REPUBLIQUE DE COTE D'IVOIRE

Union-Discipline-Travail

Ministère de l'enseignement supérieur et de la recherche Scientifique

---

CANDIDATURE A L'INSCRIPTION SUR LA LISTE

D'APTITUDE AUX FONCTIONS DE MAITRE-ASSISTANT

(LAFMA)

CTS : SCIENCES ET TECHNIQUE DE L'INGENIEUR

DISCIPLINE : INFORMATQUE

Yazid HAMBALLY YACOUBA, Amadou DIABAGATE, Hafizatou SANI YANOUSSA, Adama COULIBALY, Abdellah AZMANI, **Prediction of the Knowledge Level of Tuberculosis**. Open Journal of Applied Sciences, 15, 3401-3425. doi: 10.4236/ojapps.2025.1511219.

Yazid HAMBALLY YACOUBA

ASSISTANT

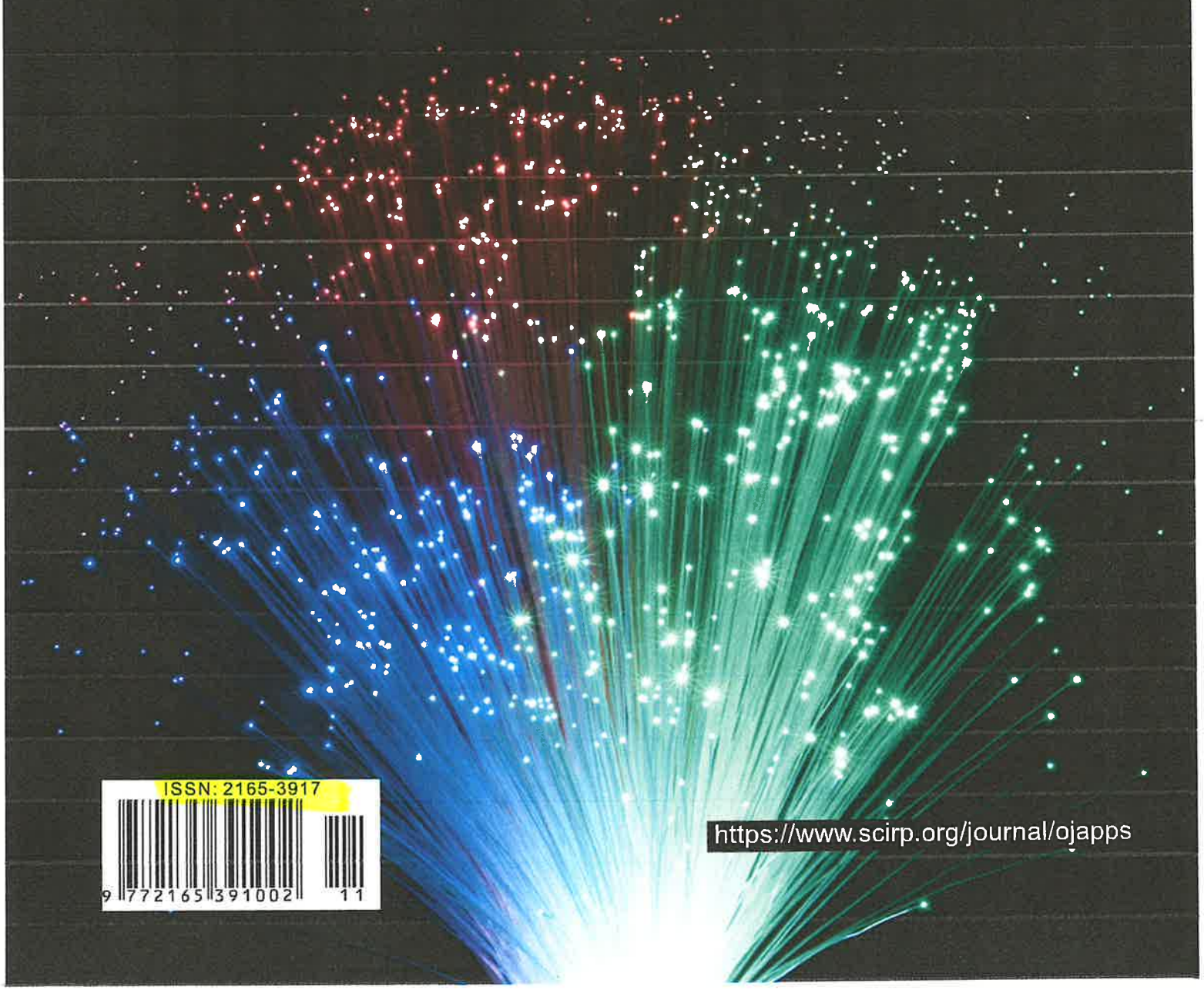
ENSAU, BP V 81 Bondoukou, Tel : +2250789404055

Email : yacouba.hambally@ubkou.edu.ci

ISSN: 2165-3917 Volume 15, Number 11, November 2025

Scientific Research Publishing

Open Journal of Applied Sciences



ISSN: 2165-3917

https://www.scirp.org/journal/ojapps

Journal Editorial Board

ISSN Print: 2165-3917 ISSN Online: 2165-3925

https://www.scirp.org/journal/ojapps

Editor-in-Chief

Editorial Board

Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(11), 3401-3794
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Table of Contents

Volume 15 Number 11 November 2025

Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(11), 3401-3794
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(11), 3401-3794
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Open Journal of Applied Sciences (OJAppS)
Journal Information

SUBSCRIPTIONS

The *Open Journal of Applied Sciences* (Online at Scientific Research Publishing, https://www.scirp.org/) is published monthly by Scientific Research Publishing, Inc., USA.

Subscription rates:
Print: $79 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: ojapps@scirp.org

Call for Papers

Open Journal of Applied Sciences

ISSN Print: 2165-3917 ISSN Online: 2165-3925

https://www.scirp.org/journal/ojapps

Open Journal of Applied Sciences **(OJAppS)** is an international peer-reviewed, open-access journal publishing in English original research studies, reviews in diverse areas of applied sciences. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in applied sciences and to keep a record of the state-of-the-art research in related areas.

Subject Coverage

The journal publishes original papers including but not limited to the following fields:

Applied Behavioral Science
Applied Biology
Applied Chemistry
Applied Ecology
Applied Economics
Applied Engineering
Applied Genetics
Applied Geography
Applied Linguistics
Applied Mathematics
Applied Mechanics
Applied Philosophy
Applied Physics
Applied Physiology
Applied Psychology
Applied Sociology
Artificial Intelligence
Computing Technology

We are also interested in: 1) Short Reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book Reviews—Comments and critiques.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submission, please access the website.

Website and E-Mail

https://www.scirp.org/journal/ojapps **E-mail: ojapps@scirp.org**

Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(11), 3401-3425
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Prediction of the Knowledge Level of Tuberculosis

Yazid Hambally Yacouba[1*], Amadou Diabagaté[2], Hafizatou Sani Yanoussa[3], Adama Coulibaly[2], Abdellah Azmani[4]

[1]National High School of Architecture and Urban Planning, University of Bondoukou, Bondoukou, Côte d'Ivoire
[2]Faculty of Mathematics and Computer Science, University Félix Houphouet-Boigny, Abidjan, Côte d'Ivoire
[3]Emy Polyclinic, Abidjan, Côte d'Ivoire
[4]Faculty of Sciences and Technologies, University Abdelmalek Essaadi, Tangier, Morocco
Email: *yazid.hambally@gmail.com, *yacouba.hambally@ubkou.edu.ci

How to cite this paper: Hambally Yacouba, Y., Diabagaté, A., Sani Yanoussa, H., Coulibaly, A. and Azmani, A. (2025) Prediction of the Knowledge Level of Tuberculosis. *Open Journal of Applied Sciences*, **15**, 3401-3425.
https://doi.org/10.4236/ojapps.2025.1511219

Received: October 9, 2025
Accepted: October 28, 2025
Published: October 31, 2025

Abstract

Predicting knowledge of tuberculosis (TB) could imply several significant changes in the management, control and prevention of this disease. These would be based on advanced technological and organizational approaches, with benefits for both the detection, treatment, and prevention of tuberculosis. The use of predictive models based on artificial intelligence (AI) and machine learning to analyze clinical, epidemiological and biological data of patients could contribute to improving screening. Assessing the level of knowledge about tuberculosis is essential to improve prevention, diagnosis and treatment of this disease. It allows for the design of targeted programs to reduce stigma and strengthen health systems. The approach adopted is to conduct a semidirected questionnaire to assess the level of knowledge of tuberculosis through multiple choice, true/false or short answer questions. The data collected is then processed by machine learning algorithms to obtain results that will be analyzed. The collection and analysis of socioeconomic, geographic and demographic data make it possible to identify the most vulnerable populations and geographic areas at high risk of spread. Machine learning can be used to predict knowledge levels based on variables such as education, geography, access to information and health behaviors.

Keywords

Machine Learning, Knowledge of Tuberculosis, Health Education, Artificial Intelligence, Decision Support

1. Introduction

Tuberculosis (TB) remains one of the world's deadliest infectious diseases, partic-

ularly in low- and middle-income countries, where social, economic, and educational barriers hinder prevention and early detection efforts. According to the World Health Organization, despite global progress, TB continues to claim over 1.3 million lives annually, with the highest burden concentrated on resource-limited settings [1]. The incidence of tuberculosis has fallen by an average of 1.5% per year since 2000, but this decline is much slower in certain countries in Africa, Asia and Eastern Europe because of poverty, HIV infection and the exodus of vulnerable populations [2].

