Journal of Theoretical Biology 597 (2025) 111988

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A comprehensive study on tuberculosis prediction models: Integrating machine learning into epidemiological analysis

Hamna Mariyam K.B. [a], Sayooj Aby Jose [a,b,*], Anuwat Jirawattanapanit [b], Karuna Mathew [c]

[a] School of Data Analytics, Mahatma Gandhi University, Kottayam, India
[b] Department of Mathematics, Faculty of Education, Phuket Rajabhat University, Phuket, Thailand
[c] Faculty of Engineering Environment and Computing, Coventry University, Coventry, United Kingdom

ARTICLE INFO

Keywords:
Tuberculosis
Predictive modeling
Machine learning models
TB incidence forecasting

ABSTRACT

Tuberculosis (TB), the second leading infectious killer globally, claimed the lives of 1.3 million individuals in 2022, after COVID-19, surpassing the toll of HIV and AIDS. With an estimated 10.6 million new TB cases worldwide in 2022, the gravity of the disease persists, necessitating urgent attention. Tuberculosis remains a critical public health crisis, and efforts to combat this infectious disease demand intensified global commitment and resources. This study utilizes predictive modeling techniques to forecast the incidence of Tuberculosis (TB), employing a range of machine learning models. Additionally, the research incorporates impactful visualizations for comprehensive data exploration, analysis and comparison. Various machine learning models are developed to anticipate TB incidence, with the optimal performing model to customize a user-defined function. This research provides valuable insights into the potential determinants influencing TB incidence, contributing to the identification of strategies for preventing the spread of Tuberculosis.

1. Introduction

Mathematical modeling involves applying mathematical concepts to address complex, real-world problems that lack clear structures (Peter and Clatworthy, 1990). In these modeling processes, mathematical methodologies are employed to derive solutions for practical issues. Real-world problems are translated into mathematical formulations and subsequently addressed through the application of mathematical techniques (Ang, 2001). It incorporates the processes of revealing the relationships, conducting mathematical analyses, obtaining results and reinterpreting the model (Dndar et al., 2012). Mathematical models play a crucial role across various domains, including epidemiological research. Initially, they aid in refining study inquiries by visually representing intricate systems, guiding literature searches, and pinpointing crucial variables. During the study design phase, models assist in testing sampling strategies, estimating sample size and power, and predicting outcomes for studies that may be impractical due to time or ethical constraints. Post data collection, models facilitate result interpretation, exploration of causal pathways, and integrated analysis of data from multiple sources. Furthermore, in applying research findings to public health practice, models are instrumental in estimating population risk, predicting intervention effects, and contributing to the ongoing program evaluation. The potential of mathematical modeling to significantly enhance epidemiology lies in its capacity to streamline the research process, serve as a communication tool for policymakers, and foster interdisciplinary collaboration (Chubb and Jacobsen, 2010). Numerous scholars employ mathematical modeling as a pivotal tool to tackle practical challenges in their research endeavors (Jose et al., 2023a, 2022, 2023b).

With the introduction of systems based on computers, the digitalization of all medical records and the evaluation of clinical data in healthcare systems have become widespread routine practices. Daily, healthcare services produce an enormous amount of data, making it increasingly complicated to analyze and handle it in “conventional ways”. Using machine learning and deep learning, this data may be properly analyzed to generate actionable insights (Badawy et al., 2023). In the present investigation, machine learning models have been employed for analysis. Machine learning systems represent innovative methodologies for mathematical modeling, grounded in the utilization of differential equations. This approach possesses a distinct advantage in enabling the development of applications capable of dynamically adjusting to evolving environments. Machine Learning (ML) is the fastest rising arena in computer science, and health informatics is of extreme challenge. The aim of Machine Learning is to develop algorithms which can learn and progress over time and can be used for predictions. Machine Learning practices are widely used in various fields and primarily health care industry has been benefitted a lot through machine

[*] Corresponding author at: Department of Mathematics, Faculty of Education, Phuket Rajabhat University, Phuket, Thailand.
E-mail address: sayooaby999@gmail.com (S.A. Jose).

https://doi.org/10.1016/j.jtbi.2024.111988
Received 15 January 2024; Received in revised form 28 October 2024; Accepted 1 November 2024
Available online 13 November 2024

learning prediction techniques (Nithya and Ilango, 2017). Machine Learning consists of training systems capable of understanding the data entered in order to predict responses or extract useful information from them. It is a subset of artificial intelligence and is closely related to statistics (Bokonda et al., 2020).

Tuberculosis (TB) stands as one of humanity's oldest diseases, with molecular evidence tracing its existence back over 17,000 years. Despite the introduction of modern approaches to diagnose and treat TB, regrettably, individuals continue to endure its impact, contributing to its status as one of the top 10 fatal infectious diseases globally (Sandhu, 2011). Tuberculosis (TB) is attributed to the bacterium *Mycobacterium tuberculosis*, and transmission occurs when individuals afflicted with TB release bacteria into the air, such as through coughing. Approximately 25% of the worldwide population is believed to have experienced TB infection (Nalunjogi et al., 2023; World Health Organization, 0000b). Tuberculosis (TB) is a preventable and usually curable disease. Yet in 2022, TB was the world's second leading cause of death from a single infectious agent, after coronavirus disease (COVID-19), and caused almost twice as many deaths as HIV/AIDS. More than 10 million people continue to fall ill with TB every year (World Health Organization, 0000b). Although it can affect people of any age, individuals with weakened immune systems, e.g., with HIV infection, are at increased risk. Since the immune system in healthy people walls off the causative bacteria, TB infection in healthy people is often asymptomatic. This bacterium lives and multiplies in the macrophages, thus avoiding the natural defense system in the patient's serum. Infection with TB can result in two stages: asymptomatic latent tuberculosis infection (LTBI) or tuberculosis disease. If left untreated, the mortality rate with this disease is over 50% (Abdualgalil et al., 2022).

