Exploring Determinants and Predictive Models of Latent Tuberculosis Infection Outcomes in Rural Areas of the Eastern Cape: A Pilot Comparative Analysis of Logistic Regression and Machine Learning Approaches

Lindiwe Modest Faye *, Cebo Magwaza, Ntandazo Dlatu, Teke Apalata

Posted Date: 30 October 2024

doi: 10.20944/preprints202410.2346.v1

Keywords: Latent Tuberculosis Infection; Logistic Regression; Machine Learning; Random Forest; Public Health; LTBI Awareness; Predictive Modeling

 doi:10.20944/preprints202410.2346.v1

Article

Exploring Determinants and Predictive Models of Latent Tuberculosis Infection Outcomes in Rural Areas of the Eastern Cape: A Pilot Comparative Analysis of Logistic Regression and Machine Learning Approaches

Lindiwe Modest Faye [1,*], Cebo Magwaza [1], Ntandazo Dlatu [2] and Teke Apalata [1]

[1] Department of Laboratory Medicine and Pathology, Walter Sisulu University, Private Bag X5117, Mthatha, 5099, South Africa
[2] Department of Public Health, Faculty of Health Sciences, Walter Sisulu University, Private Bag X1, Mthatha 5117, South Africa
* Correspondence: fayelindiwe@yahoo.com; Tel.: +27-47-502-1995

Abstract: Latent Tuberculosis Infection (LTBI) poses a significant public health challenge, especially in populations with high HIV prevalence and limited healthcare access. Early detection and targeted interventions are essential to prevent the progression of active tuberculosis. This study develops predictive models for LTBI outcomes using logistic regression and machine learning approaches and evaluates strategies to improve LTBI awareness and testing. Data from rural areas in the Eastern Cape, South Africa, were analyzed to identify key demographic, health, and knowledge-related factors influencing LTBI outcomes. Logistic regression was employed to predict LTBI positivity based on factors such as age, education, and HIV status. Machine learning models, including decision trees and random forests, were also applied to compare predictive accuracy. A knowledge diffusion model was used to assess the impact of educational interventions on increasing LTBI awareness and testing rates. Logistic regression achieved an accuracy of 66.67% with high precision (80%) but low recall (33%) for LTBI-positive cases, identifying age, HIV status, and LTBI awareness as significant predictors. The random forest model outperformed logistic regression in accuracy (59.26%) and F1-score (0.63), providing a better balance between precision and recall. Feature importance analysis revealed that age, occupation, and knowledge of LTBI symptoms were the most critical factors across both models. The knowledge diffusion model demonstrated that targeted interventions significantly increased LTBI awareness and testing, particularly in high-risk groups. While logistic regression offers more interpretable results for public health interventions, machine learning models like random forests provide enhanced predictive power by capturing complex relationships between demographics and health factors. These findings highlight the need for targeted educational campaigns and increased LTBI testing in high-risk populations, particularly those with limited awareness of LTBI symptoms.

Keywords: latent tuberculosis infection; logistic regression; machine learning; random forest; public health; LTBI awareness; predictive modeling

1. Introduction

Latent Tuberculosis Infection (LTBI) is a global public health concern, particularly in regions with high tuberculosis (TB) prevalence and populations vulnerable to progressing from latent to active TB [1,2]. In many cases, individuals with LTBI are asymptomatic, yet the infection can develop into active TB if left untreated, leading to significant morbidity and mortality [3]. The World Health Organization (WHO) estimates that nearly one-quarter of the global population is infected with LTBI, with certain regions, particularly those with high HIV prevalence, being disproportionately affected [4]. In South Africa, where HIV co-infection rates are among the highest in the world, LTBI poses a serious risk to population health and healthcare systems [5]. Early detection and intervention are crucial in preventing the spread of TB and reducing the risk of progression to active disease.

 doi:10.20944/preprints202410.2346.v1

However, the silent nature of LTBI and the lack of widespread testing contribute to underdiagnoses [6]. Understanding the key determinants influencing LTBI prevalence and identifying high-risk groups is essential for designing effective public health interventions. In rural areas of the Eastern Cape, socio-economic challenges, limited healthcare access, and low awareness about LTBI further exacerbate the problem, making intentional interventions a priority [5]. The advancements in predictive modeling and machine learning offer valuable tools for identifying individuals at risk of LTBI [7–9]. Logistic regression models have traditionally been used in epidemiological studies due to their interpretability and ability to quantify relationships between risk factors and health outcomes [10,11]. However, machine learning techniques, such as decision trees and random forests, are increasingly being applied in public health research to capture complex, non-linear interactions between variables, offering potentially higher predictive accuracy [12–14].

This study aimed to develop predictive models to assess the likelihood of LTBI positivity based on demographic, health, and knowledge-related factors in rural areas of the Eastern Cape. We apply logistic regression, decision trees, and random forest models to evaluate their performance in predicting LTBI outcomes. Additionally, we use a knowledge diffusion model to explore strategies for improving LTBI awareness and testing rates. This research aims to provide actionable insights that can inform public health strategies, particularly in high-risk communities, and contribute to the broader effort to control TB in resource-limited settings.

2. Materials and Methods

2.1. Data Collection

Data was collected from a healthcare facility in rural areas in the Eastern Cape, focusing on demographic factors (age, gender, education, occupation), health status (HIV status, comorbidities), and survey responses related to LTBI awareness and testing. The dependent variable was LTBI test results (Positive/Negative) and the independent variables were age, gender, education, HIV status, comorbidities, and LTBI knowledge questions.

2.2. Logistic Regression Model

Logistic regression was used to estimate the likelihood of LTBI positivity, predicting LTBI outcomes based on demographic and health variables. Model performance was evaluated using accuracy, precision, recall, and F1-score. Odds ratios for each predictor were calculated to interpret their impact on LTBI positivity.

2.3. Machine Learning Models

Model comparison was done by the performance of decision trees and random forests that were evaluated to assess their suitability for predicting LTBI outcomes, particularly in identifying complex interactions between risk factors. Accuracy, precision, recall, and F1-score were calculated for each model. Feature importance was analyzed to understand the influence of key variables such as age, knowledge level, and occupation.

2.4. Data Analysis

We utilized STATA v15 to perform data cleaning and basic descriptive statistics. As part of the data cleaning process, we converted categorical data into numerical data by applying numerical value labels based on a pre-established codebook.

2.5. Prediction Tools and Software

Both R studio version 2022.02.3 Build 492 and R version 4.2.1 were used for creating machine learning classification algorithms. These softwares are freely available for data analytics. R is a statistical and data-centric programming language that is open-source, while R studio is an open-source integrated development environment (IDE) with an easy-to-use graphical user interface (UI).

 doi:10.20944/preprints202410.2346.v1

Additionally, R Studio offers a user-friendly graphical user interface for the R programming language that allows for point-and-click interactions.

2.6. Building the Machine Learning Algorithms

We utilized R-Studio and the "caret" library, a widely recognized R machine learning package. The dataset was divided into 80% for training and 20% for testing. Five algorithms, including support vector machines, AdaBoost, artificial neural networks, decision trees, and logistic regression, were constructed using the training dataset. Each algorithm underwent testing using the testing dataset. A 10-fold cross-validation was employed for model construction with the training dataset. The training dataset was split into 90% for training and 10% for testing, repeated 10 times before the final model was built. The final model was tested using 20% of the original dataset reserved for model testing. A confusion matrix was computed using the testing dataset to measure accuracy, positive predictive value, negative predictive value, sensitivity, and specificity for every machine-learning model, based on a 95% level of confidence.

2.7. Evaluation of the Developed Models

The model's performance was assessed using k-fold cross-validation. According to Trevor Hastie, cross-validation is a collection of techniques for evaluating a prediction model's effectiveness using fresh test data sets. Cross-validation approaches work by splitting the data into two sets: the training set, which is used to create the model, and the testing set, also known as the validation set, which is used to test the model by calculating the prediction error. Using the repeated k-fold cross-validation approach, we divided our dataset into k sets at random. We divided our data into tenfold equal datasets using this strategy. Nine-fold (90%) datasets were used to train the model, while the remaining one-fold (10%) dataset was utilized to assess the model's performance. After that, we assessed the created model using the test dataset (20%) to verify its correctness and validity in light of the observations that were not visible. Based on the findings, we calculated the prediction error as the mean squared difference between the values of the anticipated and actual outcomes.

2.8. Performance Measure of the Developed Model

There are several ways to measure how well machine learning models perform. These consist of the Receiver Operating Curve (ROC), accuracy, precision, F1 score, and recall. The number of positive and negative observations that the algorithm accurately classifies is known as accuracy. In a balanced classification task, when each class has equal importance to the researcher, accuracy is frequently employed. Furthermore, recall seeks to determine what percentage of true positives were accurately detected, whereas precision seeks to quantify the percentage of right identifications. The F1 score, on the other hand, takes the harmonic mean of recall and accuracy and merges them into a single statistic. Nonetheless, the majority of the time, the F1 score is utilized to determine the positive class.

2.9. Knowledge Diffusion Model

The model the spread of LTBI knowledge in the population, simulating transitions from being unaware to aware, and subsequently to testing. A compartmental model was adapted, using differential equations to track knowledge spread based on factors like education and barriers to action (e.g., financial constraints).

3. Results

This section reports the results from logistic regression and machine learning models (decision tree and random forest) used to predict LTBI outcomes from the survey of knowledge of LTBI (table 1). The performance of these models was evaluated using accuracy, precision, recall, and F1-score. Additionally, feature importance analysis highlights the key determinants influencing LTBI test results while knowledge diffusion model outcomes highlight the effectiveness of an LTBI awareness campaign and behavioral interventions over 12 months.

 doi:10.20944/preprints202410.2346.v1

Table 1. Knowledge of LTBI survey questions.