In this context, understanding and enhancing public knowledge about TB symptoms, transmission, and prevention strategies is crucial for breaking the chain of infection and achieving Sustainable Development Goal 3 (SDG 3) on good health and well-being. However, traditional methods of assessing public awareness, such as manual surveys and aggregated statistics, often fail to capture granular insights and cannot easily scale for real-time or large-scale analysis. At the same time, the growing availability of survey data and the advancement of machine learning (ML) techniques provide an unprecedented opportunity to model, predict, and understand patterns in health-related knowledge including for tuberculosis.

This study contributes to this growing field by exploring how ML models can be used to predict the level of tuberculosis knowledge from structured survey data. Specifically, we:

- Compare the performance of several regression algorithms (Linear Regression, SVM, Random Forest, and Neural Networks).
- Identify the most influential variables affecting knowledge levels.
- Assess the interpretability of models using SHAP values.

By doing so, this work aims to support data-informed public health strategies, allowing for more targeted education campaigns and adaptive interventions that respond to real knowledge gaps in the population. The approach also illustrates how artificial intelligence can strengthen precision health policies in the fight against tuberculosis, especially in underserved communities.

Many studies use qualitative methods or categorical questionnaires to assess TB knowledge. However, using a numerical score allows for a quantitative and continuous measurement of this level of knowledge while providing finer granularity and greater precision. A numerical score will indeed allow interventions to be targeted according to the specific level of knowledge of individuals or groups, which will thus improve the effectiveness of TB awareness campaigns. The use of a numerical score can rely on innovative technologies such as machine learning algorithms to analyze responses and calculate scores. A numerical score can also be used to measure the impact of TB awareness campaigns more precisely by quantifying changes in the level of knowledge before and after an intervention. A new publication that uses a numerical score to assess the level of knowledge about tuberculosis provides an innovative and rigorous approach that stands out from existing publications in that it allows for more detailed analysis, easier comparisons, and more targeted interventions that will contribute to a better understanding and

more effective fight against tuberculosis.

To optimize control strategies, understanding how to predict TB in different populations can play a crucial role in public health policies. A better understanding of the disease among populations is crucial for improving prevention and treatment efforts. Artificial intelligence (AI) makes it possible to analyze complex data and predict knowledge of tuberculosis, enabling more targeted public health interventions.

TB knowledge can be predicted through various methods and approaches depending on the context and specific objectives. Questionnaires and surveys can be designed to assess individuals' knowledge about tuberculosis, including transmission, symptoms, treatment and prevention. Studying the correlation between the level of knowledge about tuberculosis and demographic (age, gender, education, etc.) and socioeconomic (income, employment, social status) factors can help predict knowledge of the disease.

The data collected as part of this work come from Niger. In Niger in 2015, the National Tuberculosis Control Program (PNLT) estimated 7115 new cases of microscopy-positive tuberculosis [3]. The capital of Niger (Niamey) had an estimated population of 1,057,347 inhabitants in 2015 according to the National Statistics Institute (INS) [4].

The general objective of this work is to study the knowledge levels of the population on tuberculosis in Niamey, and the specific objectives are as follows:

- Evaluate knowledge on the mode of transmission of tuberculosis.
- Measure the degree of knowledge about the germ causing the disease.
- Determine knowledge of symptoms suggestive of tuberculosis.
- Knowing the means of treating tuberculosis.

To achieve these objectives, we followed the plan below:

- The first part presents the problem.
- The second part is related to the literature review.
- The third part addresses methods for predicting knowledge of tuberculosis.
- The fourth part concerns the framework for collecting the data used, the results obtained and the interpretation of these results.
- The fifth part presents the results obtained.
- The sixth part presents a discussion of the results in light of the literature review.
- The seventh section is devoted to the conclusion and suggestions.

2. Problematic

Tuberculosis is a public health problem because of its severity, extent worldwide and economic weight. It is one of the main causes of death from infectious diseases worldwide according to estimates from the World Health Organization (WHO) in 2014 [1]. The incidence of tuberculosis has fallen by an average of 1.5% per year since 2000, but this decline is much slower in certain countries in Africa, Asia and Eastern Europe because of poverty, HIV infection and the exodus of vulnerable

populations [2].

Knowledge of tuberculosis varies depending on several factors around the world. Knowledge of TB is complex among young adults and adolescents due to unverified and pervasive information despite challenges in the early detection of TB [5] [6]. The authors of [7]-[9] sought to explore the epidemiology of tuberculosis and, in particular, the role of socioeconomic factors in determining how disadvantaged populations experience tuberculosis. Assessing knowledge of tuberculosis among disadvantaged populations requires taking into account the level of education, access to information and economic barriers that limit access to care and prevention. Humans respond to, understand, and treat infectious diseases worldwide through the lens of anthropology [10]. Macdonald and Harper [11] take an in-depth look at tuberculosis and show that barriers related to geographic isolation make it difficult to identify gaps in knowledge of the disease. Precarious living conditions and limited access to health care are particularly favorable for the spread of tuberculosis, especially in low- and middle-income countries [12] [13]. Ignorance of symptoms and modes of transmission and prevention strategies, especially among vulnerable populations, aggravates the tuberculosis situation in India and contributes to the persistence of the disease and delays in seeking care [14] [15].

Geographic isolation in certain regions of Niger makes access to awareness campaigns and tuberculosis screening services difficult. This hampers the collection of data on disease, especially in remote areas [16]. Popular beliefs and social stigma related to tuberculosis in the Enugu region of Nigeria may hinder recognition of symptoms of the disease and deter people from seeking treatment [17]. The assessment tools used to measure knowledge of tuberculosis are not always adapted to the specificities of the populations in eastern Algeria, particularly due to cultural differences, linguistic diversity and literacy levels [18]. The lack of health infrastructure and trained personnel in rural areas leads to low tuberculosis screening capacity, making it difficult to assess the rate of awareness of the disease among these populations [19].

The assessment of knowledge of tuberculosis is essential and complex at several levels, particularly the following:

- design educational strategies that counter misinformation and promote positive health behavior;
- help identify appropriate strategies to overcome social disparities and improve access to health;
- develop adapted interventions aimed at improving prevention, screening and adherence to treatment;
- understand not only factual knowledge but also social and cultural perceptions and health behaviors influenced by external factors such as stigma and discrimination;
- measures the effectiveness of awareness campaigns, adapts prevention and treatment strategies, and targets groups at risk.

Knowledge of tuberculosis (TB) is crucial in the fight against this infectious disease, especially in areas with poor access to health care. A good understanding of tuberculosis will make it possible to identify the levers necessary to improve the situation and fight more effectively against this preventable and treatable disease.

Machine learning algorithms can be used to predict TB knowledge by analyzing structured and unstructured data. To do this, we use specific measures and indicators that make it possible to quantify the relationships between the characteristics of individuals (demographic factors, health behaviors, etc.) and their level of knowledge about tuberculosis.

3. Literature Review

Artificial intelligence (AI) represents a significant advancement in the analysis and prediction of tuberculosis (TB) knowledge, particularly in areas with high prevalence rates. By analyzing complex data and using machine learning models, AI can identify gaps in people's understanding, predict their level of knowledge and adapt public health interventions.

This review explores recent approaches that use AI to predict and analyze TB knowledge, highlighting supervised learning models.

Machine learning (ML) is used to analyze complex data from different sources to predict TB knowledge among different populations. ML algorithms, such as decision trees, random forests, artificial neural networks, and support vector machines (SVMs), are used to identify factors influencing the understanding of TB and to predict knowledge levels in specific individuals or groups [20]-[23].

Supervised models such as random forests or SVMs make it possible to classify individuals according to their level of knowledge about tuberculosis via survey data or medical histories [20]. Neural networks can be applied to process nonlinear data and discover complex relationships between variables influencing TB knowledge [21].

The results from Harris Miriam [22] revealed that deep learning techniques can predict higher levels of diagnostic accuracy than can human radiologists in the interpretation of chest radiographs for pulmonary tuberculosis. Mohidem *et al.* [23] emphasized that neural networks can predict multifactorial phenomena such as tuberculosis, which makes them more suitable for improving public health policies.

Deep neural networks are a subcategory of machine learning algorithms capable of processing unstructured and complex data, such as images, text or behavioral signals. These models have been used to predict TB knowledge by processing large datasets, including demographic information, survey responses, and even medical histories [24]-[27].

Reference [24] is a review of recent research based on the application of artificial intelligence for the management of infectious diseases. Previous studies [25] have shown that artificial intelligence and machine learning play increasingly cen-

tral roles in the diagnosis and management of diseases. This study [26] highlights the application of deep learning techniques for tuberculosis diagnosis in India. This study [27] uses a convolutional neural network (CNN) for image feature extraction to improve pneumonia detection. This paper [28] provides an in-depth analysis of data mining for predicting TB contagion risk.