Several investigators utilized mathematical and machine learning models to evaluate the gravity of Tuberculosis, acknowledging its substantial implications for public health. A notable contribution to the understanding of tuberculosis incidence is conducted in a comprehensive study titled 'Machine Learning Prediction Model of Tuberculosis Incidence Based on Meteorological Factors and Air Pollutants' (Tang et al., 2023). Alvaro David Orjuela-Cañón implemented machine learning within the diagnostic support framework for tuberculosis, supplemented by data that can function as an alternative diagnostic tool through data processing, particularly in regions with constrained health infrastructure (Orjuela-Canñón Alvaro et al., 2022). In Tiwari and Maji (2019), the author has provided an extensive examination of the array of machine learning techniques utilized by researchers in the study of tuberculosis disease (Jose et al., 2024). The authors introduced an optimized machine learning model, which extracts optimal texture features from images related to tuberculosis, and they determined the hyperparameters of the classifiers in Hrizi et al. (2022) for tuberculosis diagnosis. In Abdualgalil et al. (2022), a comparative study on modeling and forecasting tuberculosis cases using machine learning and deep learning approaches was implemented. The study forecasted occurrences of pulmonary negative, positive, and overall TB incidence cases spanning the years 2020 to 2029, while also offering insights into the spread of Tuberculosis in Yemen. The rise of antibiotic-resistant *Mycobacterium tuberculosis* strains poses a significant threat to global tuberculosis control efforts, as it renders current treatments less effective. The increasing prevalence of drug-resistant TB has led to higher mortality rates, longer treatment durations, and increased healthcare costs, underscoring the urgent need of innovative solutions (Shamil et al., 2014). Our research methodology involves a comparative study utilizing diverse machine learning models to assess potential influential factors and propose recommendations for mitigating tuberculosis incidence, drawing on global data. The prediction models of TB incidence can help mitigate the impact of antibiotic resistance by identifying high-risk populations and areas, enabling targeted interventions, and optimizing resource allocation for preventive measures and novel treatment strategies (World Health Organization, 0000a; Anggriani et al., 2023).

Main contributions of this paper as follows:

- This research is centered on the analysis of authentic non-temporal data, with a primary emphasis on constructing diverse models for predicting tuberculosis incidence.
- The paper introduces impactful visualizations aimed at facilitating exploration, analysis, and comparison of results derived from predictive modeling of tuberculosis incidence.
- Various machine learning techniques were explored in depth for the purpose of predictive modeling within the context of tuberculosis incidence.
- A comprehensive comparative study was conducted to identify and highlight the best-performing model in this research, accompanied by the presentation of its predicted values.
- The paper proposes a user-defined function designed to offer practical suggestions for reducing the incidence of tuberculosis, based on insights derived from the selected machine learning model.

2. Methods

2.1. Data collection

In this investigation, data were obtained from the World Health Organization (WHO) official website, specifically from the World Tuberculosis Programme (World Health Organization, 0000). The dataset encompasses 18 CSV files, consisting of four files containing WHO TB burden estimates and the remaining files containing data submitted by various countries and territories to WHO. The dataset comprises a total of 637 variables. To identify suitable independent variables and the specific dependent variable, a comprehensive exploration of these variables was conducted using Microsoft Excel.

2.2. Data preprocessing

A new dataset, denoted as the "Input Dataset", was constructed in Excel, consolidating relevant columns such as Country, Year, and the values of both independent and dependent variables. Given the presence of numerous missing values in the dataset, a pragmatic decision was made to focus on the year 2022, as it exhibited a comparatively higher number of data entries than other years. The Input Dataset, thus derived, served as the foundation for subsequent machine learning analyses, enabling the impact of independent variables on the targeted outcome.

In the 'Input Dataset', certain countries exhibit missing values. To address this, mean imputation was employed to estimate the missing values of the respective variables. This imputation method involved utilizing the mean values of the variables from preceding years for the corresponding countries. After importing the 'Input Dataset' into Python, a thorough examination was conducted to identify and address potential issues, including missing values and duplicate entries. Appropriate measures were taken to resolve these data quality concerns.

2.3. Exploratory data analysis

The chosen dependent variable for prediction is 'New_Cases', representing the count of newly diagnosed cases of bacteriologically confirmed pulmonary tuberculosis. Bacteriologically confirmed cases involve laboratory tests that confirm the presence of *Mycobacterium tuberculosis* in clinical samples obtained from patients. This choice of the dependent variable stems from its recognized specificity and reliability compared to diagnoses solely based on clinical or radiological assessments. The independent variables considered in the analysis include 'hh_Contacts', representing the estimated number of household contacts of individuals newly diagnosed with bacteriologically confirmed pulmonary TB; 'Prev_Cases', representing the count of previously treated

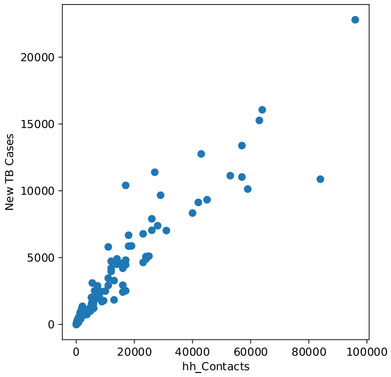

Fig. 1. Impact of independent variables.