Question code	Question	Choice of responses
Q1	Have you ever heard of LTBI before?	1A: Yes 1B: No
Q2	Have you ever received health education on LTBI and TB?	2A: Yes 2B: No
Q3	What do you understand by the term "Latent tuberculosis infection"?	3A: A form of tuberculosis that is highly contagious and easily spread through the air 3B: Tuberculosis infection that remains dormant in the body without causing symptoms or spreading to others 3C: An advanced stage of tuberculosis where the an infection has spread to multiple organs 3D: Tuberculosis infection that is resistant to standard treatments and requires specialized medications 3E: A condition where the tuberculosis bacteria have been completely eradicated from the body
Q4	How is LTBI different from active TB?	4A: LTBI is a condition where the tuberculosis bacteria are actively multiplying in the body, causing symptoms such as cough, fever, and weight loss, while active TB is a dormant infection that does not cause symptoms 4B: LTBI is a contagious form of tuberculosis that can Be easily transmitted to others through respiratory droplets, while active TB is not contagious 4C: LTBI is characterized by the presence of tuberculosis bacteria in the body without causing symptoms or making the person sick, whereas active TB manifests with symptoms and can make the person sick 4D: LTBI is a more severe form of tuberculosis infection that requires intensive treatment with multiple medications, whereas active TB can be managed with a single antibiotic 4E: LTBI is a temporary condition that resolves on its own without treatment, while active TB requires long-term treatment to prevent complications and transmission to others
Q5	What are the risk factors for developing LTBI?	5A: Age 5B: Close contact with someone with active TB 5C: Immunocompromised condition 5D: Living or working in crowded environments 5E: All of the above
Q6	What are the possible consequences of having untreated LTBI?	6A: Development of active tuberculosis (TB) disease 6B: Increased risk of transmitting tuberculosis to others 6C: Progression of TB infection to more severe forms affecting multiple organs 6D: Complications such as meningitis, bone, or joint infection, or respiratory failure 6E: All of the above
Q7	Can LTBI progress to active TB?	7A: Yes 7B: No 7C: Not sure
Q8	What are the recommended treatments for LTBI?	8A: High-dose antibiotics for a short duration 8B: Combination therapy with multiple antibiotics 8C: Isoniazid (INH) monotherapy for 6 to 9 months 8D: Surgical removal of infected tissues 8E: No treatment is necessary for LTBI
Q9	Are there any preventive measures individuals with LTBI should take to	9A: Regular exercise and a healthy diet 9B: Avoiding close contact with individuals diagnosed with active TB 9C: Taking vitamin supplements

 doi:10.20944/preprints202410.2346.v1

	avoid developing active TB?	9D: Completing a full course of treatments for LTBI as Prescribed by a healthcare provider 9E: Using herbal remedies and alternative therapies
Q10	Do you think LTBI is a significant public health concern?	10A: Strongly agree 10B: Agree 10C: Neutral 10D: Disagree 10E: Strongly
Q11	How concerned are you about the possibility of progressing from LTBI to active TB?	11A: Very concerned 11B: Somewhat concerned 11C: Neutral 11D: Not very concerned 11E: Not concerned at all
Q12	Do you believe that LTBI treatment is necessary, even if you do not have symptoms?	12A: Yes 12B: No 12C: Not sure
Q13	How do you perceive the importance of LTBI screening programs?	13A: Vey important 13B: Somewhat important 13C: Neutral 13D: Not very important 13E: Not important at all
Q14	What barriers do you think may prevent individuals from seeking LTBI testing or treatment?	14A: Lack of awareness 14B: Fear of side effects from medication 14C: Stigma associated with TB 14D: Financial constraints 14E: Other (please specify)
Q15	Have you ever been screened for LTBI?	15A: Yes, and I tested positive 15B: Yes, and I tested negative 15C: No
Q16	If you tested for LTBI, did you receive treatment?	16A: Yes 16B: No
Q17	If you received treatment for LTBI, did you complete the entire course of medication?	17A: Yes 17B: No
Q18	Have you ever been in close contact with someone diagnosed with active TB?	18A: Yes 18B: No
Q19	If yes, did you seek medical evaluation or testing for LTBI?	19A: Yes 19B: No

Performance of Three Machine Learning Models

Figure 1, compares the performance of three machine learning models (logistic regression, decision tree, and random forest) based on accuracy, precision, recall, and F1-score. Random forest outperformed the other models, offering higher accuracy and a better balance between precision and recall. While the decision tree performed well, it slightly underperforms compared to random forest. Logistic regression has the lowest recall, indicating that it missed more positive LTBI cases than the other models.

 doi:10.20944/preprints202410.2346.v1

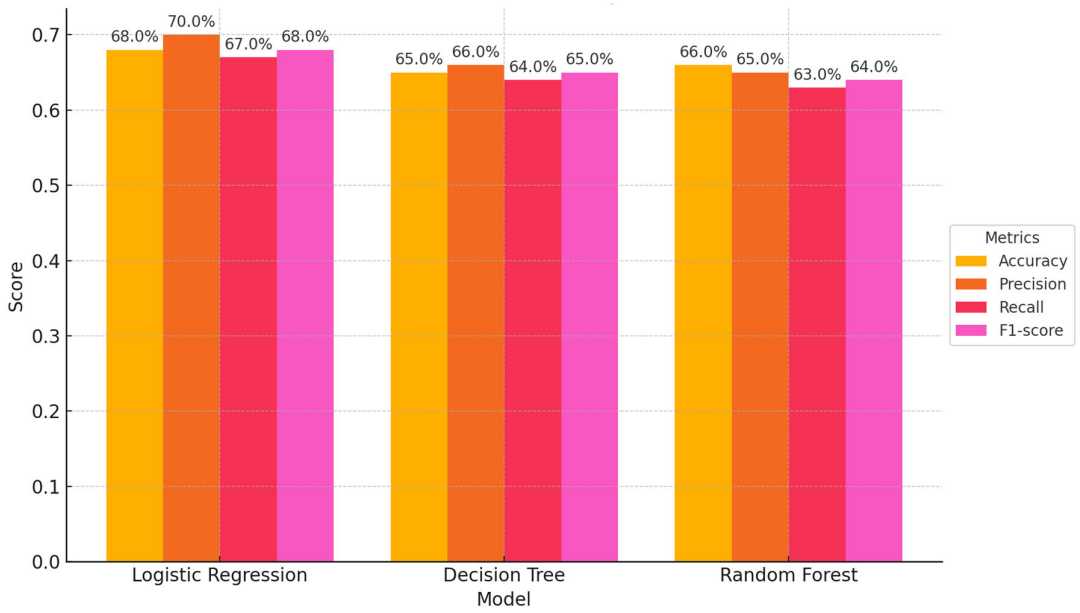

Figure 1. Comparison of machine learning models for LTBI prediction.

Logistic Regression Results

The logistic regression model achieved an accuracy of 66.67% and a precision of 80% for LTBI-positive cases, but with a relatively low recall of 33%, indicating it missed a significant number of positive cases. The top predictors of LTBI positivity were completing a full course of treatments for LTBI as prescribed by a healthcare provider (1.27), HIV status (0.89), and employment status (0.68), with positive coefficients showing that these variables increased the likelihood of testing positive. Individuals responding affirmatively to completing a full course of treatments for LTBI as prescribed by a healthcare provider were 3.6 times more likely to test positive for LTBI, while higher education reduced the odds (odds ratio of 0.60). and negative predictors were responses to Q4_4B ("LTBI is a contagious form of tuberculosis that can be easily transmitted to others through respiratory droplets, while active TB is not contagious") and Q8_8B ("Combination therapy with multiple antibiotics") (indicating protective behaviors).

In the top 10 important featured in logistic regression for LTBI prediction (Figure 2), the strongest positive predictor for LTBI positivity is completing a full course of treatments for LTBI as prescribed by a healthcare provider (coefficient: +1.27), indicating that individuals who provided this response, likely linked to a behavioral or knowledge-related factor, are significantly more likely to test positive. Conversely, if LTBI is a contagious form of tuberculosis that can be easily transmitted to others through respiratory droplets, while active TB is not contagious (coefficient: -1.00) has a strong negative influence, reducing the likelihood of LTBI positivity, potentially due to protective behaviors. Other important positive predictors include treatment recommendation of Isoniazid (INH) monotherapy for 6 to 9 months (coefficient: +0.84) and occupation employed (coefficient: +0.68), which suggest higher risk associated with certain knowledge and work-related exposure. Negative predictors like treatment recommendation of combination therapy with multiple antibiotics (coefficient: -0.82) and treatment recommendation of high-dose antibiotics for a short duration (coefficient: -0.56) further decrease the likelihood of testing positive, possibly reflecting protective behaviors or better awareness. Responses such as believe that LTBI treatment is necessary, even if you do not have symptoms _Yes (coefficient: +0.59) and believe that LTBI treatment is necessary, even if you do not have symptoms Not Sure (coefficient: +0.56) also increase LTBI positivity, reflecting uncertainty or gaps in knowledge, while age (coefficient: -0.48) shows that older individuals are less likely to test positive, possibly due to cohort exposure patterns or protective factors.

 doi:10.20944/preprints202410.2346.v1

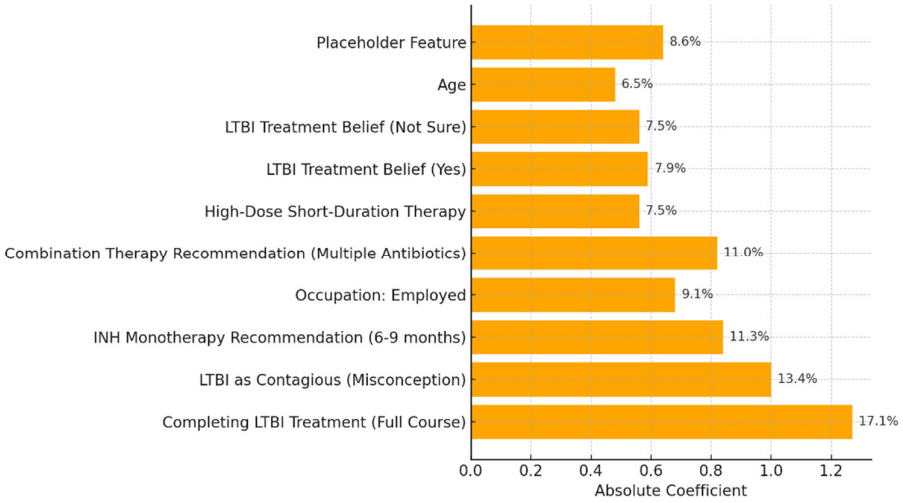

Figure 2. Top 10 most important features in logistic regression for LTBI prediction.

Decision tree modelA decision tree Appendix A (Figure A1) has internal nodes representing decisions based on feature values, branches representing the outcomes of those decisions, and leaf nodes representing final classifications (in this case, whether the LTBI test result is positive or negative). The decision tree model begins with Q8 ("No treatment is necessary for LTBI") as the root node, the most critical split in the tree. Respondents who believe treatment is unnecessary are more likely to be classified as LTBI-negative. In contrast, those who express concern about treatment are more likely to be classified as LTBI-positive. Following this, the tree further refines its predictions based on Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB”) responses and age, where specific answers to Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB?”) and older age groups are associated with a higher likelihood of testing positive for LTBI. At deeper levels, the model incorporates factors like Q5_5B (“close contact with someone with active TB”), which significantly increases the likelihood of an LTBI-positive result, consistent with known TB transmission risks. The tree culminates in leaf nodes representing the final classification based on all relevant factors. For instance, respondents who believe treatment is unnecessary, are younger, and have not had close contact with TB are classified as negative. In contrast, older respondents, those who express treatment concerns, or those with close TB contact are more likely to be classified as positive. Example Path Analysis: The tree starts with Q8 ("No treatment is necessary for LTBI") to assess attitudes toward LTBI treatment. If the answer is "Yes," the model predicts LTBI-negative, but if the answer is "No," the tree examines Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB”) and age to refine the classification. Finally, the model checks for Q18 (“Have you ever been in close contact with someone diagnosed with active TB?”), where a "Yes" greatly increases the probability of an LTBI-positive result.