Understanding the relationships between environmental factors, including weather and air quality, and the incidence of tuberculosis can help predict epidemics and implement targeted prevention measures [29]. Linear regression is used to analyze how factors such as age, gender, socioeconomic status, and education affect knowledge of TB in a given population [30]-[32]. Vimala Balakrishnan *et al.* used a support vector machine regression-based approach to improve prediction accuracy by accounting for various clinical and demographic factors [33].

An ANN model was used to predict knowledge of tuberculosis from variables such as age, gender, clinical symptoms, laboratory test results, and medical history [34].

An artificial neural network model was designed and implemented to classify patients as having TB from data that include patient information such as age, gender, clinical symptoms, laboratory test results, and medical history [35].

Random forests showed the best predictive performance, with high accuracy and a good ability to identify patients at risk of nonadherence using variables such as age, gender, education level, distance to health center, treatment side effects, and cultural beliefs from data collected from TB patients in the Mukono district, Uganda [36].

Several machine learning algorithms (logistic regression, random forests, gradient boosting, and neural networks) have been used to predict treatment success via data from pulmonary tuberculosis patients, including variables such as age, gender, comorbidities, microbiological test results, treatment regimens, and radiological data [37].

This study aims to develop a machine learning model to predict TB detection via data from trained African giant rats to improve diagnostic efficiency and accuracy [38].

4. Methodology

4.1. Data Collection Methodology

This work uses data collected during a study that took place in Niamey, capital of Niger, during the month of July 2016 to assess the population's knowledge levels of tuberculosis. The sample size of this study was limited to 507 individuals distributed across the five municipal districts of the city of Niamey as follows:

Individuals were targeted for data collection according to the following characteristics:

- At least 15 years old, having agreed to freely answer the questionnaires;
- Having all their mental capacity.

Individuals with the following characteristics were not taken into account in the study:

- Under the age of 15;
- Those who refused to participate in the study;
- Suffering from mental deficiency.

A semi-directive questionnaire was used in which the subject had to answer YES, NO or DON'T KNOW to choose a single answer from several proposed answers.

A free questionnaire was used for group interviews, in which subjects offered several answers. For general knowledge of tuberculosis, each correct answer was given one point for each of the ten questions asked.

The surveys were classified into knowledge categories, namely, good knowledge (8 - 10/10 points), average knowledge (5 - 7/10 points), and insufficient knowledge (0 - 4/10 points), on the basis of their response to this knowledge.

For the ten questions, the correct answers were as follows:

- For the definition of tuberculosis: what is tuberculosis, you had to say "YES" and suggest at least a chronic cough, fever, and weight loss (in the symptoms).
- For contagiousness, it was necessary to say "YES" and suggest that transmission takes place by air.
- For the causative agent, it was necessary to say the "koch bacillus".
- Whatever the location, it was taken into account.
- Whatever the source of information, it was taken into account.
- For curability, it was necessary to say: "YES, it is curable".
- Whatever the risk factors, they were taken into account.
- For diagnosis, it was necessary to offer sputum examination or X-ray.
- For treatment, it was necessary to say: "anti-tuberculosis drugs".
- For the duration of treatment, it was necessary to propose: "6 months, more than 6 months or 8 months".

On the basis of the scores obtained on knowledge, the respondents are classified into three levels of knowledge, and this distribution makes it possible to look for relationships between these respondents. The choice of respondents was made on the basis of neighborhoods and age groups. The group interviews took place within households with the subjects, in public places or on the street. Emphasis was placed on the fact that participation is free and voluntary.

The different variables studied are sociodemographic characteristics (gender, age, marital status, level of education, city of residence) and the general theoretical level of knowledge about tuberculosis (signs, transmission, sources of information, causal agent, locations of tuberculosis, risk factors, diagnosis, treatment).

This study takes into account the dignity, privacy and freedom of those interviewed by ensuring, in particular, the confidentiality of the identity of the respondents through an anonymous questionnaire.

4.2. Sampling Method and Distribution of Respondents

The study is based on a stratified random sampling approach grounded in the five municipal districts of Niamey. Each district was considered as a stratum, and re-

spondents were randomly selected within each stratum, ensuring a geographic representativeness of opinions across the city of Niamey. This method combines scientific rigor with territorial balance.

The 507 survey participants were distributed across the five districts as shown in **Table 1**.

Table 1. Distribution of 507 respondents.

Municipal District	Number of Respondents	Approximate Percentage
Niamey 1	99	19.5%
Niamey 2	102	20.2%
Niamey 3	100	19.7%
Niamey 4	103	20.3%
Niamey 5	103	20.3%
Total	507	100%

This distribution reveals an almost equal number of respondents across all districts, as detailed in **Table 1**, ensuring a well-balanced dataset for comparative analysis and enhancing the external validity of the study.

The chosen method guarantees homogeneous coverage of the urban territory of Niamey, minimizes selection bias, and ensures statistical reliability for extrapolating the results to the city's entire population.

4.3. Defining Predictive Variables

The table of predictive variables and their coding system is presented below in **Table 2**.

Table 2. Predictive variables and their coding system.

Predictive Variable	Definition/Description	Type of Coding
Gender	Respondent's gender	One-hot (male, female)
Age	Respondent's age group	Ordinal (15 - 25 = 1, 26 - 44 = 2, etc.)
Marital Status	Marital status (single, married, etc.)	One-hot (single, married, divorced, widower, etc.)
Town	Place of residence	One-hot (Niamey 1, Niamey 2, etc.)
Heard about tuberculosis	Has ever heard about TB	Binary: 0 = No, 1 = Yes
Perception of tuberculosis form	Personal assessment of the seriousness of the disease	One-hot (fatal disease, mild illness, etc.)
Knowledge about the causal agent of tuberculosis	Knowledge specifically related to the causal agent	Binary: 0 = No, 1 = Yes
Localization of the causal agent	Knowledge of the different organs affected by the causal agent	One-hot (pulmonary, pleural, bone, don't know, etc.)
Information sources used	Media or communication channels used	One-hot (television, radio, social networks, etc.)
Confidence in the curability of tuberculosis	Declared level of confidence in the curability of the disease	Binary: 0 = No, 1 = Yes

Continued

Knowledge about tuberculosis risk factors	History of illness or exposure	Numeric (aggregated score)
Confidence in tuberculosis diagnosis (curability)	Declared level of confidence in the effectiveness of the diagnosis	One-hot (sputum test, X-ray, etc.)
Knowledge about tuberculosis treatment	Dosage	Numeric (aggregated score)
Duration of tuberculosis treatment	Knowledge about different treatment durations	One-hot (6 months, more than 6 months, 8 months, etc.)

The dependent variable is the level of knowledge about tuberculosis, and the knowledge score results from the weighted sum of correct answers.

The coding system may vary depending on the software used for data analysis, but the underlying logic remains consistent.

4.4. Tuberculosis Knowledge Score and Modeling Approach

The tuberculosis knowledge score is calculated by summing the correct answers to a series of closed-ended questions.

This score can therefore take multiple integer values between 0 and 10, depending on the number of correct responses provided by each participant.

This score represents a graduated quantitative measure each additional point reflects an actual gain in knowledge. It is based on a fixed-interval scale and may vary considerably between individuals, without natural thresholds or predefined categories.

This type of score is suitable for modeling as a continuous variable, as it meets the conditions of an interval scale and can be treated accordingly in a linear regression analysis.

Some might consider grouping the scores into three categories (low, medium, high), but such classification would result in a loss of information. The regression approach is therefore preferable for several reasons:

1) Preserving Data Granularity

a) Keeping the variable as numeric preserves the full richness of the data.

b) Categorizing into three levels would imply arbitrary thresholds, introducing methodological noise.

2) Analytical Objective: Predicting the Effect of Explanatory Variables on the Score

a) The goal is to measure the impact of predictive factors (e.g., gender, age, occupation, perception) on the knowledge level.

b) Regression provides estimated marginal effects (e.g., men score on average 0.8 points higher than women), which classification methods cannot deliver as precisely.

3) Statistical Power Gain

a) Treating the scores as continuous variables increases the sensitivity of statis-

tical tests.

b) Less information loss = better ability to detect real effects.

4) Reducing Interpretation Bias

a) Class-based methods impose arbitrary thresholds that are difficult to justify empirically.

b) Regression yields coefficients directly interpretable on the original scale.

A classification approach would only be suitable if:

- The objective was to categorize respondents according to policy or decision thresholds (e.g., urgent need for tuberculosis training if score < 4).
- The model focused on probabilities of group membership rather than score progression.

But that is not the case here, where the objective is to measure the effect of predictors on the knowledge level measured by a score.

4.5. Proposed Methodology for Predicting Tuberculosis Knowledge Levels

1) Input

a) A dataset of 507 observations collected via a survey on knowledge of tuberculosis.

2) Output

a) Trained regression models capable of predicting knowledge levels.

b) Interpretability of the predictions, identifying the most influential features.