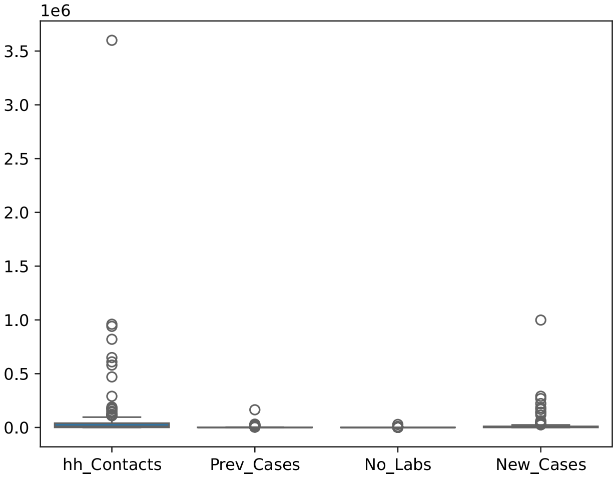

Fig. 2. Boxplot for detecting outliers.

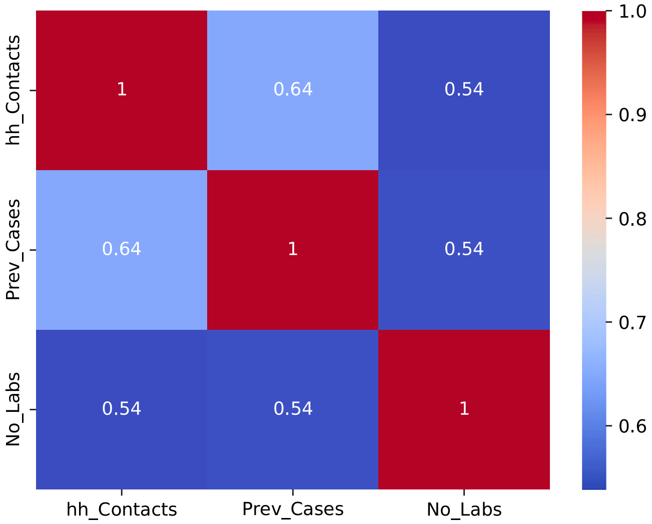

Fig. 3. Correlation matrix for detecting multicollinearity.

bacteriologically confirmed pulmonary TB cases; and ‘No_Labs’, representing the total number of sites providing laboratory diagnostic testing for TB at the end of the reporting year 2022.

The rationale behind choosing these independent variables is based on their direct impact on the transmission dynamics and detection capacity of TB. ‘hh_Contacts’ is crucial because household contacts are at a higher risk of contracting TB from infected individuals, making it a significant predictor of new cases. ‘Prev_Cases’ provides insight into the recurrence and ongoing transmission within the population, as previously treated cases can contribute to the spread of the disease if not adequately managed. ‘No_Labs’ reflects the diagnostic capacity and accessibility of TB testing, which directly influences the detection rates of new cases. These variables collectively capture important aspects of TB transmission, recurrence, and detection, making them essential for accurately predicting new TB cases. Fig. 1 illustrates the impact of individual independent variables on the dependent variable using a scatter plot.

The analysis of Fig. 1 reveals a robust positive correlation between ‘hh_Contacts’ and ‘New_Cases’, surpassing the correlation observed with other independent variables such as ‘Prev_Cases’ and ‘No_Labs’. Specifically, as the magnitude of the independent variables increases, there is a corresponding tendency for the dependent variable to exhibit an increase as well. Fig. 2 illustrates the identification of outliers through the visualization of a boxplot.

Upon scrutinizing the findings depicted in Fig. 2, it becomes apparent that the majority of predicted values closely align with the actual values, with the exception of some data points, which can be identified as outliers. The presence of these outliers has the potential to exert an influence on the accuracy of predictions. To address this issue, outliers were eliminated from the dataset through the application of the Interquartile Range (IQR) Method. The IQR method is renowned for its robust statistical approach and is employed to detect and exclude outliers in the dataset. This method is particularly effective in identifying extreme values that deviate significantly from the central tendency of the data, which can otherwise skew the results and reduce the accuracy of predictive models. By removing these outliers, the IQR method enhances the reliability of the data analysis process by mitigating the influence of extreme values on statistical measures. The rationale behind using the IQR method lies in its ability to improve model performance. Outliers can distort parameter estimates and lead to overfitting, where the model learns noise instead of the underlying pattern. By excluding outliers, we aim to achieve a more generalized model that better captures the true relationship between the variables, thereby improving prediction accuracy. This process involves calculating the IQR (the range between the first and third quartiles) and removing data points that fall below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR. This criterion is widely accepted in statistical analysis for its balance between sensitivity and specificity in outlier detection. Thus, the application of the IQR method not only helps in refining the dataset but also plays a crucial role in enhancing the overall accuracy and robustness of the predictive models used in our study. Fig. 3 displays a correlation matrix, employed for the assessment of multicollinearity within the dataset.