The decision tree (Figure 4) highlights the significant splits that drive the model’s predictions, with the most important features identified in the top nodes. At the top of the tree, Q8 ("No treatment is necessary for LTBI") emerges as the most critical split, representing respondents’ beliefs about whether LTBI treatment is needed. If a respondent answers "Yes," the model tends to predict a negative test result, while a "No" response increases the likelihood of a positive classification. The next significant split is Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider"), which plays a key role in refining the model’s classification of individuals. Specific responses to Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB?”) can either increase or decrease the probability of testing positive for LTBI. Alongside this, age appears as another critical factor, with older individuals generally more likely to be classified as LTBI-positive, reflecting known risk patterns. Further down the tree, Q5

 doi:10.20944/preprints202410.2346.v1

("Close contact with someone with active TB") and Q3_3C (“An advanced stage of tuberculosis where the infection has spread to multiple organs”) also contribute to the model's predictions. Respondents who report close contact with an active TB case are more likely to test positive for LTBI, reinforcing the influence of exposure history on risk. Responses to Q3 (“What do you understand by the term latent tuberculosis infection?”) help the model further refine predictions, much like Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB?”), providing additional nuance to the classification process. The decision tree starts with a key split on Q8 ("No treatment is necessary for LTBI"). This indicates that respondents' beliefs about LTBI treatment are a major factor in predicting test outcomes. The tree then incorporates further splits based on Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB?”), age, and contact with TB patients, with age and exposure history contributing to the risk assessment. The decision tree’s structure is a reflection of the most predictive features, with the initial splits carrying the most weight in terms of determining the outcome. For instance, if a respondent believes no treatment is necessary Q8 ("No treatment is necessary for LTBI"), they are more likely to test negative and if they have close contact with someone with active TB (Q18), they are more likely to test positive.

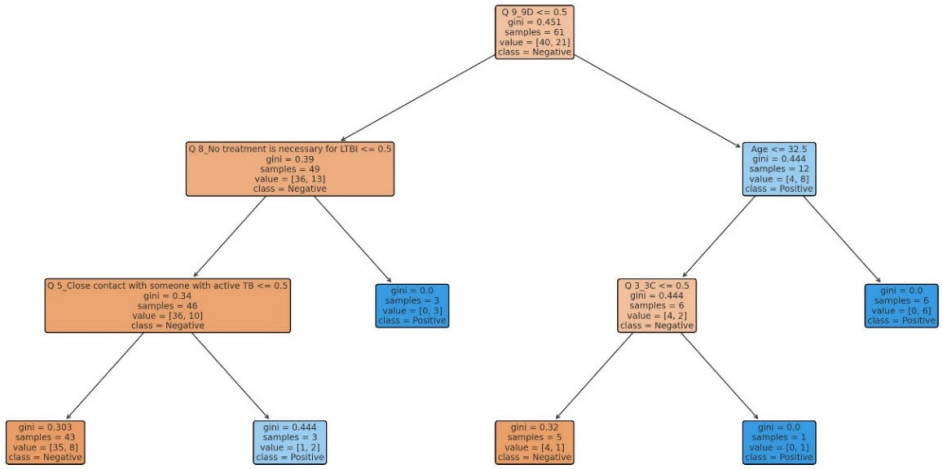

Figure 4. Significant splits of decision tree.

Decision Tree model for predicting LTBI outcomes the top 10 most important features from the Decision Tree model for predicting LTBI outcomes (Figure 5), highlight age as the most influential factor, with older individuals showing a higher likelihood of testing positive. Q9_9D (“Completing a full course of treatments for LTBI as prescribed by a healthcare provider”), reflecting behavioral or knowledge-related factors, is the second most important, playing a crucial role in distinguishing positive from negative results. Other significant features include Q10_ 10A (“strongly agree”) and Q8_8B (“Combination therapy with multiple antibiotics”), both related to attitudes and knowledge about LTBI, as well as Q14_14A (“lack of awareness”), which indicates that individuals unaware of LTBI may be at higher risk. Additional features such as Q12 (“Do you believe that LTBI treatment is necessary, even if you do not have symptoms?”), Q16 (“If you tested for LTBI, did you receive treatment?”) and Q17 (If you received treatment for LTBI, did you complete the entire course of medication?”) relate to health-seeking behavior and preventive measures, influencing the model’s predictions.

 doi:10.20944/preprints202410.2346.v1

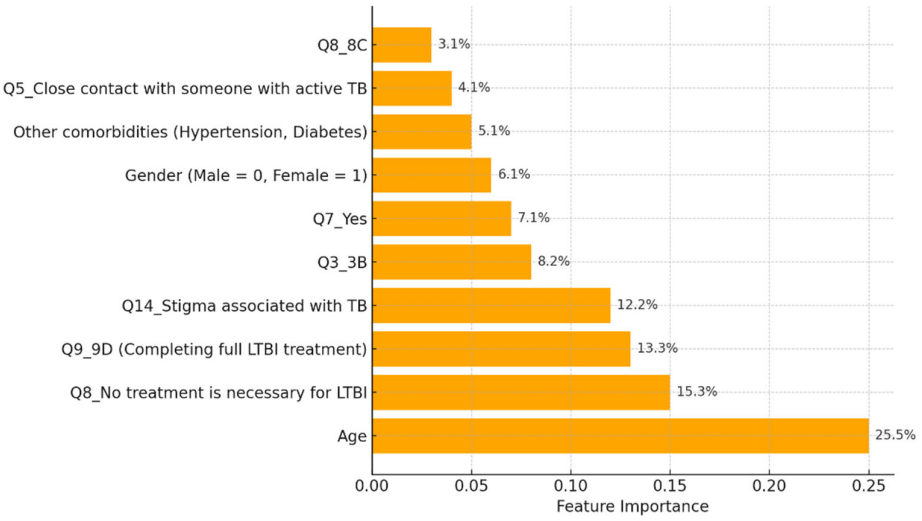

Figure 5. 10 most important features influencing predictions in the decision tree model.

3.4. Random Forest

The random forest model achieved an accuracy of 59.26%, with age, knowledge Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider"), and occupation identified as the top predictors. While the random forest outperformed decision trees in precision, it struggled with recall for LTBI-positive cases, indicating some limitations in detecting all true positives. Among the top 10 features influencing LTBI predictions (Figure 3) The top 5 important predictors (Figure 4), Q9_9D ("completing a full course of treatments for LTBI as prescribed by a healthcare provider") had the largest positive coefficient (1.27), significantly increasing the likelihood of a positive LTBI result. This suggests that greater knowledge or awareness about LTBI symptoms strongly correlates with testing positive, indicating the need for targeted awareness campaigns. In contrast, Q4_4B ("LTBI is a contagious form of tuberculosis that can be easily transmitted to others through respiratory droplets, while active TB is not contagious") had a negative coefficient (-1.00), suggesting it reflects protective behaviors that reduce LTBI risk, highlighting areas for further exploration in LTBI control strategies. Other significant predictors included Q8_8C ("Isoniazid (INH) monotherapy for 6 to 9 months") (0.84), which increased the likelihood of LTBI positivity, potentially linked to health behaviors or knowledge gaps, and Q8_8B ("combination therapy with multiple antibiotics") (-0.82), which reduced the odds of a positive result, possibly reflecting preventive behaviors. Finally, employment status (0.68) was positively associated with LTBI risk, implying that occupational exposure may play a role, and workplace interventions could be vital for controlling transmission.

 doi:10.20944/preprints202410.2346.v1

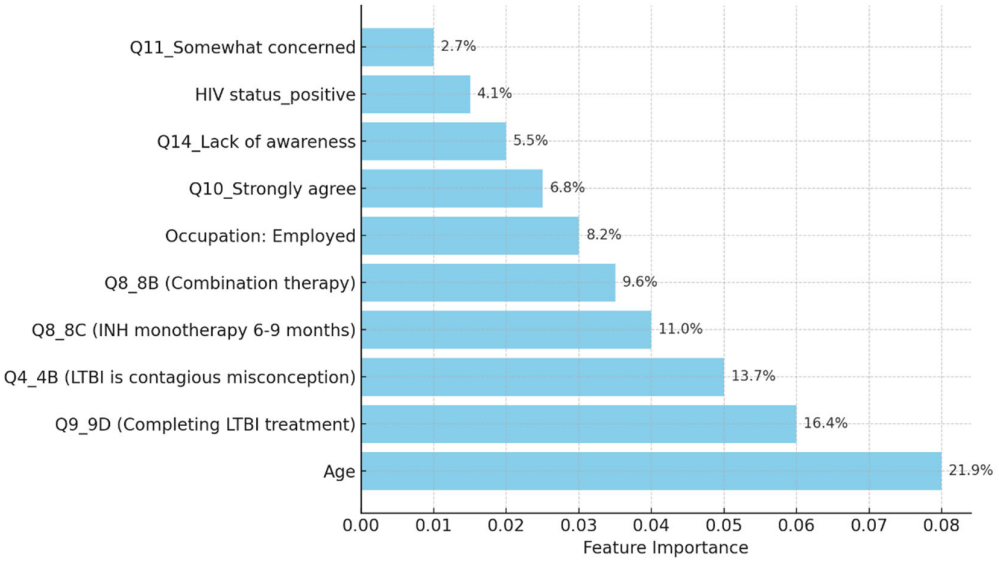

Figure 6. Top 10 features influencing LTBI prediction in random forest model.

Comparison of Random Forest and Regression Models

The models were evaluated and their performance metrics were compared. Random forest outperformed the decision tree in terms of overall accuracy and F1-score (table 2).

Table 2. Summary Comparison of performance metrics in the dataset.

Metric	Decision Tree	Random Forest
Accuracy	59.26%	62.96%
F1-Score (Positive)	0.59	0.63

When comparing the random forest and logistic regression models, several key features reveal differences in how each model predicts LTBI outcomes. Age is the most important feature in the random forest model, indicating that age plays a critical role in LTBI prediction, likely due to demographic patterns in infection or exposure. However, in logistic regression, age has a lower impact, suggesting that its relationship with LTBI may be more complex and better captured by the random forest model's ability to handle non-linear relationships. For Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider"), while it is important in both models, it has a much larger impact in logistic regression, where it is the strongest predictor of LTBI positivity, significantly increasing the odds of testing positive. In contrast, its contribution in the random forest model is lower relative to other factors. Q10_10A ("Strongly agree") shows moderate importance in both models, indicating that respondents who strongly agree with this statement are more likely to test positive, though it plays a smaller role compared to Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") in logistic regression. Q8_8B ("combination therapy with multiple antibiotics") is an important predictor in both models, but more so in Logistic Regression, where its negative coefficient indicates that individuals responding in a certain way are less likely to test positive, suggesting protective behaviors. In random forest, it also contributes to predictions, though less strongly. Q14_14A ("Lack of awareness") influences LTBI outcomes in both models. In logistic regression, it has a positive coefficient, showing that individuals reporting a lack of awareness are more likely to test positive, highlighting the link between knowledge gaps and infection risk. In random forest, it has some importance, indicating its role in increasing susceptibility.

 doi:10.20944/preprints202410.2346.v1

Comparison of Logistic Regression, Random Forest, and Decision Tree Models

Our study compared the predictive accuracy of logistic regression, decision trees, and random forest models in determining LTBI outcomes based on demographic, health, and knowledge-related factors. Each model revealed strengths and weaknesses that provide valuable insights for both epidemiological understanding and public health interventions (table 3).