3) Step-by-Step Methodological Workflow

a) Import Libraries and Packages

- Load Python libraries essential for data science (e.g., pandas, numpy, scikit-learn, TensorFlow, SHAP).

b) Load Dataset

- Import the raw CSV or Excel file containing the 507 survey records.

c) Data Cleaning

- Handle missing values or outliers through imputation or filtering.

d) Feature Engineering

- Encode categorical variables using one-hot or label encoding.
- Standardize numerical features to normalize their ranges (e.g., z-score).

e) Dataset Splitting

- Divide the final cleaned dataset into:
  - Training set (80%)
  - Testing set (20%)

f) Model Selection

Define and prepare five candidate models for comparison:

- Dummy Regressor (baseline)
- Linear Regression
- Random Forest Regressor
- SVM Regressor (Support Vector Machine)

- Artificial Neural Network (ANN)

g) Hyperparameter Tuning

- Use RandomizedSearchCV to fine-tune model parameters for all models except Dummy Regressor.

h) Training and Validation

- Apply Stratified K-Fold Cross-Validation to train and validate each model robustly and reduce overfitting risk.

i) Performance Evaluation

Evaluate all models using five regression metrics:

- R^2 (explained variance)
- RMSE (root mean squared error)
- MAE (mean absolute error)
- MSE (mean squared error)
- MAPE (mean absolute percentage error)

j) Model Interpretability with SHAP

- Use SHAP (SHapley Additive exPlanations) to interpret the best-performing model.
  - Identify the most influential variables.
  - Visualize their impact (positive or negative) on prediction outcomes.

This robust, multi-model regression pipeline ensures predictive accuracy and interpretability, making it suitable for public health decision-making related to tuberculosis knowledge dissemination. The inclusion of SHAP enhances transparency, allowing stakeholders to understand which factors most influence knowledge levels across the population.

5. Results

5.1. Presentation of Indicators

The mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) metrics are specific indicators that help quantify how accurate and reliable the model is in predicting TB knowledge levels, especially when working with continuous variables or regression tasks. The MAE is the average of the absolute differences between the actual values and the values predicted by the model. This gives an idea of the average error in absolute terms without taking into account the direction (positive or negative) of the error.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

where y_i is the actual value and where $\hat{y}_i$ is the predicted value.

For the prediction of a knowledge score between 0 and 10 to assess knowledge of tuberculosis, the MAE measures the average difference between the actual knowledge score and the predictive score of the model. The random forest regressor, SVM regressor, and linear regression record the three lowest MAEs, indicating that these models make small errors in predicting the knowledge level.

The MSE is the average of the squares of the errors, which penalizes large errors (those that are far from the true value) more because the errors are squared.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{2}$$

where y_i is the actual value and where $\hat{y}_i$ is the predicted value.

For predicting a knowledge score, the MSE gives you an idea of the error variance. A low MSE means that the errors are generally small, whereas a high MSE indicates that the errors are large and can destabilize the model. The random forest regressor, linear regression and SVM regressor yield the three smallest errors in predicting the knowledge level.

The RMSE is the square root of the MSE. It represents the average error in a more intuitive form by putting the error scale in the same unit as the target variable, which, in our case, is the knowledge score.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{3}$$

The RMSE is a measure that is often more understandable than the MSE because it gives the average error in the same units as the target variable, which here is a knowledge score between 0 and 10. A low RMSE means that the model predictions are close to the actual values, and a high RMSE indicates that the model predictions deviate greatly from the actual values.

These indicators measure different aspects of model performance; the MAE, MSE, and RMSE measure the accuracy of predictions (the lower the error is, the better). In the case of tuberculosis, a good prediction model must minimize errors so that it can effectively identify cases of tuberculosis without producing too many false positives or false negatives.

The Mean Absolute Percentage Error (MAPE) measures the average percentage difference between predicted values and actual (true) values. It is commonly used in regression analysis to assess forecast accuracy or model prediction error in percentage terms.

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

where: n = total number of observations, y_i = actual value, $\hat{y}_i$ = predicted value. MAPE expresses prediction errors as a percentage of the actual values, which makes it easy to interpret across datasets and domains.

5.2. Hyperparameter Optimization Results, Training Durations and Cross-Validation

Hyperparameter tuning was carried out using the RandomizedSearchCV method, as stated in step g) Hyperparameter Tuning. This random search process over the possible hyperparameter combinations was applied to all models, except the Dummy Regressor, for which only the baseline strategy was adjusted. **Table 3** summarizes the training times and the best hyperparameter settings for each model.

Table 3. Hyperparameters and training times.

Model	Training Time (s)	Best Hyperparameters
Linear Regression	0.0018	{'fit_intercept': False}
Random Forest	0.248	{'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200}
SVM Regressor	0.011	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}
Artificial Neural Network	0.407	{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50,)}

AANN and Random Forest provide higher complexity and accuracy but require more training time. Linear and SVM models are more time-efficient but may perform worse in complex prediction tasks. This hyperparameter tuning confirms that model performance gains often come with a training time cost, especially for ensemble and deep learning models.

1) Fastest Training

a) Linear Regression trained the fastest (0.0018 s), reflecting its simplicity and low computational cost.

b) SVM also trained quickly (0.011 s), despite using a non-linear kernel (rbf), due to its optimization efficiency on smaller datasets.

2) Longest Training

a) ANN had the longest training time (0.407 s), which is typical due to the iterative learning and complexity of neural networks.

b) Random Forest took 0.248 s, which is relatively efficient given the number of estimators (200) and tree depth.

3) Model Complexity & Parameters

a) The Random Forest model benefited from fine-tuned parameters that control overfitting (max_depth, min_samples_split) and model stability (n_estimators).

b) The SVM Regressor used a radial basis function kernel with gamma = 'auto', suitable for capturing non-linear relationships.

c) The ANN model was configured with a single hidden layer of 50 neurons, ReLU activation, and L2 regularization (alpha = 0.0001).

5.3. Results and Interpretation

Quantitative approaches, such as machine learning models, are used to predict TB knowledge on the basis of multiple variables. These models can identify groups at risk of lack of knowledge and guide interventions more precisely.

Cross-validation was performed using the K-fold method, as indicated in step h) Training and Validation of the data processing methodology. The number of folds (k) used was 10, which is a standard practice to ensure a good balance between bias and variance, while maintaining sufficient data representativeness in each training and test subsample. This is consistent with the cross-validation results presented in the performance **Table 4**, which reflect the average performance

across the different folds.

Table 4. Performance with cross-validation.

Model	Best_R^2	Best_RMSE	Best_MAE	Best_MSE	Best_MAPE
Dummy Regressor	−0.000003	0.694385	0.554172	0.482171	9.973556
Linear Regression	0.560736	0.701756	0.555878	0.492462	8.964243
Random Forest	0.794882	0.597130	0.439553	0.356565	7.151338
SVM Regressor	0.574384	0.537905	0.394127	0.289342	6.929802
Artificial Neural Network	0.706541	0.606632	0.411178	0.368002	6.890633

The evaluation compared five models: Dummy Regressor, Linear Regression, Random Forest, Support Vector Machine (SVM) Regressor, and Artificial Neural Network (ANN), based on five performance metrics (R^2, RMSE, MAE, MSE, and MAPE).

- Random Forest achieved the highest explanatory power with an R^2 of 0.795, showing strong ability to model the variance in knowledge levels.
- SVM Regressor recorded the lowest prediction errors across RMSE (0.538), MAE (0.394), and MSE (0.289), indicating superior accuracy in absolute terms.
- ANN provided a strong balance between explanation and precision, with an R^2 of 0.707 and the best MAPE (6.89%), suggesting high reliability in percentage-based error.
- Linear Regression performed poorly, with higher error rates and low explanatory power compared to non-linear models.
- The Dummy Regressor, as expected, served as a weak baseline.

The best-performing models in this context are SVM Regressor and ANN, due to their lower error margins and robust predictive ability. Random Forest remains the top choice for interpretation and variance explanation. These results support a hybrid or ensemble strategy for enhanced accuracy and interpretability in public health knowledge modeling.

5.4. Feature Importance and Interpretability

To understand what drives the knowledge score, we combined two complementary perspectives. The Random Forest's global importance summarizes which variables contribute most to the model's predictive power, while SHAP values reveal not only how important each variable is overall but also the direction of its effect for individual observations. Both views are based on the same final model and the same preprocessing pipeline, ensuring consistency.

Figure 1 shows that Treatment Duration is by far the dominant driver of the model's predictions, with Town clearly in second position. Beyond these two, Age, Mild Illness, and Fasting Treatment contribute meaningfully, whereas Marital Status, Gender, Fatal Disease, Form of TB, and Heard about TB play smaller roles. This ranking reflects how much each feature helps the trees separate higher from

lower knowledge scores; it does not, however, indicate whether a higher value pushes the prediction up or down.

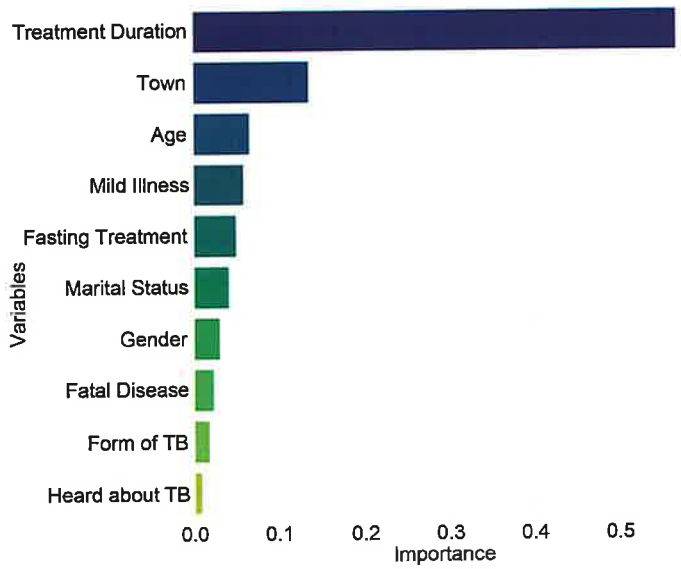


Figure 1. Random forest global feature importance.

