Ibom	1223	25
Cross River	1531	31
Imo	882	18
Rivers	1232	25
Age Group	**4868**	**100**
<19 yrs	1934	40
19-24 yrs	424	9
25-44 yrs	1557	32
45-60 yrs	600	12
>60 yrs	353	7
Gender	**4868**	**100**
Male	2175	45
Female	2693	55

3.3 Experimental Settings

Evaluating a supervised learning model is a crucial step in assessing its performance. For machine learning predictive models, an error table, commonly referred to as a confusion matrix, is typically used. Additionally, we consider related metrics such as false positive (FP), false negative (FN), true positive (TP), and true negative (TN), which provide further insights into model performances. We outline several common criteria for evaluating predictive models, including accuracy, sensitivity, specificity, and the harmonic mean (also known as the F1 score).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$\text{F-1 score} = 2 \times \frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

A tenfold cross-validation testing method was employed across all experiments. In this approach, the dataset is divided into ten subsets, with nine subsets used for training and one subset reserved for testing the model's predictions. This process is repeated ten times, ensuring that each subset is used for testing exactly once. By applying this method, overfitting is minimized, and the classifiers are evaluated more fairly and reliably.

The performances of these models were evaluated and compared using key performance metrics, allowing us to make recommendations based on their predictive capabilities. Figure 1 below provides a summary of the entire process used in the study, highlighting the workflow.

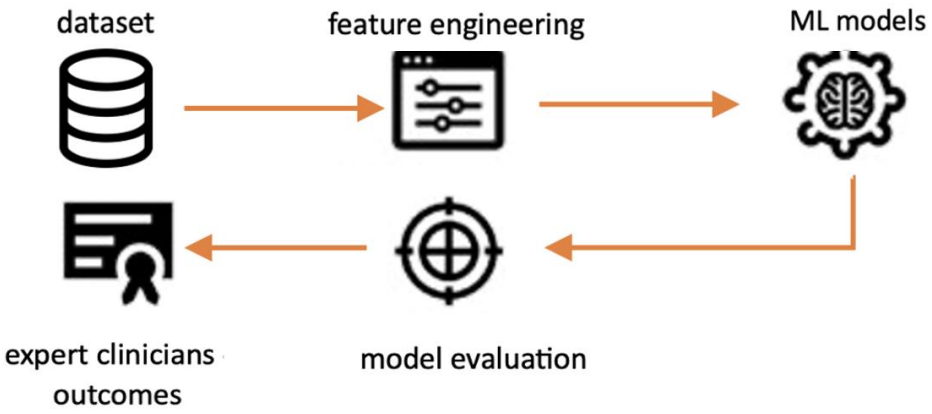

Figure 1. Workflow of ML application on TB detection

The dataset consisted of 19 attributes (features). These features are highlighted in table 2 below.

Table 2. Features in the Dataset

*	**Feature**	**Description**
Biological Factors		
1	GNCN	Genetic Condition
2	HIBP	High blood Pressure
3	HICOL	High cholesterol
4	UNCHRIL	Underlying chronic illness
5	ALG	Allergies
Environmental and Socioeconomic Factors		
6	STRVEN	Street vendor
7	PPHYG	Poor personal hygiene
8	PECON	Poor environmental conditions
9	OVCRW	Overcrowding
10	IVDRUS	Intravenous drug use
11	TRVENRG	Travel to endemic region
12	SKPUPR	Skin puncture procedure
13	DRCOIFPS	Direct contact with infected person
14	LWFLIN	Low fluid intake
15	EXPMQBT	Exposure to mosquito bites
16	SMEXSM	Smoking or exposure to smoking
17	EXIDARPOL	Exposure to indoor air pollution
18	Severity	Severity of TB
19	Class	Medical practitioner confidence level

Table 3 below highlights the overall performance metrics of the 3 classifiers used in this empirical study. We are comparing 3 tree-based classifiers (Decision Table, RIPPER and PART).

Table 3. Performance Metrics of the Models in Predicting High/Low TB Symptoms

Classifier	Accuracy	Recall	Precision	F-Measure	ROC-AUC
Decision Table	0.894	0.632	0.841	0.675	0.717
RIPPER	0.900	0.689	0.816	0.729	0.684
PART	0.888	0.680	0.761	0.709	0.768