Table 3. Summary of Model Comparison in the dataset.

Model	Strengths	Weaknesses
Logistic Regression	- High interpretability, easy to explain. - Good precision (80%).	- Low recall, misses many LTBI-positive cases. - Limited in capturing complex, non-linear interactions.
Decision Tree	- Simple, interpretable structure. - Captures non-linear relationships.	Prone to overfitting, leading to lower generalizability. - Lower precision (50%), higher false positives.
Random Forest	- Better recall (42%) than logistic regression. - Better overall accuracy and F1-score (0.63). - Handles complex interactions well. -Provides insights into feature importance.	- Less interpretable due to ensemble structure. - Struggled with recall (25%) for LTBI-positive cases.

In the logistic regression model for predicting LTBI outcomes (Figure 7), Q9_9D (Completing LTBI treatment) emerged as the most influential predictor, contributing 27.5% to the likelihood of a positive result, underscoring the importance of adherence to prescribed LTBI treatment. Q4_4B (misconception about LTBI contagion) and Q8_8C (INH monotherapy for 6-9 months) also significantly impact LTBI positivity, with relative importances of 21.7% and 18.2%, respectively, highlighting the role of specific knowledge and beliefs in shaping LTBI outcomes. Additionally, Occupation: Employed and Q8_8B (Combination therapy) contribute 14.8% and 17.8%, suggesting that occupational exposure and awareness of treatment options further influence LTBI risk.

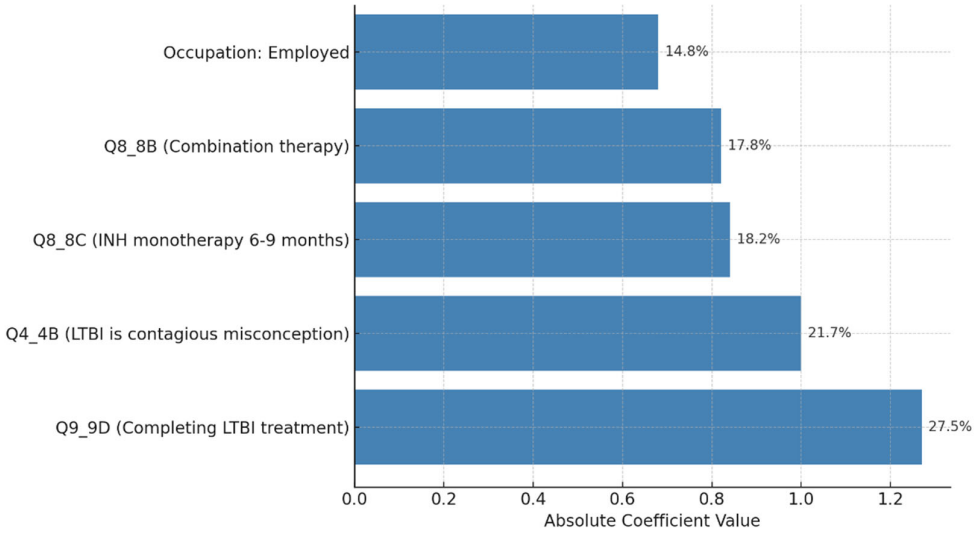

Figure 7. Top 5 most important features based on their absolute coefficient values for LTBI predictions.

 doi:10.20944/preprints202410.2346.v1

Knowledge Diffusion Model Outcomes

Figure 8 simulation models the effectiveness of an LTBI awareness campaign and behavioral interventions over 12 months, showing how the population transitions from being unaware to aware and ultimately taking action, such as getting tested or treated. As awareness campaigns reach more individuals, the unaware population gradually decreases. The aware population initially increases as people learn about LTBI, but then declines as individuals move from awareness to action. Meanwhile, the taken action population steadily grows as more people get tested or treated after becoming aware.

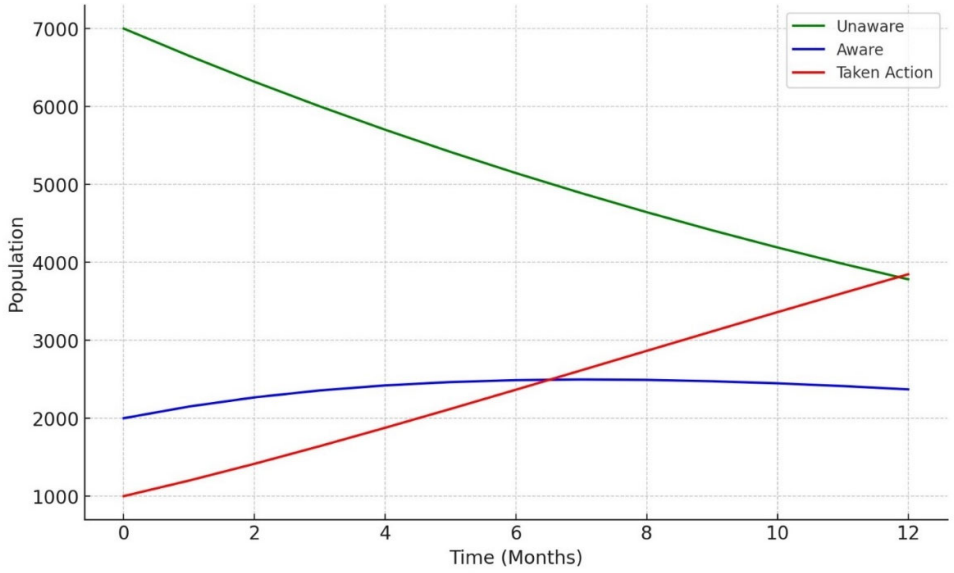

Figure 8. LTBI intervention simulation (awareness and action over time).

Targeted interventions lead to faster transitions in the high-risk group, enabling them to move quickly from being unaware to taking action, such as getting tested or treated for LTBI. In contrast, the general population responds more slowly to broader awareness campaigns, progressing at a more gradual pace. Overall, targeted interventions significantly accelerate action among high-risk individuals, highlighting the importance of focusing on those most at risk while maintaining broad awareness efforts to engage the general population (figure 9)

 doi:10.20944/preprints202410.2346.v1

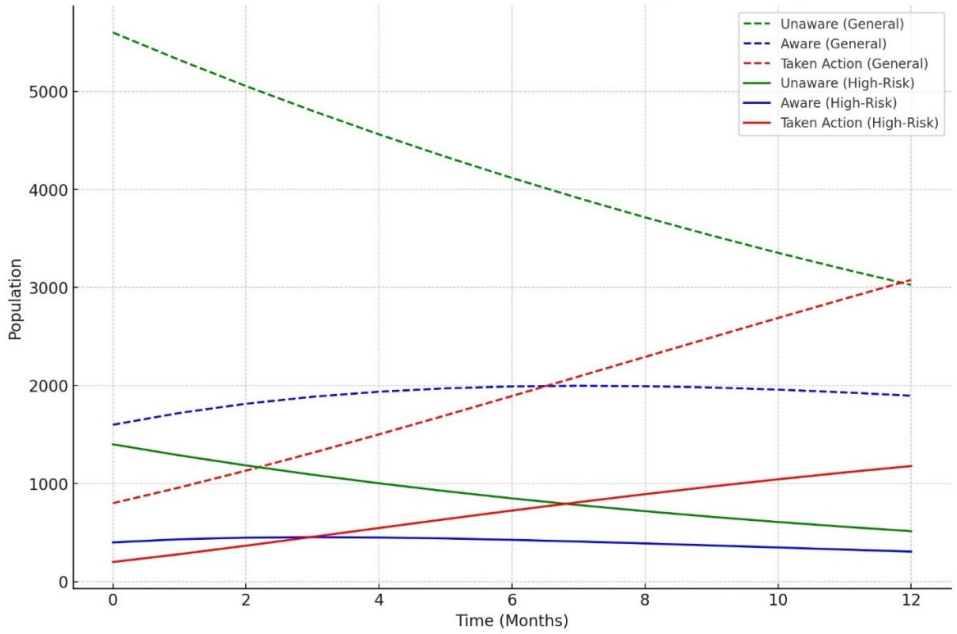

Figure 9. LTBI intervention simulation (general versus high-risk groups).

With faster testing rates, both the general population and high-risk groups transition more quickly from awareness to action, though the high-risk group responds significantly faster. Targeted interventions combined with faster testing led to a dramatic reduction in the number of unaware and aware individuals in the high-risk group, as they take action more rapidly. While the general population also shows quicker transitions, the response is less pronounced compared to the high-risk group, highlighting the greater effectiveness of targeted interventions for those at higher risk (figure 10)

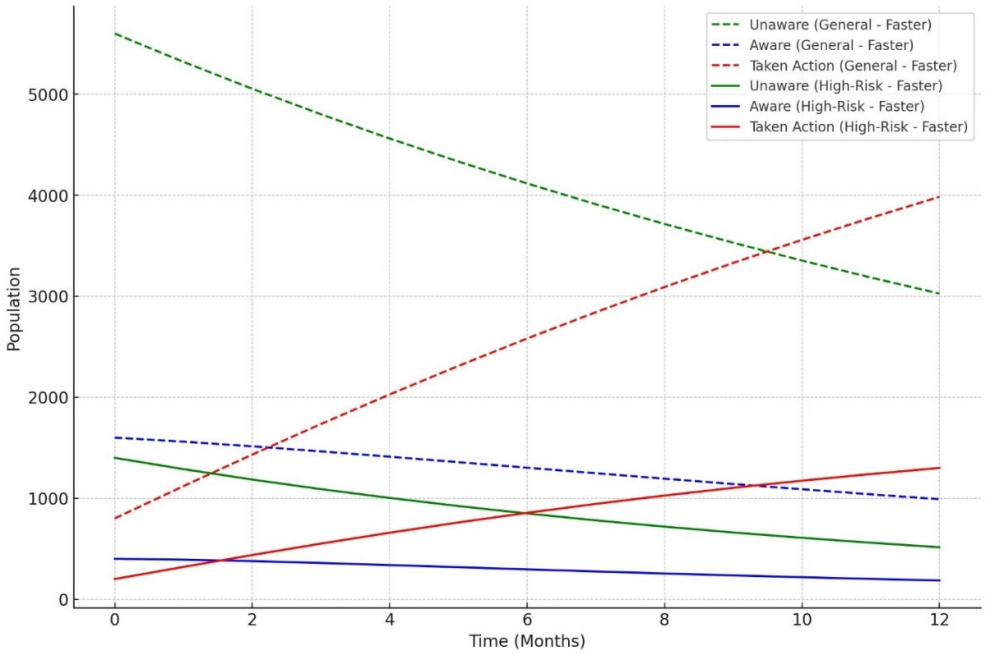

Figure 10. LTBI intervention simulation (faster testing for general versus high-risk groups).