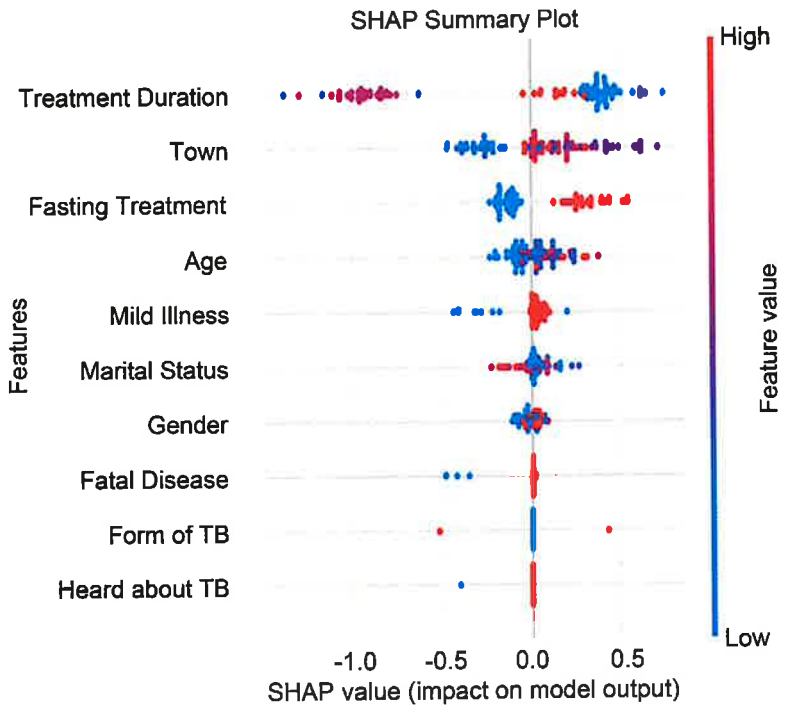


Figure 2. SHAP summary for the final model.

Figure 2 complements this picture by adding direction. Longer perceived Treatment Duration is associated with higher predicted knowledge, suggesting that respondents who recognize that treatment takes time tend to score better.

Geographic context also matters: the Town variable shows heterogeneous effects, consistent with local differences in awareness and access to information. Two misconceptions, considering tuberculosis a Mild Illness and believing treatment should be taken on an empty stomach, are linked to lower predicted knowledge, while Age exerts only a modest, non-monotonic influence. Taken together, the SHAP summary reinforces the Random Forest ranking and clarifies how these predictors shape the outcome.

5.5. Discussion of Error Distribution

Figure 3 presents histograms of the residuals (prediction errors = predicted value − actual value) for each model used to estimate knowledge levels on tuberculosis.

1) Dummy Regressor

a) The residuals are widely and irregularly spread around −1, with no clear central tendency.

b) This reflects the model's inability to learn from the data, acting as a random baseline.

c) Distribution is non-normal and asymmetric, with high variance.

2) Linear Regression

a) The residuals are centered around 0, showing a symmetric, bell-shaped distribution.

b) However, the spread is relatively large, indicating higher error variance.

c) Some outliers are visible on both ends, suggesting sensitivity to extreme values.

3) Random Forest

a) The residuals show a narrow and steep distribution around 0.

b) The shape is more peaked (leptokurtic), suggesting high precision and low bias.

c) The curve is well-centered and exhibits fewer extreme residuals compared to others.

4) SVM Regressor

a) The error distribution is tightly centered around zero and almost perfectly symmetrical.

b) It demonstrates the least spread and smoothest distribution, confirming the model's excellent generalization capacity.

c) Indicates minimal bias and consistent error performance across instances.

5) Artificial Neural Network (ANN)

a) The residuals are centered at 0 with a smooth, symmetrical distribution.

b) The spread is slightly broader than SVM but tighter than Linear Regression.

c) The ANN shows a balanced trade-off between accuracy and generalization, with relatively few extreme errors.

SVM Regressor and Random Forest display the most desirable residual patterns: tight, centered, and normally distributed, indicating high model reliability. ANN follows closely with a smooth and symmetrical distribution, slightly more

dispersed. Linear Regression and Dummy Regressor show less optimal residual behaviors, with wider spreads and signs of underfitting or oversimplification.

The residual distribution analysis reinforces the numerical evaluation:

- SVM and Random Forest offer the most consistent and accurate predictions.
- ANN is also strong, with minimal bias and decent error variance.
- Linear Regression fails to capture non-linearities, while Dummy Regressor confirms its role as a weak benchmark [39]-[41].

These visual diagnostics suggest that non-linear models with regularization and ensemble learning (SVM, Random Forest) are better suited for predicting complex patterns such as knowledge levels in public health datasets.

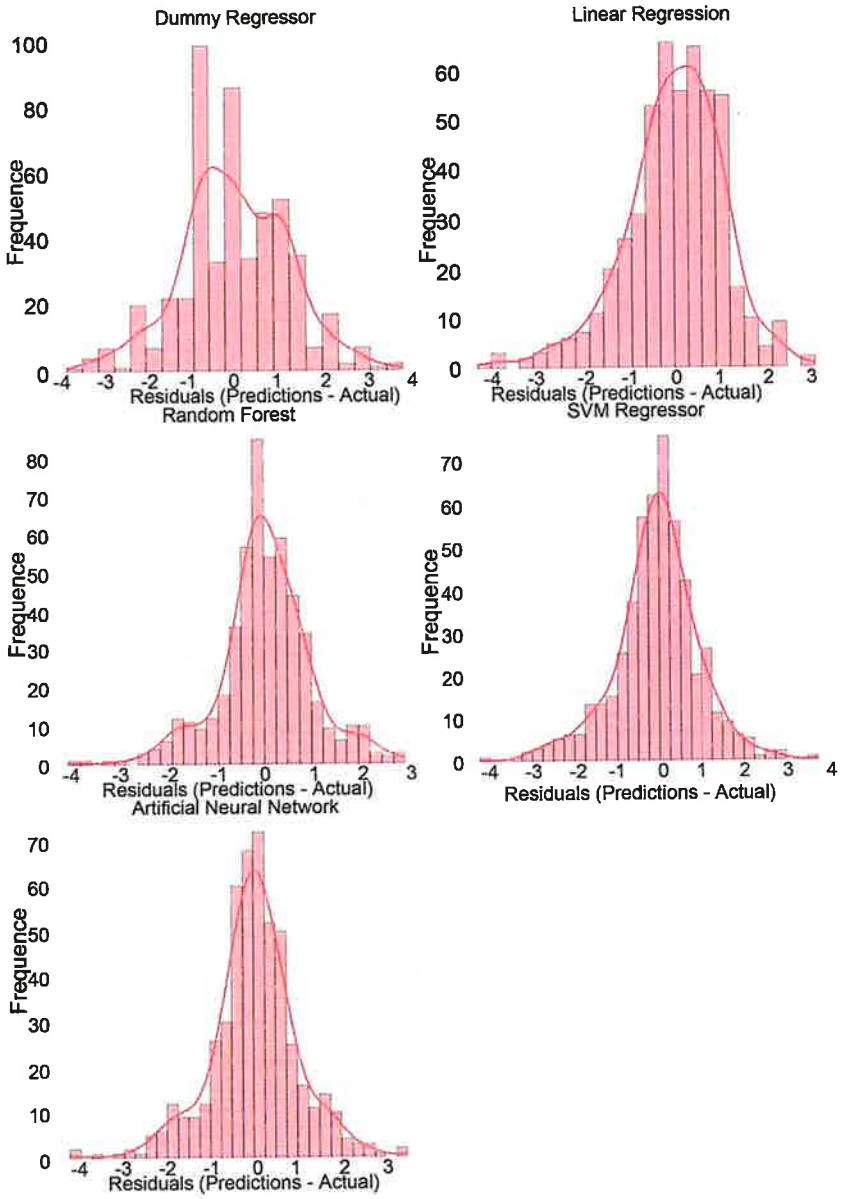


Figure 3. Error distribution.

5.6. Performance Visualization Results Based on the Metrics Charts, Heatmaps, Bar Plots, and Radar Charts

As shown in **Figure 4**, Random Forest attains the highest R^2 (0.795), indicating that it explains the largest share of variance, with ANN close behind (R^2 = 0.707). SVM Regressor and Linear Regression perform at moderate levels (≈0.57 and 0.56), while the Dummy model is near 0, as expected. For error magnitude, **Figure 4** also shows SVM achieving the lowest RMSE (0.538), with Random Forest and ANN remaining competitive (0.597 and 0.607). On absolute error metrics, **Figure 4** indicates that SVM leads on MAE (0.394) and MAPE (6.93%), with ANN just behind (MAE = 0.411; MAPE = 6.89%). Random Forest is moderate (MAE = 0.44; MAPE = 7.15%), whereas Linear Regression and the Dummy model exhibit the largest errors (MAE > 0.55; MAPE ≈ 9% - 10%).