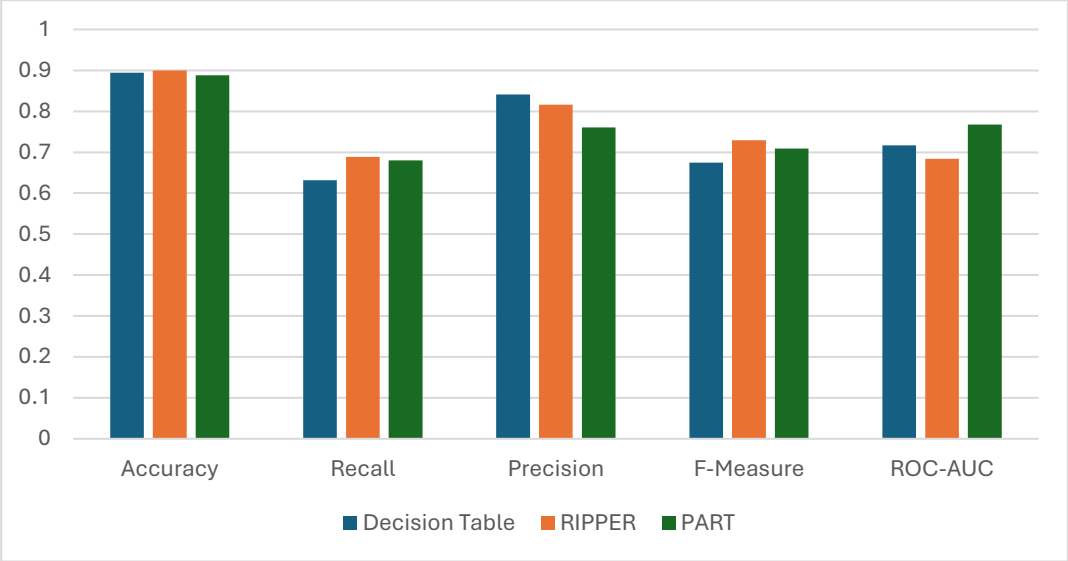

Figure 2. Performance of Classifiers

According to Figure 2, within the recall metric, the classifier RIPPER was found to most accurately identify actual positive cases, as indicated by its value of (0.689). PART and Decision Table followed closely, with values of (0.680) and (0.632), respectively.

Additionally, Decision Table achieved the best results in precision, the metric that determines the proportion of true positive entries among all entries classified as positive. This is an important factor for reducing false positives. With a precision score of (0.841), Decision Table is the best option for correctly identifying medium to high-risk tuberculosis cases. RIPPER follows with a precision of (0.816), performing at an intermediate level in identifying positive TB cases. PART, with a precision of (0.761), generates the most false positives compared to the other classifiers'-measure, which evaluates a model's performance by balancing both precision and recall, shows that Decision Table had the lowest performance (0.675). Meanwhile, PART and RIPPER show moderate improvement, with F-measure scores of (0.709) and (0.729), respectively. In terms of ROC-AUC, which measures a model's ability to distinguish between positive and negative cases, PART performed best with a score of (0.768). Decision Table followed with (0.717), while RIPPER had the lowest score at (0.684).

In summary, if a balanced classifier is the priority, RIPPER, with the highest recall (0.689) and a strong F-measure (0.729), would be the most effective choice. For applications requiring strong class differentiation, PART, with its high ROC-AUC (0.768), solid recall (0.680), and F-measure (0.709), is a strong contender. Finally, Decision Table demonstrates competitive accuracy (0.894) and exceptional precision (0.841), making it the most suitable option in scenarios where minimizing false positives is critical.

4. Conclusion

This paper emphasizes the importance of including non-clinical variables such as biological, environmental, and socio-economic factors in machine learning models for predicting the risk of Tuberculosis (TB). The results section presents an evaluation of three classifiers: Decision Table, RIPPER, and PART. Each classifier demonstrates specific strengths and weaknesses, making them suitable for different use cases depending on the goals of the client or decision-maker. If achieving balanced classification is the primary objective, RIPPER is the most effective model. It provides a strong F-measure, with balanced recall and precision, making it a suitable choice for a wide range of predictive tasks. PART achieves the highest ROC-AUC score (0.768), which reflects its strong ability to differentiate between positive and negative cases. It also delivers solid performance in recall (0.680) and F-measure (0.709), making it a well-rounded option. In contrast, Decision Table shows the highest precision, which is particularly valuable in situations where minimizing false positives is critical. In healthcare settings where resources are limited, timely and accurate risk assessments are essential to enable early diagnosis and intervention. Integrating non-clinical variables into machine learning models can improve diagnostic accuracy and contribute to reducing the spread of TB in underserved areas.

Research Funding

This research was sponsored by the Mount Royal University Internal Research Grant Fund (IRGF) Award no: 104422.

Data Availability Statement
Not applicable.

Conflicts of Interest
The author declares no conflicts of interest.

Ethical Approval and Consent to Participate
Not applicable.