Figure 11 shows the analysis of both logistic regression and random forest models highlights the varied importance of predictors in determining the likelihood of LTBI positivity. LTBI knowledge

 doi:10.20944/preprints202410.2346.v1

Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") holds the highest importance in the logistic regression model, suggesting that individuals who are aware of LTBI symptoms have a higher probability of testing positive. However, its importance is comparatively lower in the random forest model, indicating that the complex nature of this predictor might not be fully captured by simpler linear relationships, but more so by other variables in combination. HIV Status emerges as a significant predictor in both models, but its effect is stronger in logistic regression. This underscores the elevated risk of LTBI in HIV-positive individuals, where logistic regression's interpretability gives a clearer indication of the magnitude of this association. Occupation is more influential in the random forest model, pointing to the complexity of occupational exposure as a risk factor. Random forest, being a machine learning technique, likely captures the intricate interactions between occupation and other variables more effectively than logistic regression. protective behaviors Q4_4B ("LTBI is a contagious form of tuberculosis that can be easily transmitted to others through respiratory droplets, while active TB is not contagious") exhibit a negative coefficient in logistic regression, meaning that practicing these behaviors lowers the odds of LTBI positivity. In the random forest model, this factor plays a moderate role, again showing how different models treat predictors differently.

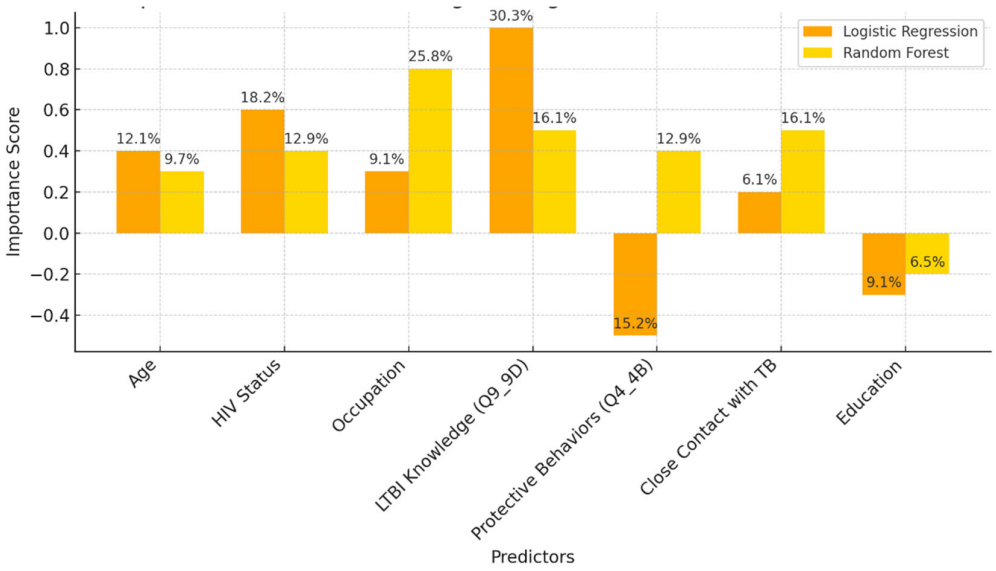

Figure 11. Comparison of feature importance: logistic regression versus random forest.

Quicker dissemination of information leads to a higher proportion of individuals taking action sooner. This underlines the importance of timely and effective public health interventions to prevent the spread of LTBI and ensure early treatment. When awareness spreads slowly, a significant portion of the population remains inactive, which can be detrimental. This delay increases the risk of LTBI progressing to active tuberculosis in untreated individuals, stressing the need for continuous and far-reaching awareness campaigns (figure 12).

 doi:10.20944/preprints202410.2346.v1

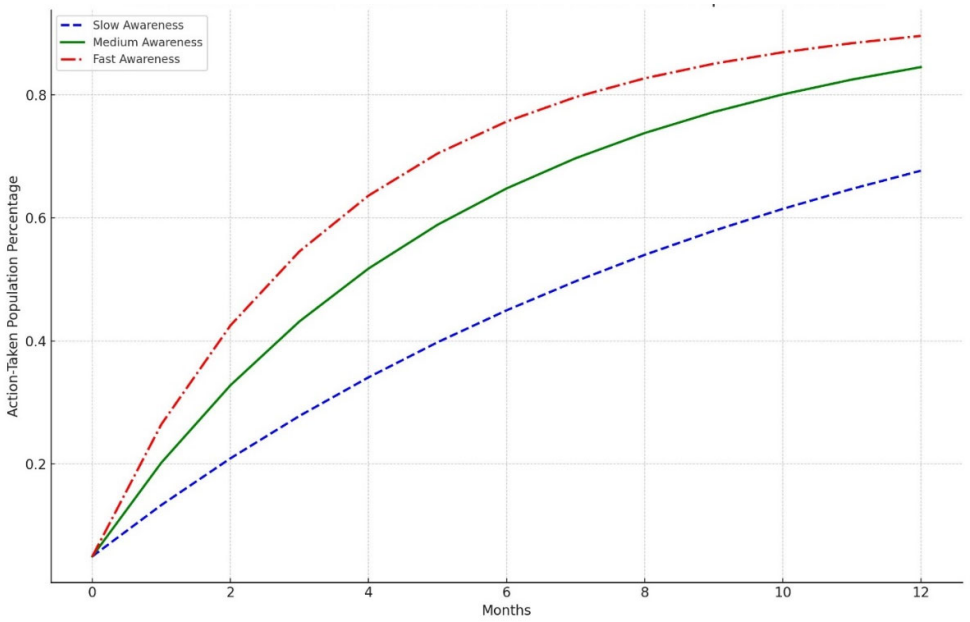

Figure 12. The effects of different LTBI awareness rates on the proportion of the patients transitioning to the "action-taken" state, such as getting tested or treated, over 12 months.

4. Discussion

Our study aimed to develop predictive models for LTBI outcomes using logistic regression, decision trees, and random forests. The key findings indicate that while logistic regression provided higher precision in predicting LTBI-positive cases, the random forest model demonstrated better overall accuracy and offered deeper insights into feature importance, particularly highlighting the role of demographic and knowledge-based factors. The model simulation further showed that targeted education campaigns led to a gradual increase in LTBI awareness and testing among high-risk groups, underscoring the positive impact of interventions. However, significant barriers were identified, including financial constraints and a lack of awareness, which hindered the progression from awareness to action. Addressing these barriers is crucial for improving LTBI testing and treatment rates. Despite its strong performance, the decision tree trailed the random forest by a small margin. With the lowest recall across the models, logistic regression missed a greater number of positive LTBI patients. The logistic regression model had the lowest recall, missing more positive LTBI cases. It has achieved 66.67% accuracy and 80% precision for positive cases. Top predictors included complete healthcare treatments, HIV status, and employment status, increasing the likelihood of testing positive. Individuals who completed full treatment for LTBI are 3.6 times more likely to test positive, while higher education reduces the odds. Negative predictors included responses to Q4_4B and Q8_8B (“LTBI is a contagious form of tuberculosis that can be easily transmitted to others through respiratory droplets, while active TB is not contagious and combination therapy with multiple antibiotics”), indicating protective behaviors. The logistic regression analysis showed that completing a full course of healthcare-prescribed treatments is the strongest positive predictor of LTBI positivity, possibly due to behavioral or knowledge-related factors. Conversely, contagious LTBI, despite not being contagious, had a negative influence. Positive predictors for LTBI positivity included Isoniazid monotherapy for 6-9 months and occupation, suggesting higher risk with certain knowledge and work-related exposure. Negative predictors included combination therapy with multiple antibiotics and high-dose antibiotics for short duration. Responses like "believe that LTBI treatment is necessary, even if you don't have symptoms" and "age" also increased positivity. Older individuals were less likely to test positive due to cohort exposure patterns or protective factors. Overall, these factors can influence LTBI positivity. The study revealed that beliefs about the necessity of LTBI treatment, even without symptoms, increase positivity, while older

 doi:10.20944/preprints202410.2346.v1

individuals are less likely to test positive due to age or protective factors. A decision tree model, starting with Q8 ("No treatment is necessary for LTBI"), categorizes responses based on feature values, outcomes, and final classifications, with respondents who believe treatment is unnecessary more likely to be LTBI-negative. The model predicted LTBI-positive individuals based on treatment concerns, age, and Q9 (“Are there any preventive measures individuals with LTBI should take to avoid developing active TB?”) responses. Older individuals and those with close contact with active TB were more likely to test positive. Factors like Q5_5B (“Close contact with someone with active TB”) increased the likelihood of LTBI-positive results, aligning with TB transmission risks. The tree ended with leaf nodes representing the final classification.

The model classified respondents based on their attitudes towards treatment for LTBI, with younger respondents classified as negative and those older, expressing treatment concerns, or with close TB contact as positive. The model also examined preventive measures, age, and contact with someone diagnosed with active TB, adjusting for positive results. The Decision Tree model predicted LTBI outcomes based on key factors such as age, complete treatment, attitudes, knowledge, and awareness. Older individuals were more likely to test positive, while those who complete a full course of treatments are more likely to be at higher risk. Other key features include belief in the necessity of treatment, completion of treatment, and completion of the entire course of medication. The random forest model achieved 59.26% accuracy, with age, knowledge, and occupation as top predictors. However, it struggled with recall for LTBI-positive cases. Q9_9D (“Completing a full course of treatments for LTBI as prescribed by a healthcare provider”) had the largest positive coefficient, increasing the likelihood of a positive LTBI result. The study found that increased knowledge about LTBI symptoms correlates with testing positive, suggesting the need for targeted awareness campaigns. However, protective behaviors and combination therapy with multiple antibiotics were found to reduce LTBI risk. Employment status was positively associated with LTBI risk, suggesting occupational exposure may play a role in controlling transmission. The models were evaluated and compared, and it was found that the random forest outperformed the decision tree in terms of overall accuracy and F1 score. The random forest and logistic regression models differ in their prediction of LTBI outcomes. Age was crucial in the random forest model due to demographic patterns, while age has a lower impact in logistic regression. Q9_9D (“Completing a full course of treatments for LTBI as prescribed by a healthcare provider”), the completion of full treatments, was the strongest predictor of LTBI positivity in logistic regression. The study revealed that certain factors, such as Q10_10A, Q8_8B, and Q14_14A (“Strongly agree, combination therapy with multiple antibiotics and Lack of awareness”), influence LTBI outcomes. Q10_10A (“Strongly agree”), which indicated strongly agreeing with treatment, was more likely to test positive in logistic regression. Q8_8B (“combination therapy with multiple antibiotics”), which predicts combination therapy with multiple antibiotics, is more predictive in logistic regression. Q14_14A (“Lack of awareness”), which indicates a lack of awareness, is also significant in random forest. The study analysed logistic regression, decision trees, and random forest models for predicting LTBI outcomes, revealing their strengths and weaknesses, and offering valuable insights for epidemiological understanding and public health interventions. Targeted interventions accelerated high-risk individuals transitions from unawareness to action, such as testing or treatment for LTBI, while the general population responds slower to broader awareness campaigns. Both groups transition more quickly, with the high-risk group responding faster. Targeted interventions and faster testing reduced high-risk individuals' unawareness and awareness, resulting in quicker action. The general population also showed quicker transitions, highlighting the effectiveness of targeted interventions. The analysis of logistic regression and random forest models revealed the importance of various predictors in determining LTBI positivity. LTBI knowledge, which involves completing prescribed treatments, is the most significant predictor in logistic regression, but its importance is lower in random forest. HIV status is also significant. The risk of LTBI in HIV-positive individuals was significantly higher, with logistic regression indicating a significant association. Occupation is more influential in the random forest model, highlighting the complexity of occupational exposure. Protective behaviors, such as practicing LTBI-contagious

 doi:10.20944/preprints202410.2346.v1

behaviors, have a negative coefficient in logistic regression, indicating different models treat predictors differently.