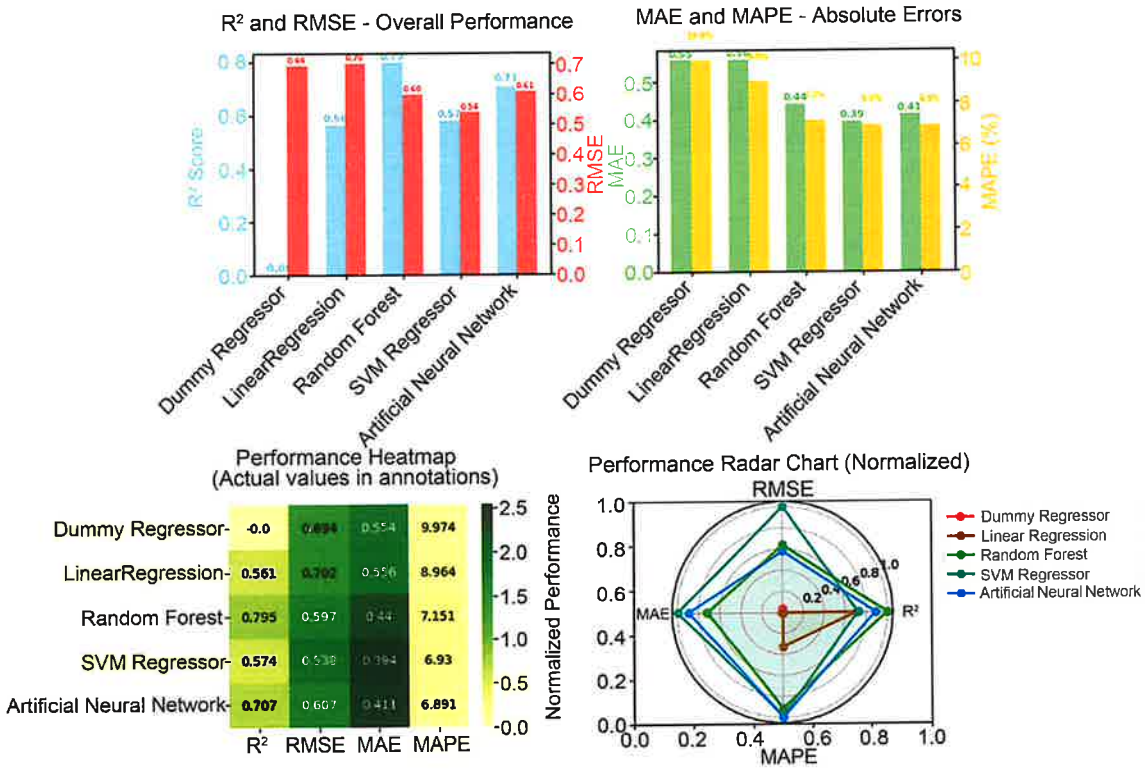


Figure 4. Performance visualizaion results—part I.

The radar chart in **Figure 4** highlights SVM's broad, balanced profile across normalized axes; ANN and Random Forest also perform strongly, while the Dummy and Linear models lag on error-related metrics. The heatmap and ranking panel in **Figure 5** summarize these patterns: SVM ranks first on three of four metrics (MAE, MAPE, RMSE), ANN ranks first on MAPE and second on two metrics, and Random Forest leads on R^2, suggesting strong explanatory power but not necessarily the lowest errors. The trade-off plot (R^2 vs MAPE) in **Figure 5** further illustrates that ANN offers the best balance between high R^2 and low

MAPE; Random Forest provides excellent variance explanation with slightly higher MAPE, and SVM minimizes errors while accepting a modestly lower R^2 than ANN.

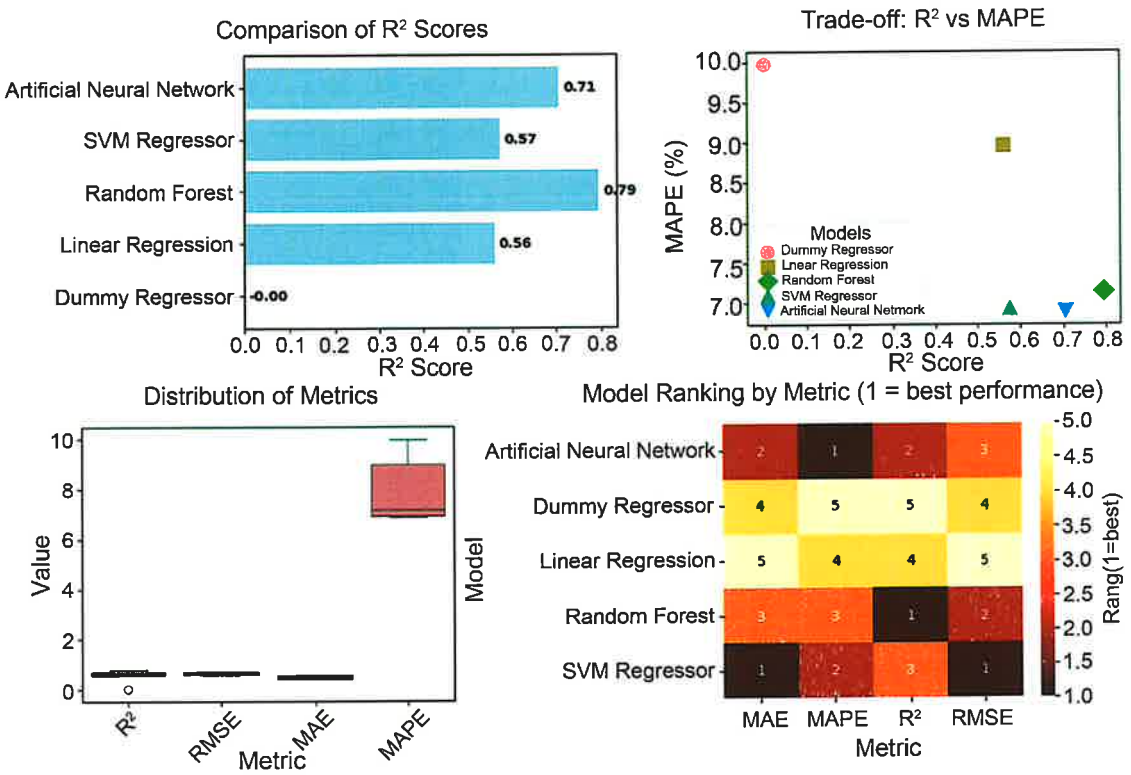


Figure 5. Performance visualizaion results—part II.

SVM Regressor delivers the most accurate predictions in terms of absolute error reduction, whereas ANN offers the best overall balance across metrics [42]. Random Forest is preferable when explanatory power and variance explanation are prioritized. Linear Regression and the Dummy model are not adequate for reliable prediction in this context. Depending on priorities (lowest errors vs. best balance), SVM or ANN is recommended; a stacking ensemble combining SVM and ANN could further enhance robustness.

5.7. Actual vs. Predicted Results

Figure 6 places the observed values against the model predictions to show how well each method tracks reality. The Dummy Regressor ($R^2 = -0.0032$) collapses into a flat band, which simply confirms that it does not learn from the data and performs worse than predicting the mean.

With Linear Regression ($R^2 = 0.2358$), **Figure 6** shows a gentle upward trend toward the 45° line, but the cloud of points spreads widely, signaling limited variance capture and sensitivity to outliers and heteroscedasticity. The Random Forest panel is noticeably tighter around the perfect-prediction line; its R^2 of 0.4104

reflects the strongest alignment between actual and predicted values among all candidates.