References

[1]. Bagcchi, Sanjeet. (2022). WHO's Global Tuberculosis Report. The Lancet Microbe, Volume 4, Issue 1, e20
[2]. KNVC Nigeria (2024). TB fact in Nigeria. https://kncvnigeria.org/nigeria-is-among-the-14-high-burden-countries-for-tb/
[3]. WHO AFRO (2024) Intensifying new initiatives for TB case-finding in Nigeria https://www.afro.who.int/countries/nigeria/news/intensifying-new-initiatives-tb-case-finding-nigeria
[4]. Asuquo, D., Attai, K., Obot, O., Ekpenyong, M., Akwaowo, C., Kiirya, A., & Uzoka, F. Febrile disease modeling and diagnosis system for optimizing medical decisions in resource-scarce settings. Clinical eHealth, 7, 52–76. 2024. https://doi.org/10.1016/j.ceh.2024.05.001.
[5]. Attai, K., Ekpenyong, M., Amannah, C., Asuquo, D., Ajuga, P., Obot, O., Johnson, E., John, A., Maduka, O., Akwaowo, C., & Uzoka, F.M. (2024). Enhancing the interpretability of malaria and typhoid diagnosis with explainable AI and large language models. Tropical Medicine and Infectious Disease, 9(9), 216. https://doi.org/10.3390/tropicalmed9090216.
[6]. Dukuzumuremyi, A. Machine learning based prediction of malaria outbreak using environment data in Rwanda (Doctoral dissertation, University of Rwanda). 2020.
[7]. Mbunge, E., Millham, R. C., Sibiya, M. N., & Takavarasha, S. (2022, March). Application of machine learning models to predict malaria using malaria cases and environmental risk factors. In 2022 Conference on Information Communications Technology and Society (ICTAS) (pp. 1–5).
[8]. Xiong Y, Ba X, Hou A, Zhang K, Chen L, Li T. "Automatic detection of mycobacterium tuberculosis using artificial intelligence". J Thorac Dis 2018;10(3):1936-1940. doi: 10.21037/jtd.2018.01.91

[9]. S. S. Meraj, R. Yaakob, A. Azman, S. N. M. Rum, and A. S. A. Nazri, "Artificial Intelligence in Diagnosing Tuberculosis: A Review," International Journal on Advanced Science, Engineering and Information Technology, vol. 9, no. 1, pp. 81–91, 2019, doi: 10.18517/ijaseit.9.1.7567.

[10]. D. Stuckler, S. Basu, M. McKee, and M. Lurie, "Mining and Risk of Tuberculosis in Sub-Saharan Africa," American Journal of Public Health, vol. 101, no. 3, pp. 524–530, Mar. 2011. [Online]. Available: https://doi.org/10.2105/AJPH.2009.175646

[11]. S. M. Phahlamohlaka, Environmental Factors Associated with Tuberculosis Incidence and Mortality in Tshwane District, Gauteng Province, South Africa, M.S. thesis, Univ. of South Africa, 2024. [Online]. Available: https://www.proquest.com/docview/3132878162

[12]. C. Im and Y. Kim, "Spatial pattern of tuberculosis (TB) and related socio-environmental factors in South Korea, 2008–2016," PLOS ONE, vol. 16, no. 8, pp. 1–14, Aug. 2021, doi: 10.1371/journal.pone.0255727.

[13]. M. Murray, O. Oxlade, and H.-H. Lin, "Modeling social, environmental and biological determinants of tuberculosis," Int. J. Tuberc. Lung Dis., vol. 15, suppl. 2, pp. S64–S70, Jun. 2011, doi: 10.5588/ijtld.10.0535.

[14]. H. Qureshi, Z. Shah, M. A. Z. Raja, M. Y. Alshahrani, W. A. Khan, and M. Shoaib, "Machine learning investigation of tuberculosis with medicine immunity impact," Diagnostic Microbiology and Infectious Disease, vol. 110, 116472, Aug. 2024. doi: 10.1016/j.diagmicrobio.2024.116472.

[15]. S. Baadel, K. Attai, B. Bassey, E. Attai, A. Majeed, and F.M. Uzoka. Comparative Analysis of Machine Learning Classifiers for Yellow Fever Diagnosis Using Causative Data: Evaluating Naïve Bayes, KNN, Ripper, and PART. 40th Int. Conference on Computers and Their Application (CATA 2025). 2025. San Francisco, USA. Springer.

[16]. S. Baadel, B. Bassey, and F.M. Uzoka. 2025. Enhancing Public Health Outcomes: Machine Learning for Early Dengue Fever Detection in LMICs. In Proceedings of IST – Africa 2025. Nairobi, Kenya. IEEE.

[17]. S. Baadel, C. Akwaowo, J.C. Obi, M. Baadel, et. al. 2025. Beyond Mosquito bites: Analyzing Malaria Risk Factors in Southern Nigeria. Problems of Infectious and Parasitic Diseases. 53 (1). 28-33.

[18]. Nunnally J. C. 1978. Psychometric theory (2nd ed.). New York, NY: McGraw-Hill.