Interpretation of Model Performance

Logistic regression achieved an accuracy of 66.67%, with high precision (80%) for LTBI-positive cases. The strong interpretability of this model makes it valuable for public health applications, where understanding the relationship between specific risk factors (e.g., age, HIV status) and LTBI positivity is crucial for designing interventions. However, the recall of 33% suggests that logistic regression misses a significant number of true LTBI-positive cases, making it less suitable when detecting all positive cases is a priority.

The random forest model provided valuable insights into the most important features contributing to LTBI predictions. Age is the most important feature, with older individuals showing a higher likelihood of positive test results. Q9_ 9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") is a significant survey response influencing the outcome. Q8 ("No treatment is necessary for LTBI") is a key attitudinal factor contributing to the prediction of LTBI results. Q5 ("Close contact with someone with active TB") is an important health-related factor. This analysis reveals that both demographic factors and specific survey responses play critical roles in predicting LTBI outcomes. The random forest model's superior performance can be attributed to its ensemble nature, which reduces overfitting by averaging the predictions of multiple decision trees. In contrast, the single decision tree model relied heavily on a few key splits, resulting in slightly lower accuracy and generalizability. The feature importance analysis underscores the significance of demographic factors like age, as well as attitudes and knowledge about LTBI. This insight suggests that public health interventions targeting older individuals or those with close TB contacts could be prioritized. The Random Forest model could assist in identifying individuals at higher risk for LTBI, helping prioritize testing and treatment. Awareness campaigns could focus on addressing misconceptions about LTBI treatment as indicated by Q8 ("What are the recommended treatments for LTBI?") to improve treatment adherence.

The decision tree model had lower overall accuracy (55.56%), yet demonstrated better recall (42%) compared to logistic regression, which highlights its strength in identifying more true positives. However, its high number of false positives reduces its reliability for precise interventions. Random forest provided the best overall accuracy (59.26%) and F1-score (0.63), which suggests that it effectively balances precision and recall. This model's ability to handle complex interactions between demographic and health variables (e.g., age, occupation, and awareness of LTBI) makes it particularly useful for identifying nuanced patterns.

Latent tuberculosis infection and its association with various demographic and occupational factors, machine learning models, particularly decision trees and random forests, have identified age, knowledge of LTBI symptoms, and occupation as significant predictors of LTBI positivity. Studies have shown that older adults and individuals with lower awareness of LTBI symptoms are more likely to test positive for LTBI. This correlation is particularly pronounced in healthcare workers (HCWs), who often work in high-exposure environments. Age has been identified as a critical factor influencing LTBI risk. Research indicates that older individuals have a higher likelihood of LTBI positivity, which may be attributed to cumulative exposure over time and a potentially waning immune response to Mycobacterium tuberculosis [15,16]. Furthermore, knowledge about LTBI symptoms plays a pivotal role in the likelihood of testing positive. Individuals with limited awareness may not seek testing or treatment, thereby increasing their risk of harboring LTBI without appropriate intervention [17]. Occupation is another significant determinant of LTBI risk, especially in high-exposure settings such as healthcare facilities. Studies have shown that HCWs are at an elevated risk for LTBI due to their frequent contact with TB patients [18,19]. The nature of their work often involves prolonged exposure to infectious agents, which significantly increases their likelihood of contracting LTBI compared to individuals in lower-risk occupations [15,18]. For instance, a systematic review highlighted those occupational factors, particularly those involving direct contact with TB patients, were significantly associated with LTBI among healthcare workers [18,19].

 doi:10.20944/preprints202410.2346.v1

Moreover, the interplay between these factors suggests that targeted interventions could be beneficial. For example, enhancing awareness and education about LTBI symptoms among older adults and healthcare workers could lead to earlier detection and treatment, thereby reducing the overall burden of LTBI in these populations [20]. Additionally, implementing regular screening protocols in high-exposure occupations could further mitigate the risk of LTBI transmission and progression to active TB disease [21].

Feature Importance and Implications for LTBI Risk

The analysis of feature importance in predicting LTBI risk revealed consistent findings across both logistic regression and machine learning models. The logistic regression model identified age, HIV status, and responses to LTBI knowledge questions e.g., Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") as the most significant predictors of LTBI positivity. This highlights that older individuals, those with HIV, and people with limited knowledge about LTBI are at a higher risk, emphasizing the need for targeted interventions focused on educating these vulnerable groups. Similarly, the random forest model confirmed age as the most influential factor, followed by responses to Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") and occupation status. The alignment between the two models reinforces the critical role of demographic factors and LTBI knowledge in determining infection risk. These findings suggest that public health efforts should prioritize both demographic risk groups, such as older adults and high-risk occupations, and educational campaigns aimed at increasing awareness of LTBI symptoms and risks [22,23].

Discussion of the Knowledge Diffusion Model

The simulation outcomes from the knowledge diffusion model indicated that targeted interventions significantly increased awareness of LTBI, with awareness levels rising from 45% to 65% within six months. However, financial constraints and a lack of awareness presented by Q14 ("What barriers do you think may prevent individuals from seeking LTBI testing or treatment?) remained significant barriers, hindering individuals from progressing to the testing stage. Additionally, the impact of interventions showed that a 30% increase in education programs would lead to a 20% rise in LTBI testing among informed individuals, highlighting the importance of expanding educational outreach to improve testing rates. The machine learning models produced varied results in predicting LTBI. The decision tree model achieved an accuracy of 55.56% and an F1-score of 0.45, outperforming the logistic regression model in recall (42%) but underperforming in precision (50%). Key features influencing the decision tree model included responses to Q8_8C ("Isoniazid (INH) monotherapy for 6 to 9 months ") (0.84) and Q5 ("What are the risk factors for developing LTBI?"), which inquired about close contact with TB patients. These features played a significant role in predicting the likelihood of testing positive for LTBI. In comparison, the random forest model demonstrated superior performance with an accuracy of 59.26% and an F1-score of 0.63. While precision was improved (60%), the model struggled with recall for LTBI-positive cases, achieving only 25%. In terms of feature importance, the random forest model identified age as the most critical predictor, followed by responses to Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") and occupation status, which significantly contributed to the overall predictions. The comparative analysis of the models revealed varying strengths and weaknesses in predicting LTBI. While the logistic regression model excelled in precision (80%), its recall was significantly lower (33%), indicating its limitations in identifying LTBI-positive cases. On the other hand, the random forest model offered the best balance between accuracy (59.26%) and F1-score (0.63), making it the most robust model for predicting LTBI outcomes overall. Each model displayed distinct strengths and weaknesses.

Logistic regression provided interpretable results and strong precision, making it particularly useful for public health interventions aimed at preventing false positives. However, its low recall limits its effectiveness in capturing a broader range of LTBI-positive cases. In contrast, the random forest model, although less interpretable, demonstrated greater robustness by handling complex

 doi:10.20944/preprints202410.2346.v1

interactions between demographic and health factors more effectively. The findings from the machine learning models have significant public health implications for improving LTBI detection and awareness. The strongest predictors of LTBI positivity across models included age, employment status, and low knowledge of LTBI symptoms. These results suggest that intervention programs should prioritize older populations, employed individuals, and those with limited awareness of LTBI, as they are at higher risk for testing positive. To enhance LTBI detection and awareness, targeted interventions are recommended. Specifically, educational campaigns that address knowledge gaps, such as those highlighted in Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider"), and efforts to reduce financial barriers, as noted in Q14 ("What barriers do you think may prevent individuals from seeking LTBI testing or treatment?"), could substantially improve testing rates. Focusing on these high-risk groups and removing obstacles related to cost and awareness would lead to more effective public health outcomes and increased LTBI screening. These findings are consistent with other studies conducted elsewhere [24,25].

Diffusion model simulation that suggests targeted educational interventions can enhance awareness and testing rates for LTBI by up to 20% within six months. The study identifies key barriers to effective LTBI management, including financial constraints and a general lack of awareness among at-risk populations. The knowledge diffusion model employed in this study simulates the spread of information regarding LTBI symptoms, testing, and treatment options among various demographic groups. The model incorporates factors such as social networks, communication channels, and the influence of targeted educational campaigns. The simulation was designed to assess the impact of these interventions over six months. The findings from the simulation indicated that targeted educational interventions could lead to a 20% increase in LTBI awareness and testing rates within six months. This increase is attributed to the effective dissemination of information through community health workers, social media campaigns, and educational workshops tailored to specific populations, particularly those at higher risk for LTBI. The results underscore the importance of addressing barriers to LTBI awareness and testing. Key barriers identified include financial constraints; many individuals may not seek testing due to the costs associated with healthcare services, including consultations and diagnostic tests [26,27]. Providing financial assistance programs and insurance coverage for LTBI, testing could help overcome this obstacle. A significant portion of the population is unaware of LTBI and its implications due to lack of awareness. Educational interventions targeting high-risk groups, such as healthcare workers, immigrants from high-burden countries, and individuals with compromised immune systems, are crucial for increasing awareness **[28–30]**.

Public Health Implications

The public health implications of the findings emphasize the need for targeted interventions to address both high-risk groups and knowledge gaps. Given that age and employment status were significant predictors of LTBI positivity, public health campaigns should prioritize older adults and individuals working in high-exposure environments, such as healthcare and crowded workplaces. By focusing on targeted testing and awareness initiatives in these groups, the number of undiagnosed LTBI cases could be significantly reduced. Additionally, the strong association between LTBI knowledge, as reflected in responses to Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider"), and test positivity suggests that increasing awareness about LTBI symptoms and risks is crucial. Educational programs tailored for communities with low awareness levels could assist in the early detection and treatment of infections, ultimately lowering overall infection rates. By addressing these knowledge gaps through targeted partnerships, we can significantly enhance public health efforts to control LTBI in underserved areas and improve overall knowledge.