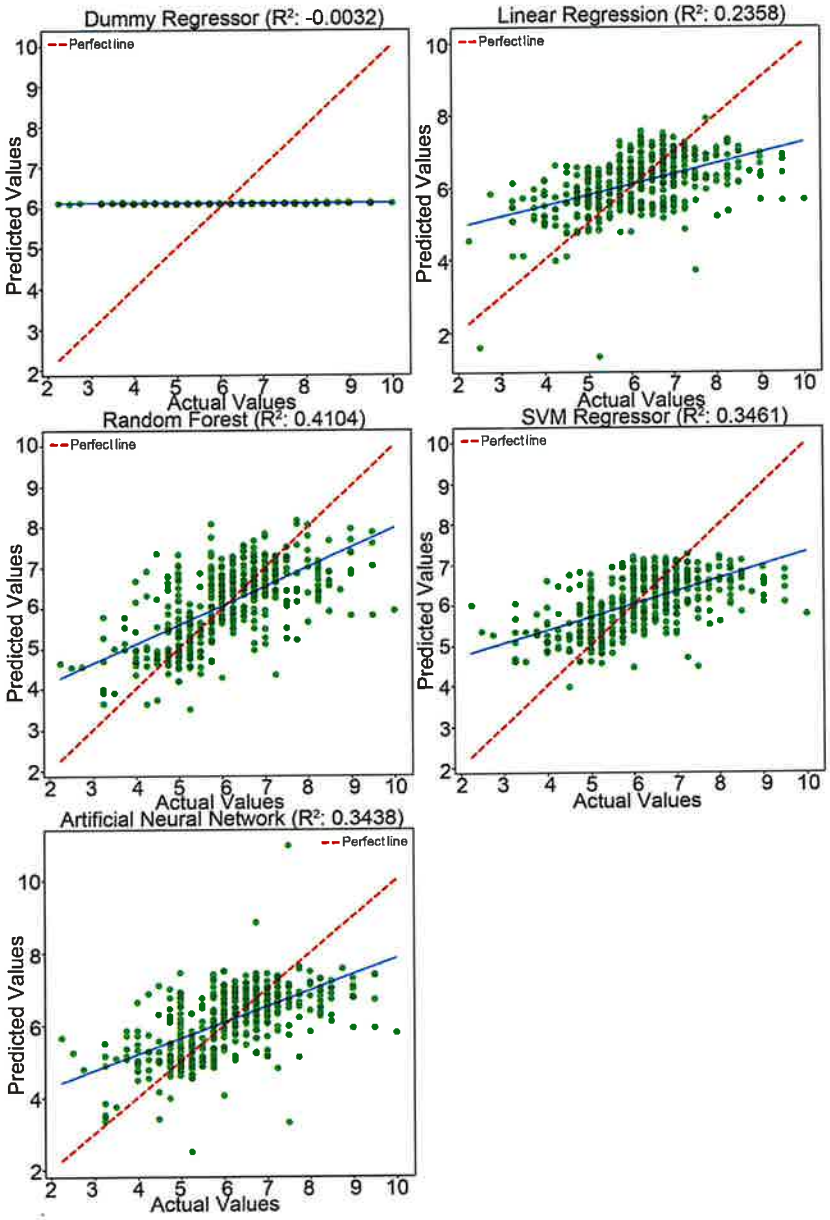


Figure 6. Actual vs. predicted results.

The SVM Regressor ($R^2 = 0.3461$) appears more structured than Linear Regression in **Figure 6**, yet a visible spread remains, indicating good generalization with slightly less fidelity than Random Forest. The Artificial Neural Network ($R^2 = 0.3438$) also pulls predictions closer to the reference line than the linear baseline, landing roughly on par with SVM and just behind Random Forest.

Overall, **Figure 6** supports a clear hierarchy: Random Forest offers the best balance of explanatory power and reliability, followed by SVM and ANN. Linear Re-

gression and the Dummy model lag behind, consistent with their limited ability to capture the non-linear patterns present in this task.

6. Discussion

This study introduced a numerical score to assess the level of knowledge of tuberculosis and used machine learning algorithms to validate and optimize this score. The results show that the numerical score allows a more accurate and granular assessment of knowledge than traditional methods based on categorical responses do. The machine learning algorithms used demonstrated a high ability to predict the level of knowledge on the basis of questionnaire responses, with accuracy.

Our study provides a methodological innovation by combining a numerical score with machine learning techniques to assess knowledge of tuberculosis. Unlike traditional methods, which are limited to qualitative or categorical assessments, our approach allows a quantitative and aggregated measurement of the level of knowledge. In addition, the use of machine learning algorithms made it possible to predict the level of knowledge on the basis of questionnaire responses via different indicators. This method can be adapted to other areas of public health where the overall assessment of knowledge is crucial.

Our results are consistent with those of previous studies that identified gaps in TB knowledge, particularly in populations with low education levels or those living in rural areas [43]. However, unlike existing studies that rely on qualitative methods or categorical questionnaires, our numerical score offers a more precise and reproducible measure. For example, a recent study reported that only 50% of 1,200 randomly surveyed Hainan University students knew specific prevention methods and that 60% believed that TB could be completely cured [44]. Our score, in contrast, allows for more nuanced distinctions between knowledge levels, identifying specific subgroups in need of tailored interventions. Several limitations should be considered. First, the numerical score relies on a self-report questionnaire, which may introduce social desirability bias. Second, although machine learning algorithms have shown high performance, their application requires quality data and technical expertise, which may limit their use in some contexts. Third, our study was conducted in a specific population, and the generalizability of the results to other contexts requires further validation.

To overcome these limitations, future research could explore the use of more advanced machine learning techniques, such as deep neural networks, to enable a more fine-grained analysis of the factors influencing the level of knowledge. Finally, longitudinal studies are needed to assess the impact of TB awareness on the evolution of the numerical score over time.

7. Conclusions

Prediction plays a crucial role in the management of tuberculosis, facilitating early detection, reducing transmission, improving treatments, and contributing to a more effective health response to this disease. Predicting TB could reveal how the

disease is managed globally. For these changes to occur, it is necessary to invest in data collection and analysis technologies, strengthen cooperation between public and private health institutions, and build resilient health systems capable of responding quickly to the challenges posed by the disease. This change, although complex, could have a major impact on reducing the incidence of tuberculosis by increasing the rate of early detection, improving treatments and limiting the spread of the disease.

This study made it possible to assess the knowledge levels of the population of Niamey on tuberculosis. It constitutes a basic tool on which subsequent studies can be based with the aim of improving the levels of knowledge of a population about tuberculosis.

This study successfully demonstrates the potential of supervised machine learning algorithms to predict individuals' level of knowledge about tuberculosis based on survey data. By comparing multiple regression models including Linear Regression, Support Vector Machine (SVM), Random Forest, and Artificial Neural Networks (ANN), the analysis highlights the effectiveness of non-linear approaches in capturing complex patterns within the data.

Among all tested models, Random Forest achieved the best performance in terms of variance explanation ($R^2 = 0.795$), while the SVM Regressor and ANN stood out for their low prediction errors, with MAPE values below 7%. These results confirm the added value of advanced algorithms in enhancing the predictive accuracy of public health assessments. A value of $k = 10$ was used for stratified cross-validation.

Hyperparameter tuning was conducted via RandomizedSearchCV, only on the training set. No data leakage occurred: the data split, step sequencing, and final evaluation all adhered to best practices in predictive modeling.

Moreover, the use of SHAP interpretability techniques made it possible to identify the most influential features contributing to knowledge prediction, ensuring transparency and facilitating actionable insights for public health stakeholders.

In light of these findings, this work paves the way for data-driven decision-making in tuberculosis education and awareness campaigns; personalized public health strategies, targeting populations with the lowest predicted knowledge levels; scalable AI-based screening tools in epidemiological surveys.

Further research could enrich the model by integrating behavioral, spatial, or clinical data, and by deploying real-time applications in community-based health programs. Ultimately, this study underscores the promise of explainable artificial intelligence in promoting precision public health, especially in resource-limited settings where tuberculosis remains a major concern.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (WHO) (2024) Operational Handbook on Tuberculosis

(Module 1—Prevention): Tuberculosis Preventive Treatment. Global Programme on Tuberculosis and Lung Health (GTB), Second Edition.

[2] WHO (2024) Global Tuberculosis Report 2024. Global Tuberculosis Programme (GTB).

[3] PNLT (2016) Evaluation of the National Tuberculosis Control Program. INS.

[4] INS (2016) Demographic Projections for Niger.

[5] Kigozi, N.G., Heunis, J.C., Engelbrecht, M.C., Janse van Rensburg, A.P. and van Rensburg, H.C.J.D. (2017) Tuberculosis Knowledge, Attitudes and Practices of Patients at Primary Health Care Facilities in a South African Metropolitan: Research Towards Improved Health Education. *BMC Public Health*, **17**, Article No. 795. https://doi.org/10.1186/s12889-017-4825-3

[6] Laycock, K.M., Enane, L.A. and Steenhoff, A.P. (2021) Tuberculosis in Adolescents and Young Adults: Emerging Data on TB Transmission and Prevention among Vulnerable Young People. *Tropical Medicine and Infectious Disease*, **6**, Article 148. https://doi.org/10.3390/tropicalmed6030148

[7] Költringer, F.A., Annerstedt, K.S., Boccia, D., Carter, D.J. and Rudgard, W.E. (2023) The Social Determinants of National Tuberculosis Incidence Rates in 116 Countries: A Longitudinal Ecological Study between 2005-2015. *BMC Public Health*, **23**, Article No. 337. https://doi.org/10.1186/s12889-023-15213-w

[8] Ma, E., Ren, L., Wang, W., Takahashi, H., Wagatsuma, Y., Ren, Y., *et al.* (2015) Demographic and Socioeconomic Disparity in Knowledge about Tuberculosis in Inner Mongolia, China. *Journal of Epidemiology*, **25**, 312-320. https://doi.org/10.2188/jea.je20140033

[9] Banu, S., Rahman, M.T., Uddin, M.K.M., Khatun, R., Ahmed, T., Rahman, M.M., *et al.* (2013) Epidemiology of Tuberculosis in an Urban Slum of Dhaka City, Bangladesh. *PLOS ONE*, **8**, e77721. https://doi.org/10.1371/journal.pone.0077721

[10] Brown, P.J. and Inhorn, M.C. (1998) The Anthropology of Infectious Disease: International Health Perspectives. Routledge.

[11] Macdonald, H. and Harper, I. (2019) Understanding Tuberculosis and its Control, Anthropological and Ethnographic Approaches. Routledge.

[12] World Health Organization (2014) Global Tuberculosis Report 2014. Global Tuberculosis Programme (GTB).