The examination of Fig. 3 suggests the existence of moderate multicollinearity; however, caution should be exercised in drawing definitive

Table 1
VIF results.

Variable	VIF
hh_Contacts	2.685449
Prev_Cases	2.655770
No_Labs	2.281032

conclusions based solely on this correlation assessment. To provide a conclusive evaluation of multicollinearity, the Variance Inflation Factor (VIF) for the independent variables was computed, and the corresponding values are presented in Table 1.

All independent variables exhibit VIF values below 10, indicating the absence of substantial multicollinearity concerns within this context. Therefore, it can be concluded that multicollinearity is not prevalent in the dataset under examination.

3. Performing model selection: Application of six machine learning techniques

In this section, we developed multiple machine learning models using a dataset comprising 152 instances. The dataset was divided into 70% for training purposes and 30% for testing. This investigation utilized a diverse set of machine learning techniques, including K-Nearest Neighbors, Quantile Regression, Random Forest, XGBoost, LightGBM, and CatBoost. Various libraries such as numpy, pandas, matplotlib, scikit-learn, and seaborn in Python were employed for the analysis. To ensure that the features contribute equally to the model's performance, the variables in both the training and test sets were standardized using the StandardScaler. This process involved calculating the mean and standard deviation for each feature from the training set and then applying this scaling to both the training and test sets to ensure they have a mean of 0 and a standard deviation of 1. Prior to model training, an exploratory data analysis (EDA) was conducted to understand the underlying properties of the data, including checking for homoskedasticity versus heteroskedasticity, which are fundamental assumptions in regression modeling. Homoskedasticity implies constant variance of the residuals, a key assumption for linear models, while heteroskedasticity indicates varying variance, which can affect model performance and inference. In the context of Quantile Regression, this study investigated the presence of heteroskedasticity using the Breusch–Pagan test. The results indicated the presence of heteroskedasticity, which quantile regression can handle effectively as it does not rely on the homoskedasticity assumption. For ensemble methods like Random Forest, XGBoost, LightGBM, and CatBoost, the assumption of homoskedasticity is not a strict requirement. These models are robust to heteroskedasticity and can handle complex, non-linear relationships in the data. During the training stage, the model learns the relationships between the input features and the target variable by adjusting its parameters to minimize the difference between its predictions and the actual target values. Once the model is trained, we use the test set to evaluate its performance. The model's predictions on the test set are compared to the actual values, and various metrics (such as Mean Squared Error or R-squared) are used to assess how well the model generalizes to new, unseen data.

3.1. K-nearest neighbors model

K-nearest neighbors (KNN) is a supervised machine learning method that can be utilized for both classification and regression tasks (Ozturk Kiyak et al., 2023). Fig. 4 illustrates the cross-validation technique employed to determine the optimal value of k. The analysis of Fig. 4 indicates that the optimal value of k is 1.

Fig. 5 presents a multiline chart depicting the fluctuations in observed and predicted values of TB incidence using the K-Nearest Neighbors model on the test data. The visual representation in Fig. 5 indicates that, at certain instances, the predicted values closely correspond to the actual values.

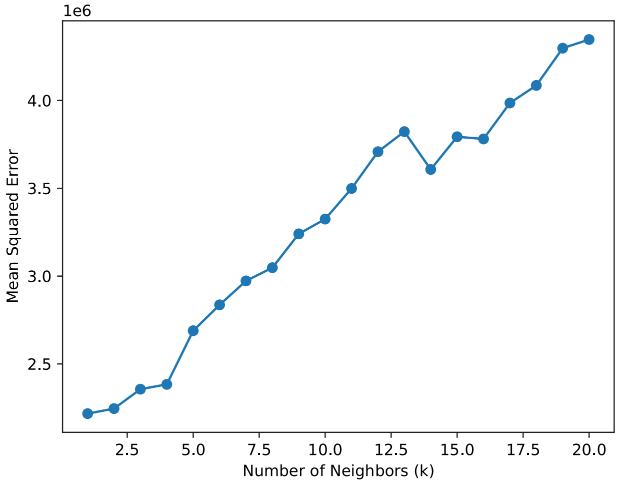

Fig. 4. Cross-validation mean squared error for different k values.

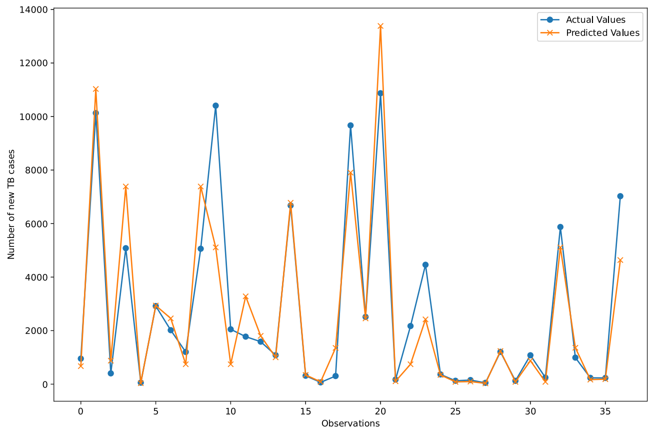

Fig. 5. Actual vs. Predicted values.

Table 2
Evaluation metrics.