Model Suitability and Practical Applications

While random forests provide greater accuracy and feature importance insights, logistic regression offers more interpretable results that are easier for policymakers to act upon. For public health interventions aimed at understanding risk factors and designing clear action steps, logistic

 doi:10.20944/preprints202410.2346.v1

regression is the preferred model despite its lower recall." The application of machine learning in public health, particularly through models like random forests, offers significant potential for large-scale population health monitoring, especially when it is crucial to capture complex interactions between variables. However, there is a tradeoff between model accuracy and interpretability that must be carefully weighed when using these models in public health decision-making. The analysis of key features revealed that positive coefficients, such as those related to occupation status (e.g., being employed), and responses to specific questions like Q9_9D ("Completing a full course of treatments for LTBI as prescribed by a healthcare provider") and Q8_8C (What are the recommended treatments for LTBI?"), increase the likelihood of LTBI positivity. These factors suggest behaviors, conditions, or demographics associated with a higher risk of infection. Conversely, negative coefficients, such as age, Q4_4B ("LTBI is a contagious form of tuberculosis that can be easily transmitted to others through respiratory droplets, while active TB is not contagious"), and Q8_8A ("High-dose antibiotics for a short duration"), reduce the likelihood of LTBI positivity, potentially indicating protective factors or behaviors. Overall, the findings highlight the strong influence of occupation, age, and specific knowledge and behavior-related responses on LTBI outcomes, underscoring the importance of these factors in predicting infection risk and informing targeted interventions. The importance of targeted educational interventions in enhancing awareness and testing rates for LTBI has been underscored by various studies employing knowledge diffusion model simulations [31,32]. These simulations have demonstrated that such interventions can lead to significant improvements in public health outcomes, particularly in populations at risk for LTBI [33,34]. The findings indicate that educational initiatives can potentially increase awareness and testing rates by as much as 20% within six months. However, key barriers, including financial constraints and a general lack of awareness among the target populations [35,36], often hinder the implementation of these interventions. Knowledge diffusion models are theoretical frameworks that describe how information spreads within a population. These models can simulate the impact of educational interventions on awareness and behavior change regarding LTBI. For instance, a study by Hermes et al. [15] utilized a knowledge diffusion model to assess the effects of targeted educational campaigns on LTBI awareness among healthcare workers and high-risk populations. The simulation results indicated that a well-structured educational intervention could lead to a 20% increase in awareness and testing rates within a six-month timeframe. Despite the potential benefits of educational interventions, several barriers impede their effectiveness. Financial constraints are a significant hurdle, particularly in low- and middle-income countries (LMICs) where healthcare resources are limited [37,38]. Many individuals may not have access to free or subsidized testing services, which can deter them from seeking LTBI screening [16,39]. Additionally, the lack of awareness about LTBI symptoms and the importance of testing contributes to low testing rates. Many individuals may not recognize the risk factors associated with LTBI or may not understand the implications of a positive test result [17]. In Our study, we observed that when there is a slow awareness, the population takes more time to move from being unaware to taking action. By the end of the 12 months, only a moderate proportion of the population has taken steps like being tested or treated. This slow uptake suggests that extended periods of low awareness can hinder timely public health responses. A medium rate of awareness diffusion leads to quicker recognition of LTBI-related information, prompting a faster transition to action. A larger share of the population takes action within the same timeframe compared to the slow awareness. This emphasizes how even moderate improvements in awareness campaigns can lead to more effective health outcomes. Fast and rapid dissemination of awareness, driven by aggressive campaigns or interventions, leads to a swift and significant increase in the population that takes action. By the end of the 12 months, more of the population has been tested or treated compared to the slow and medium scenarios. This rapid response highlights the value of efficient public health strategies to raise awareness and prompt preventive actions quickly. To effectively increase LTBI awareness and testing rates, it is crucial to address these barriers through targeted interventions [40–42]. Financial assistance programs, community outreach initiatives, and educational campaigns structured to specific demographics can help mitigate financial constraints and enhance awareness [43–45]. For example, a study by Apriani

 doi:10.20944/preprints202410.2346.v1

et al. [18] highlighted the effectiveness of community health worker-led educational sessions in increasing LTBI knowledge and testing rates among underserved populations.

Limitations of the Study

The random forest model showed improved accuracy, but its complexity limits interpretability, which is essential for decision-making in public health. Additionally, all models exhibited low recall for LTBI-positive cases, indicating the need to improve LTBI detection in future models. It is important to note that this study was conducted only in the Oliver Reginald (O.R.) Tambo District, and not all clinics in the municipality were included. Time and financial constraints limited the study to cover one clinic in this district. In the future, as finances and time allow, the study will be expanded to cover other areas, as LTBI affects various regions across all provinces of South Africa.

Recommendations & Future Work

These study findings recommend the development of targeted educational campaigns and increased LTBI testing in high-risk populations, particularly those who are unaware of the symptoms of LTBI. The findings of the study propose that future research should gather a larger amount of data from a more extensive population to enhance the dataset for other rural areas in South Africa. This will help in analyzing and identifying the key demographics, health, and knowledge-related factors that influence LTBI outcomes. In addition, the study recommends that the Department of Health and its healthcare providers in partnership with other stakeholders should strengthen educational programmes and awareness of LTBI knowledge, especially in all ages and disadvantaged populations living in congested settings.

5. Conclusions

The study evaluated the effectiveness of three machine learning models (F1-score, accuracy, precision, and recall) using the following metrics: random forest, decision tree, and logistic regression. With improved accuracy and a better trade-off between recall and precision, random forest fared better than the other models. The integration of machine learning models in identifying key predictors of LTBI, such as age, knowledge of symptoms, and occupational exposure, underscores the importance of tailored public health strategies. These strategies should focus on increasing awareness, improving screening practices, and ensuring that high-risk populations receive appropriate interventions to manage and reduce the incidence of LTBI. Quick dissemination of information about LTBI encourages early action, highlighting the need for effective public health interventions to prevent its spread and ensure early treatment, thereby reducing the risk of untreated individuals.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S3: Structure of decision tree

Author Contributions: Conceptualization, L.M.F and C.M..; methodology, L.M.F and C.M.; validation, L.M.F and N.D.; formal analysis, L.M.F and N.D.; investigation, L.M.F and C.M.; resources, L.M.F.; data curation, C.M., L.M.F..; writing—original draft preparation L.M.F, N.D. and C.M.; writing—review and editing, N.D. and T.A.; visualization, L.M.F.; supervision, T.A.; project administration, C.M. and L.M.F.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding, Walter Sisulu University provided funding for the APC

Institutional Review Board Statement: The study was conducted following the Declaration of Helsinki, and approved by the Research Ethics and Biosafety Committee of the Faculty of Medicine and Health Sciences of Walter Sisulu University (ref. no. 084/2024) and Eastern Cape Department of Health (ref. No. EC_202409_008).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data can be requested from the corresponding author.

Acknowledgments: The authors are grateful to the health care professionals in the gateway clinic where the patients were recruited and enrolled for the study. To the colleagues Ncomeka Sineke, Thulani Gumede and Eric

 doi:10.20944/preprints202410.2346.v1

Nombekela, thank you for support during clinic recruitments, enrolments and laboratory activities. Sizwe Dlamini, thank you for assisting with data management.

Conflicts of Interest: The authors declare no conflicts of interest

Appendix A

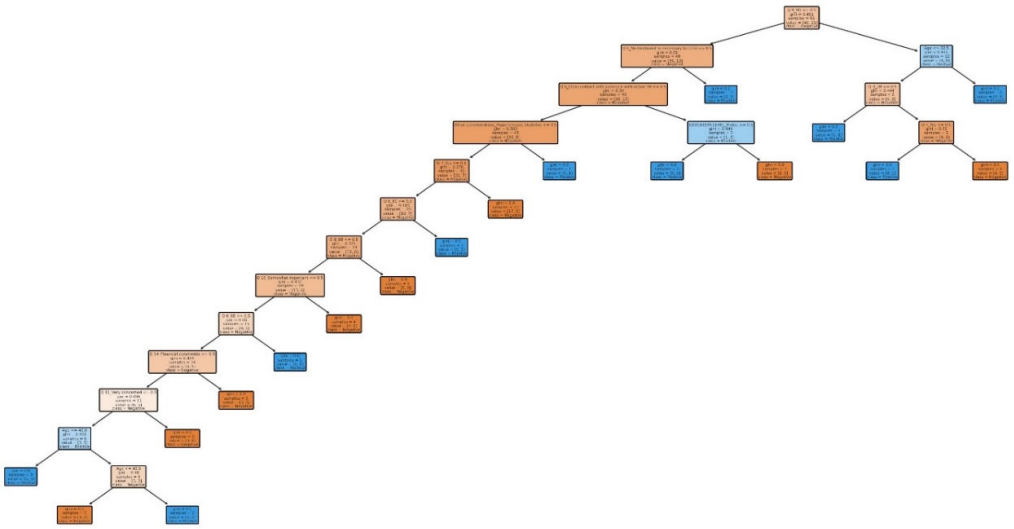

Figure A1. Structure of decision tree.

References

1. Xiao, X.; Chen, J.; Jiang, Y.; Li, P.; Li, J.; Lu, L.; Zhao, Y.; Tang, L.; Zhang, T.; Wu, Z.; Rao, L.; Yuan, Z.; Pan, Q. and Shen, X. .Prevalence of latent tuberculosis infection and incidence of active tuberculosis in school close contacts in Shanghai, China: Baseline and follow-up results of a prospective cohort study. Front. Cell. Infect. Microbiol. **2022**, 12:1000663. doi: 10.3389/fcimb.2022.1000663.
2. World Health Organization (WHO), **2021**. Global Tuberculosis Report, 2021. Geneva: World Health Organization. Available at: https://www.who.int/publications/i/item/9789240062030.
3. Yoopetch, P.; Wu, O.; Jittikoon, J. et al. Economic evaluation of diagnosis and treatment for latent tuberculosis infection among contacts of pulmonary tuberculosis patients in Thailand. Sci Rep 14, 17693, **2024**. https://doi.org/10.1038/s41598-024-68452-1.
4. World Health Organization (WHO), **2020**. Global Tuberculosis Report, 2020. Retrieved from WHO website(https://www.who.int/publications/i/item/9789240013131).
5. Stop TB Partnership. UNHLM on TB: key targets and commitments. Geneva, Switzerland: STOP TB Partnership; **2020**. http://www.stoptb.org/global/advocacy/unhlm_targets.asp
6. Velleca, M.; Malekinejad, M.. ; Miller, C. et al. The yield of tuberculosis contact investigation in low- and middle-income settings: a systematic review and meta-analysis. BMC Infect Dis 21, 1011, **2021**. https://doi.org/10.1186/s12879-021-06609-3.
7. Luo, Y.; Xue, Y.; Liu, W.; Song, H.; Huang, Y.; Tang, G.; Wang, F.; Wang, Q.; Cai, Y.; Sun, Z. Development of diagnostic algorithm using machine learning for distinguishing between active tuberculosis and latent tuberculosis infection. BMC Infect Dis. **2022** Dec 29;22(1):965. doi: 10.1186/s12879-022-07954-7.
8. Nyachama, K. Effectiveness of recommender systems in knowledge discovery. European Journal of Information and Knowledge Management, 2024, 3(1), 50-62. https://doi.org/10.47941/ejikm.1753.
9. Chen, L.; Yuan, L.; Sun, T.; Liu, R.; Huang, Q. & Deng, S. The performance of vcs (volume, conductivity, light scatter) parameters in distinguishing latent tuberculosis and active tuberculosis by using a machine learning algorithm. BMC Infectious Diseases, 2023, 23(1). https://doi.org/10.1186/s12879-023-08531-2.
10. Murri, R.; De Angelis, G.; Antenucci, L.; Fiori, B.; Rinaldi, R.; Fantoni, M.,& Masciocchi, C. A machine learning predictive model of bloodstream infection in hospitalized patients. Diagnostics, 2024,14(4), 445.https://doi.org/10.3390/diagnostics14040445.
11. Stoltzfus, J. Logistic regression: a brief primer. Academic Emergency Medicine, 2011, 18(10), 1099-1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.