[13] Ashaba, C., Musoke, D., Tsebeni Wafula, S. and Konde-Lule, J. (2021) Stigma among Tuberculosis Patients and Associated Factors in Urban Slum Populations in Uganda. *African Health Sciences*, **21**, 1640-1650. https://doi.org/10.4314/ahs.v21i4.18

[14] Sharma, S.K. and Mohan, A. (2006) Multidrug-Resistant Tuberculosis: A Menace That Threatens to Destabilize Tuberculosis Control. *Chest*, **130**, 261-272. https://doi.org/10.1016/s0012-3692(15)50981-1

[15] Erving, G. (1964) Stigma: Notes on the Management of Spoiled Identity. *Social Forces*, **43**, 127-128.

[16] World Health Organization (WHO) (2015) Global Tuberculosis Report 2015: Global Tuberculosis Programme (GTB). 20th Edition, WHO.

[17] Onyeonoro, U.U., Chukwu, J.N., Oshi, D.C., Nwafor, C.C. and Meka, A.O. (2014) Assessment of Tuberculosis-Related Knowledge, Attitudes and Practices in Enugu, South East Nigeria. *Journal of Infectious Diseases and Immunity*, **6**, 1-9. https://doi.org/10.5897/jidi2011.0020

[18] Ait Mouhoub, W. (2020) Profil épidémiologique de la tuberculose dans une wilaya de l'est d'Algérie. *Revue des Maladies Respiratoires Actualités*, **12**, 268.

https://doi.org/10.1016/j.rmra.2019.11.609

[19] World Health Organization (WHO) (2019) Global Tuberculosis Report 2019. Global Tuberculosis Programme.

[20] Smith, J.P., Milligan, K., McCarthy, K.D., Mchembere, W., Okeyo, E., Musau, S.K., *et al.* (2023) Machine Learning to Predict Bacteriologic Confirmation of Mycobacterium Tuberculosis in Infants and Very Young Children. *PLOS Digital Health*, **2**, e0000249. https://doi.org/10.1371/journal.pdig.0000249

[21] Khan, M.T., Kaushik, A.C., Ji, L., Malik, S.I., Ali, S. and Wei, D. (2019) Artificial Neural Networks for Prediction of Tuberculosis Disease. *Frontiers in Microbiology*, **10**, Article 395. https://doi.org/10.3389/fmicb.2019.00395

[22] Harris, M., Qi, A., Jeagal, L., Torabi, N., Menzies, D., Korobitsyn, A., *et al.* (2019) A Systematic Review of the Diagnostic Accuracy of Artificial Intelligence-Based Computer Programs to Analyze Chest X-Rays for Pulmonary Tuberculosis. *PLOS ONE*, **14**, e0221339. https://doi.org/10.1371/journal.pone.0221339

[23] Mohidem, N.A., Osman, M., Muharam, F.M., Elias, S.M., Shaharudin, R. and Hashim, Z. (2021) Prediction of Tuberculosis Cases Based on Sociodemographic and Environmental Factors in Gombak, Selangor, Malaysia: A Comparative Assessment of Multiple Linear Regression and Artificial Neural Net-Work Models. *The International Journal of Mycobacteriology*, **10**, 442-456. https://doi.org/10.4103/ijmy.ijmy_182_21

[24] Srivastava, V., Kumar, R., Wani, M.Y., Robinson, K. and Ahmad, A. (2024) Role of Artificial Intelligence in Early Diagnosis and Treatment of Infectious Diseases. *Infectious Diseases*, **57**, 1-26. https://doi.org/10.1080/23744235.2024.2425712

[25] Ahmed, K.M., Chandra Das, B., Saadati, Y. and Amini, M.H. (2024) A Comprehensive Review of Artificial Intelligence and Machine Learning Methods for Modern Healthcare Systems. In: Amini, M.H., Ed., *Distributed Machine Learning and Computing*, Springer, 71-110. https://doi.org/10.1007/978-3-031-57567-9_4

[26] Singh, M., Pujar, G.V., Kumar, S.A., Bhagyalalitha, M., Akshatha, H.S., Abuhaija, B., *et al.* (2022) Evolution of Machine Learning in Tuberculosis Diagnosis: A Review of Deep Learning-Based Medical Applications. *Electronics*, **11**, Article 2634. https://doi.org/10.3390/electronics11172634

[27] Hansun, S., Argha, A., Bakhshayeshi, I., Wicaksana, A., Alinejad-Rokny, H., Fox, G.J., *et al.* (2025) Diagnostic Performance of Artificial Intelligence-Based Methods for Tuberculosis Detection: Systematic Review. *Journal of Medical Internet Research*, **27**, e69068. https://doi.org/10.2196/69068

[28] Althomsons, S.P., Winglee, K., Heilig, C.M., Talarico, S., Silk, B., Wortham, J., *et al.* (2022) Using Machine Learning Techniques and National Tuberculosis Surveillance Data to Predict Excess Growth in Genotyped Tuberculosis Clusters. *American Journal of Epidemiology*, **191**, 1936-1943. https://doi.org/10.1093/aje/kwac117

[29] Tang, N., Yuan, M., Chen, Z., Ma, J., Sun, R., Yang, Y., *et al.* (2023) Machine Learning Prediction Model of Tuberculosis Incidence Based on Meteorological Factors and Air Pollutants. *International Journal of Environmental Research and Public Health*, **20**, Article 3910. https://doi.org/10.3390/ijerph20053910

[30] Maharani, R., Karima, U.Q. and Kamilia, K. (2022) Socio-Demographic and Behavioral Factors Relationship with Pulmonary Tuberculosis: A Case-Control Study. *Open Access Macedonian Journal of Medical Sciences*, **10**, 130-135. https://doi.org/10.3889/oamjms.2022.8157

[31] Dorji, T., Tshering, T. and Wangdi, K. (2020) Assessment of Knowledge, Attitude and Practice on Tuberculosis among Teacher Trainees of Samtse College of Educa-

tion, Bhutan. *PLOS ONE*, **15**, e0241923. https://doi.org/10.1371/journal.pone.0241923

[32] Manoharan, A., Chellaiyan, V.G., M., J. and Liaquathali, F. (2019) Impact of Educational Intervention on the Tuberculosis Knowledge among the Medical Students, Chennai. *International Journal of Community Medicine and Public Health*, **6**, 5317-5320. https://doi.org/10.18203/2394-6040.ijcmph20195491

[33] Balakrishnan, V., Ramanathan, G., Zhou, S. and Wong, C.K. (2023) Optimized Support Vector Regression Predicting Treatment Duration among Tuberculosis Patients in Malaysia. *Multimedia Tools and Applications*, **83**, 11831-11844. https://doi.org/10.1007/s11042-023-16028-y

[34] Khan, M.T., Kaushik, A.C., Ji, L., Malik, S.I., Ali, S. and Wei, D. (2019) Artificial Neural Networks for Prediction of Tuberculosis Disease. *Frontiers in Microbiology*, **10**, Article 395. https://doi.org/10.3389/fmicb.2019.00395

[35] El-Solh, A.A., Hsiao, C., Goodnough, S., Serghani, J. and Grant, B.J.B. (1999) Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network. *Chest*, **116**, 968-973. https://doi.org/10.1378/chest.116.4.968

[36] Gichuhi, H.W., Magumba, M., Kumar, M. and Mayega, R.W. (2023) A Machine Learning Approach to Explore Individual Risk Factors for Tuberculosis Treatment Non-Adherence in Mukono District. *PLOS Global Public Health*, **3**, e0001466. https://doi.org/10.1371/journal.pgph.0001466

[37] Ahamed Fayaz, S., Babu, L., Paridayal, L., Vasantha, M., Paramasivam, P., Sundarakumar, K., *et al.* (2024) Machine Learning Algorithms to Predict Treatment Success for Patients with Pulmonary Tuberculosis. *PLOS ONE*, **19**, e0309151. https://doi.org/10.1371/journal.pone.0309151

[38] Jonathan, J., Barakabitze, A.A., Fast, C.D. and Cox, C. (2024) Machine Learning for Prediction of Tuberculosis Detection: Case Study of Trained African Giant Pouched Rats. *Online Journal of Public Health Informatics*, **16**, e50771. https://doi.org/10.2196/50771

[39] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[40] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. *R News*, **2**, 18-22.

[41] Cutler, A., Cutler, D.R. and Stevens, J.R. (2012) Random Forests. In: Zhang, C. and Ma, Y., Eds., *Ensemble Machine Learning*, Springer, 157-175. https://doi.org/10.1007/978-1-4419-9326-7_5

[42] Basheer, I.A. and Hajmeer, M. (2000) Artificial Neural Networks: Fundamentals, Computing, Design, and Application. *Journal of Microbiological Methods*, **43**, 3-31. https://doi.org/10.1016/s0167-7012(00)00201-3

[43] Anguyo, R., Mukama, S., Bindeeba, D., Senyimba, C., Ezajobo, S., Nakawesi, J., *et al.* (2025) Knowledge of Tuberculosis Prevention across Eight Districts in Central Uganda: An Analysis of Lot Quality Assurance Sampling Survey Data. *Risk Management and Healthcare Policy*, **18**, 719-738. https://doi.org/10.2147/rmhp.s494335

[44] Xie, H., Wang, W., Chen, X., Huang, D., Yu, Q. and Luo, L. (2025) An Analysis of Knowledge, Attitudes, Practice and Influencing Factors for Tuberculosis Prevention and Control among Hainan University Students. *Frontiers in Public Health*, **13**, Article 1478251. https://doi.org/10.3389/fpubh.2025.1478251