Metric	Value
Mean squared error	1 825 604.5135135136
R-squared	0.8290852427928085

Fig. 6 exhibits the residual values derived from the K-Nearest Neighbors model on the test data. The examination of Fig. 6 indicates a notable prevalence of high residual values for the majority of predictions made by this model during the evaluation on the test data. Consequently, this analysis suggests the consideration of employing an alternative machine learning technique.

Table 2 displays the values of evaluation metrics for the performance of K-Nearest Neighbors Model on test data. Even though R-squared value is high, the larger Mean Squared Error (MSE) suggests that the K-Nearest Neighbor technique does not generalize well to new and unseen data.

3.2. Quantile regression

Quantile Regression is a Machine Learning Model that is more resistant to outliers and can handle data with a wide range of distributions. It has emerged as a promising approach for practical applications, offering a more comprehensive view than mean regression (Kumar

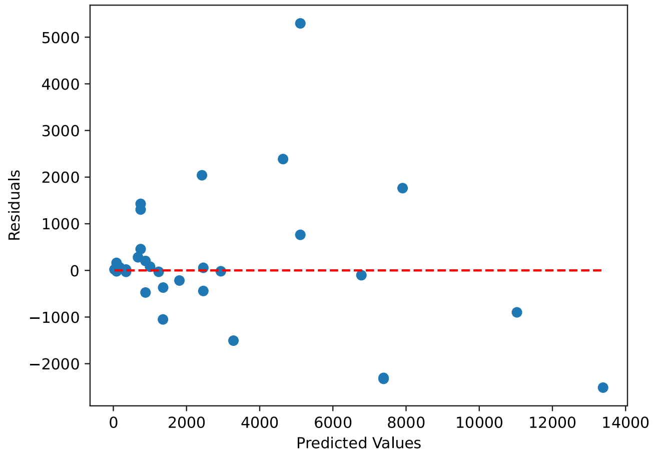

Fig. 6. Residual analysis.

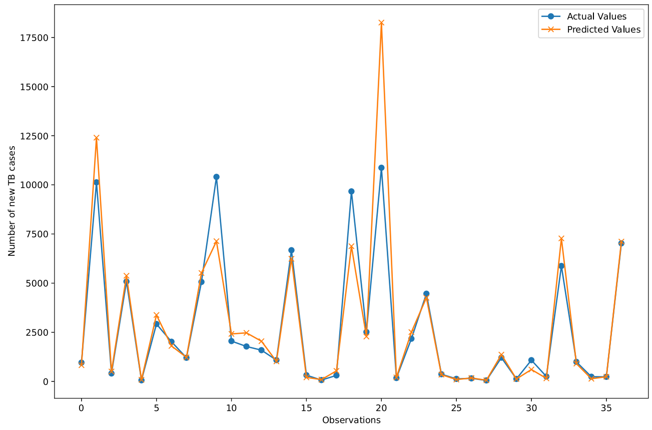

Fig. 7. Actual vs. Predicted values.

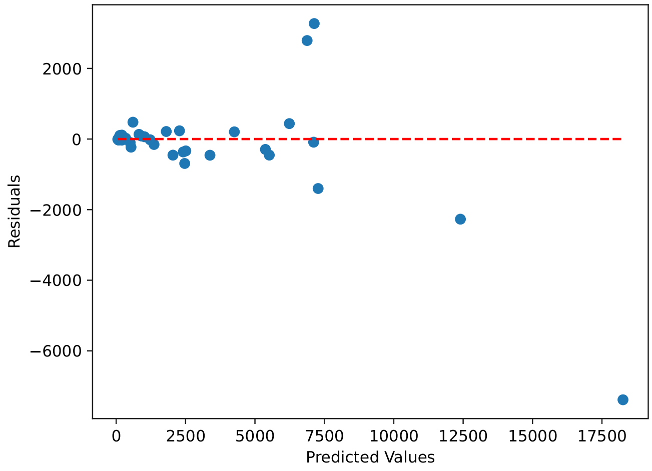

Fig. 8. Residual analysis.

et al., 2023). The Fig. 7 displays the trajectory of Actual and Predicted values across the observations of test data.

The Fig. 8 portrays the Residual Analysis of the Quantile Regression Model on test data. In Fig. 8, Most of the predicted values have low value of residuals except five.

Table 3 presents the evaluation metrics values illustrating the performance of the Quantile Regression (QR) Model on the test data.

Table 3
Evaluation metrics.

Metric	Value
Mean squared error	2 225 342.1865850654
R-squared	0.791661437782542

Table 4
Evaluation metrics.

Metric	Value
Mean squared error	1 547 136.267245946
R-squared	0.8551556936207015

To assess the assumptions of the Quantile Regression model, we performed the Breusch–Pagan test to evaluate the presence of heteroskedasticity. The Breusch–Pagan test yielded a p-value of 0.02963 and an f_p-value of 0.02808. These results indicate that the null hypothesis of homoskedasticity is rejected, suggesting that the model exhibits heteroskedasticity. Heteroskedasticity implies that the variance of the residuals is not constant across all levels of the independent variables, which can affect the efficiency of the estimates and the validity of statistical tests. Despite this, Quantile Regression remains a robust technique for modeling data with non-constant variance, as it does not assume homoskedasticity and can provide valuable insights across different quantiles of the dependent variable.