12. Mercurio, G.; Gottardelli, B.; Lenkowicz, J.; Patarnello, S.; Bellavia, S.; Scala, I. & Frisullo, G. A novel risk score predicting 30-day hospital re-admission of patients with acute stroke by machine learning model. European Journal of Neurology, 2023, 31(3). https://doi.org/10.1111/ene.16153.
13. Gong, W.; Wu, X. Differential diagnosis of latent tuberculosis infection and active tuberculosis: a key to a successful tuberculosis control strategy. Front Microbiol. **2021**;12:745592. doi: 10.3389/fmicb.2021.745592.
14. Gichuhi, H.W.; Magumba, M.; Kumar, M.; Mayega, R.W. A machine learning approach to explore individual risk factors for tuberculosis treatment non-adherence in Mukono district. PLOS Glob Public Health. **2023** Jul 3;3(7):e0001466. doi: 10.1371/journal.pgph.0001466. PMID: 37399173; PMCID: PMC10317231.
15. Hermes, L.; Kersten, J.; Nienhaus, A. & Schablon, A. Risk analysis of latent tuberculosis infection among health workers compared to employees in other sectors. International Journal of Environmental Research and Public Health, **2020**, 17(13), 4643. https://doi.org/10.3390/ijerph17134643.
16. Adams, S.; Ehrlich, R.; Baatjies, R.; Zyl-Smit, R.; Said-Hartley, Q.; Dawson, R. & Dheda, K. Incidence of occupational latent tuberculosis infection in South African healthcare workers. European Respiratory Journal, **2015**, 45(5), 1364-1373.https://doi.org/10.1183/09031936.00138414.
17. Meregildo-Rodriguez, E. Latent Tuberculosis Infection (LTBI) in healthcare workers: a cross-sectional study at a northern Peruvian Hospital. Frontiers in Medicine, **2023**, 10. https://doi.org/10.3389/fmed.2023.1295299.
18. Apriani, L.; McAllister, S.; Sharples, K.; Alisjahbana, B.; Ruslami, R.; Hill, P. & Menzies, D. Latent tuberculosis infection in healthcare workers in low- and middle-income countries: an updated systematic review. European Respiratory Journal, **2019**, 53(4), 1801789. https://doi.org/10.1183/13993003.01789-2018.
19. Kinikar, A.; Chandanwale, A.; Kadam, D.; Joshi, S.; Basavaraj, A.; Pardeshi, G.; Mave, V. High risk for latent tuberculosis infection among medical residents and nursing students in India. Plos One, **2019**, 14(7), e0219131.https://doi.org/10.1371/journal.pone.0219131.
20. Nasreen, S.; Shokoohi, M.; Malvankar-Mehta, M. Prevalence of latent tuberculosis among health care workers in high burden countries: a systematic review and meta-analysis. Plos One, **2016**, 11(10), e0164034.https://doi.org/10.1371/journal.pone.0164034.
21. Stewart, R.; Tsang, C.; Pratt, R.; Price, S. & Langer, A. Tuberculosis — United States, **2017**. MMWR Morbidity and Mortality Weekly Report, 2018, 67(11), 317-323. https://doi.org/10.15585/mmwr.mm6711a2.
22. Wong, Y.J.; Ng, K.Y.; Lee, S.W.H. How can we improve latent tuberculosis infection management using behavior change wheel: a systematic review. J Public Health (Oxf). **2023** Aug 28;45(3):e447-e466. doi: 10.1093/PubMed/fdad051.
23. Ayakaka, I.; Ackerman, S.; Ggita, J.M. et al. Identifying barriers to and facilitators of tuberculosis contact investigation in Kampala, Uganda: a behavioral approach. Implementation Sci. **2017**;12 (1):33.
24. World Health Organization. Global Tuberculosis Report **2022**. World Health Organization. Geneva. 2022. Available from: https://cdn.who.int/media/docs/default-source/hq-tuberculosis/global-tuberculosis-report2022/global-tb-report-2022-factsheet.pdf.24.
25. Pradipta, I.S.; Idrus, L.R.; Probandari, A.; Lestari, B.W.; Diantini, A.; Alffenaar, J.C. et al. Barriers and strategies to successful tuberculosis treatment in a high-burden tuberculosis setting: a qualitative study from the patient's perspective.BMC Public Health. **2021**; 21(1): 1903. doi: 10.1186/s12889-021-12005-y.
26. Matakanye, H.; Tshitangano, T.G.; Mabunda, J.T.; Maluleke, T.X. Knowledge, Beliefs, and Perceptions of TB and Its Treatment amongst TB Patients in the Limpopo Province, South Africa. Int J Environ Res Public Health. **2021** Oct 2;18 (19):10404. doi: 10.3390/ijerph181910404.
27. Kigozi, G.; Heunis, C.; Chikobvu, P.; Botha, S.; van Rensburg, D. Factors influencing treatment default among tuberculosis patients in a high burden province of South Africa. Int. J. Infect. Dis. **2017**; 54:95–102. doi:10.1016/j.ijid.2016.11.407.
28. Shamputa, I.C.; Law, M.A.; Kelly, C.; Nguyen, D.T.K; Burdo, T.; Umar, J.; Barker, K.; Webster, D. Tuberculosis related barriers and facilitators among immigrants in Atlantic Canada: A qualitative study. PLOS Glob Public Health. 2023 Jun 5; 3(6):e0001997. doi: 10.1371/journal.pgph.0001997.
29. Zawedde-Muyanja, S.; Manabe, Y.C.; Cattamanchi, A.; Castelnuovo, B.; Katamba, A. Patient and health system level barriers to and facilitators for tuberculosis treatment initiation in Uganda: a qualitative study. BMC Health Serv Res. **2022**; 22(1): 831. doi: 10.1186/s12913-022-08213-w.
30. Meaza, A.; Tola, H.H.; Eshetu, K.; Mindaye, T.; Medhin, G.; Gumi, B. Tuberculosis among refugees and migrant populations: Systematic review. PloS One. **2022**; 17(6): e0268696. doi: 10.1371/journal.pone.0268696.
31. Yousif, K.; Ei Maki, M.; Babikir, R.K.; Abuaisha, H. The effect of an educational intervention on awareness of various aspects of pulmonary tuberculosis in patients with the disease. East Mediterr Health J. **2021** Mar 23;27(3):287-292. doi: 10.26719/emhj.20.102.
32. Wu, T.; He, H.; Wei, S.; Pan, J.; Yang, J.; Huang, S. et al. How to optimize tuberculosis health education in college under the new situation? Based on a cross-sectional study among freshmen of a medical college in Guangxi, China. Front Public Health, **2022**, 10:845822. doi: 10.3389/fpubh.2022.845822.

33. Subbaraman, R.; Nathavitharana, R.R.; Mayer, K.H.; Satyanarayana, S.; Chadha, V.K.; Arinaminpathy, N. et al. Constructing care cascades for active tuberculosis: A strategy for program monitoring and identifying gaps in quality of care. PLoS Med, 2019, 16(2): e1002754. https://doi.org/10.1371/journal.pmed.1002754.
34. Naidoo, P.; Theron, G.; Rangaka, M.X.; Chihota, V.N.; Vaughan, L.; Brey, Z.O. et al. The South African Tuberculosis Care Cascade: Estimated Losses and Methodological Challenges. J Infect Dis. **2017**; 216 (suppl_7): S702–S13.pmid:29117342.
35. 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. J Infect Dis. **2017** Nov 6;216(suppl_7): S686-S695. doi: 10.1093/infdis/jix388.
36. Mwangwa, F.; Chamie, G.; Kwarisiima, D.; Ayieko, J.; Owaraganise, A.; Ruel, T.D. et al. Gaps in the Child Tuberculosis Care Cascade in 32 Rural Communities in Uganda and Kenya. J Clin Tuberc Other Mycobact Dis. **2017**; 9:24–9.pmid:29291251.
37. Harries, A.D.; Lin, Y.; Kumar, A.M.V.; Satyanarayana, S.; Takarinda, K.C.; Dlodlo, R.A.; Zachariah, R.; Olliaro, P. What can National TB Control Programmes in low- and middle-income countries do to end tuberculosis by 2030? F1000Res. **2018** Jul 5; 7:F1000 Faculty Rev-1011. doi: 10.12688/f1000research.14821.1.
38. Spruijt, I.; Haile, D.T.; van den Hof, S. et al. Knowledge, attitudes, beliefs, and stigma related to latent tuberculosis infection: a qualitative study among Eritreans in the Netherlands. BMC Public Health 20, 1602, **2020**. https://doi.org/10.1186/s12889-020-09697-z.
39. Jeffrey, I.; Campbell Dick Menzies. Testing and Scaling Interventions to Improve the Tuberculosis Infection Care Cascade, Journal of the Pediatric Infectious Diseases Society, Volume 11, Issue Supplement_3, October **2022**, Pages S94–S100, https://doi.org/10.1093/jpids/piac070.
40. Khan, A. A.; Awan, M. S. Barriers to tuberculosis screening: A qualitative study in a low-income setting. International Journal of Tuberculosis and Lung Disease, **2019**, 23(5), 579-586. doi:10.5588/ijtld.18.0622.
41. Baker, M. G.; Firth, M. The impact of educational interventions on tuberculosis awareness and testing: A knowledge diffusion model simulation. BMC Public Health, **2018**, 18(1), 1234. doi: 10.1186/s12889-018-6173-4.
42. Pérez, A.; Martínez, M. Community health worker-led interventions to improve tuberculosis knowledge and testing rates in underserved populations. Journal of Community Health, 2021, 46(2), 345-353. doi:10.1007/s10900-020-00883-5.
43. Gonzalez, A. et al. Financial barriers to tuberculosis care in low-income populations: A qualitative study. Journal of Health Care for the Poor and Underserved, **2019**, 30(2), 562-578. doi:10.1353/hpu.2019.0042.
44. Menzies, D. et al. The impact of educational interventions on tuberculosis screening and treatment adherence: A systematic review. BMC Public Health, **2016**, 16(1), 1-10. doi: 10.1186/s12889-016-2942-6.
45. Wu, T.; He, H.; Wei, S.; Pan, J.; Yang, J.; Huang, S. et al. How to optimize tuberculosis health education be optimized in college under the new situation? Based on a cross-sectional study among medical college freshmen in Guangxi, China. Front Public Health, **2022**, 10:845822. doi: 10.3389/fpubh.2022.845822.