3.3. *Random forest*

In Machine Learning, Random forests are a combination of tree predictors in which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. This machine learning model mitigates overfitting (Breiman, 2001). As Random forests is an ensemble of decision trees, it is challenging to visualize the entire Random forest model. Fig. 9 depicts a visual representation of a single decision tree from the Random Forest model.

Fig. 10 presents a multiline chart depicting the variations in both observed and predicted values of the dependent variable using the Random Forest model on the test data. A visual examination of Fig. 10 reveals that, for the most part, the predicted values align well with the corresponding actual values, with only a few exceptions.

Fig. 11 illustrates the residual plot reflecting the performance of the Random Forest machine learning model on the test data. Analysis of Fig. 11 reveals that while there is a notable difference between predictions and actual values for a few data points, overall, the model demonstrates relatively accurate predictions.

Table 4 presents the evaluation metric values for the Random Forest model, illustrating the model's performance on the test data. Analysis of Table 4 highlights the relatively low Mean Squared Error (MSE), indicating favorable performance in the context of this study. Additionally, the coefficient of determination value in Table 4 is notably high, indicating a well-fitted model.

3.4. *XGBoost model*

Extreme Gradient Boosting (XGBoost) is a scalable end-to-end tree boosting system which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges (Chen and Guestrin, 2016). It is a type of gradient boosting with additional features and optimizations such as regularization, parallel processing, and tree-pruning techniques. Fig. 12 provides a visual depiction of a tree of the XGBoost model.

Fig. 13 displays a multiline chart illustrating the fluctuations in both observed and predicted values of the dependent variable using the XGBoost model on the test data. A detailed analysis of Fig. 13 reveals

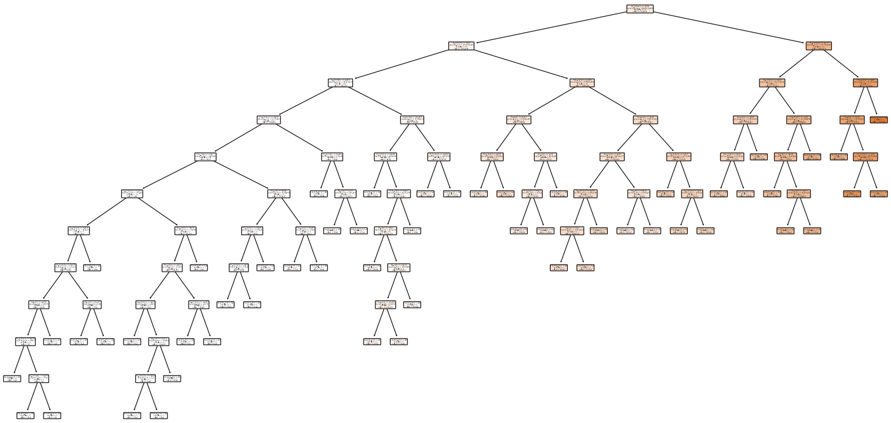

Fig. 9. A single decision tree from random forest model.

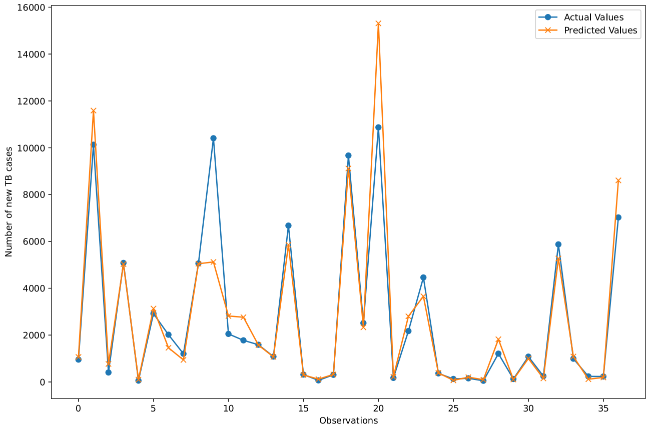

Fig. 10. Actual vs. Predicted values.

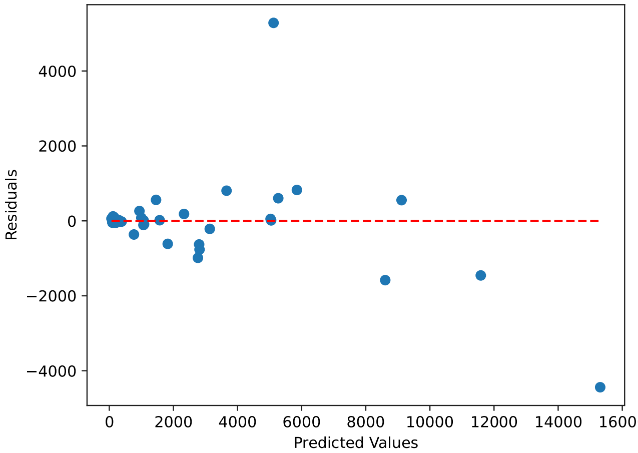

Fig. 11. Residual analysis.

that, for the most part, the predictions of the dependent variable closely align with the corresponding actual values, with a few exceptions. This observation suggests a well-fitted model. To further validate this insight, a residual analysis is conducted, and evaluation metrics are employed.

The visualization of residual analysis for the XGBoost model on test data is displayed in Fig. 14. A thorough examination of Fig. 14 makes sense that the occurrence of large residuals is rare, indicating a well fitted model.

Table 5
Evaluation metrics.

Metric	Value
Mean squared error	1 477 982.538665361
R-squared	0.8616299286715207

Table 6
Evaluation metrics.

Metric	Value
Mean squared error	2 771 451.6797361965
R-squared	0.7405341696696737

The Table 5 provides values of evaluation metrics to assess the performance of the model to generalize well to new and unseen data. The analysis of Table 5 suggests that XGBoost Model performs well on test data with low Mean Squared Error (MSE) and high R-squared (Coefficient of Determination) value.

3.5. LightGBM Model

LightGBM is a fast, distributed, high-performance gradient boosting method based on decision tree algorithms, used for machine learning tasks (Machado et al., 2019; Ke et al., 2017). The following Fig. 15 displays the variation of predicted values of dependent variables across the observations of test data along with the variation of corresponding actual values. The visual perception of Fig. 15 provides an intuitive information that some observations have a noticeable difference between actual and predicted values of dependent variable.

In Fig. 16, a graphical representation of residual analysis of LightGBM model on test data is plotted. Fig. 16 highlights that, upon visual inspection, certain predicted values exhibit higher residuals. This indicates comparatively moderate variation of predicted values of dependent variable from corresponding actual values.

Table 6 provides values of evaluation metrics of the LightGBM model depicting the performance of the model on test data. The analysis of Table 6 confirms that this model is comparable with previously constructed models with a reasonable value of coefficient of determination and Mean Squared Error (MSE).

3.6. CatBoost model

CatBoost is an open-source Gradient Boosted Decision Tree (GBDT) implementation for supervised machine learning (Hancock and Khoshgoftaar, 2020). It is an ensemble of decision trees to deal with large amounts of data effectively. The Fig. 17 provides intuitive information

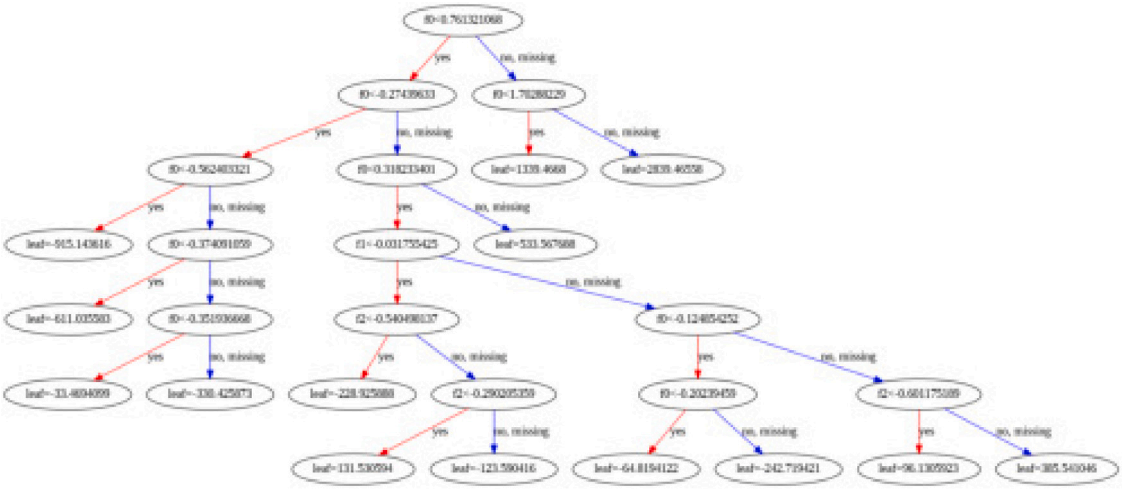

Fig. 12. Visualization of one of the trees in XGBoost model.

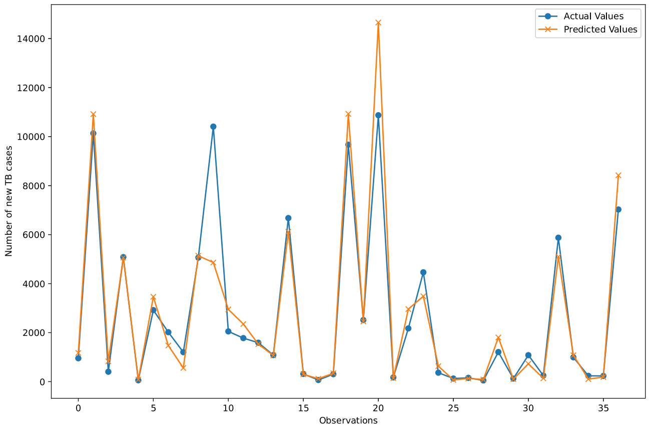

Fig. 13. Actual vs. Predicted values.

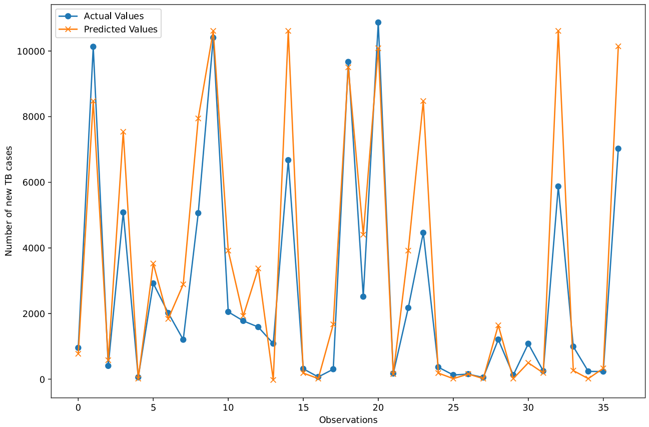

Fig. 15. Actual vs. Predicted values.

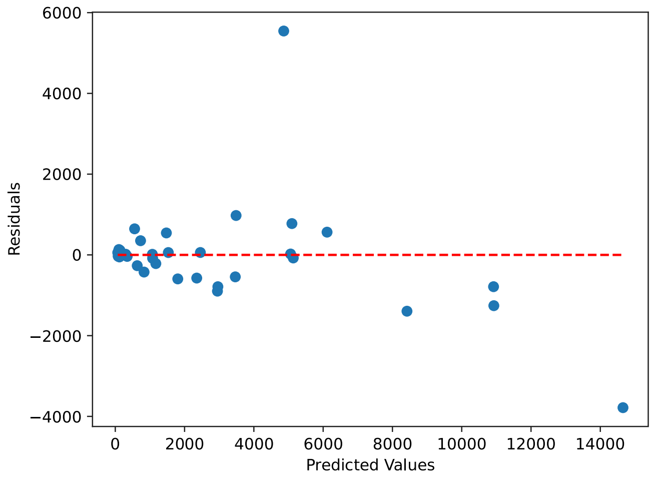

Fig. 14. Residual analysis.

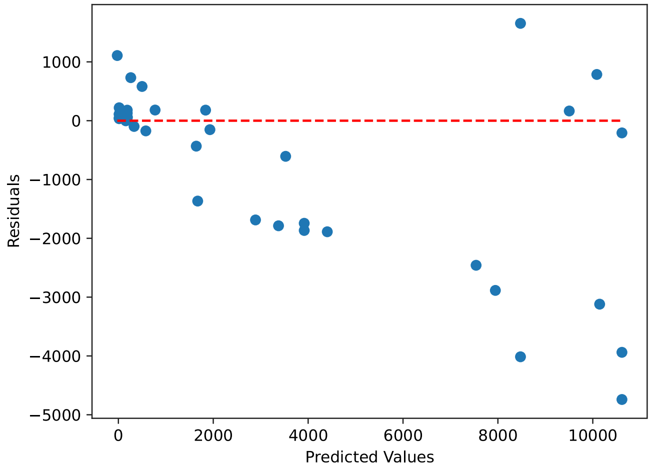

Fig. 16. Residual analysis.

about the fluctuations of predicted values of dependent variable with corresponding actual values in the test data.

Fig. 18 portrays the Residual Analysis of the CatBoost model on the test data. The occurrence of high-valued residuals is relatively small in this case, providing a comparable model. The detailed comparison of models is provided in the section ‘Model Comparison and Selection of Optimal Model’.

The accuracy of the performance of the CatBoost model on test data is measured using Mean Squared Error (MSE) and R-squared value. The resultant values are displayed in the Table 7. The metric values in Table 7 indicates the presence of a well-fitted model.

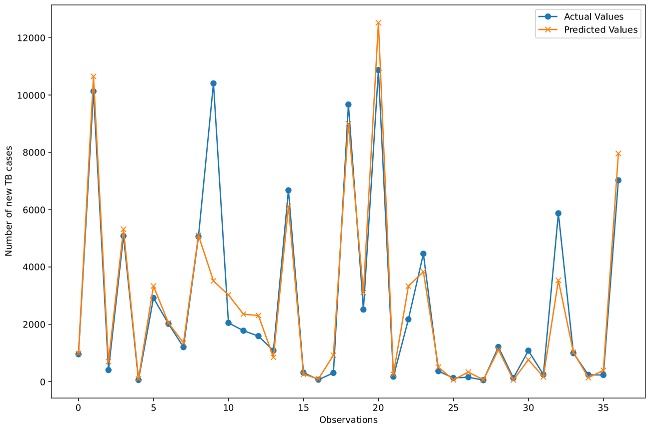

Fig. 17. Actual vs. Predicted values.

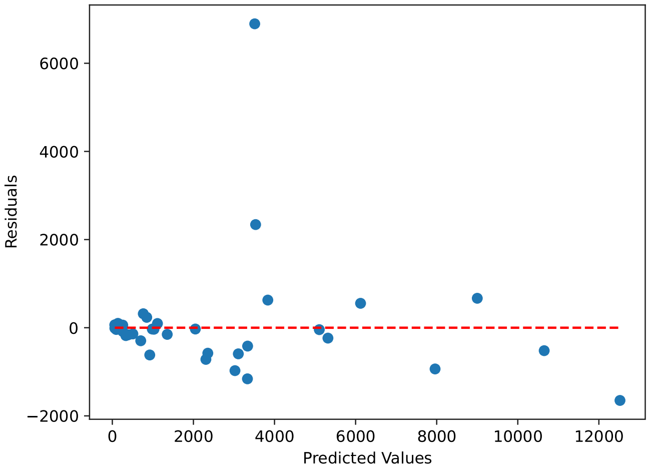

Fig. 18. Residual analysis.

Table 7
Evaluation metrics.

Metric	Value
Mean squared error	1 690 019.4090710224
R-squared	0.8417788437535669

4. Model comparison, selection of optimal model and tailoring a user-defined function

Various machine learning models were constructed to evaluate the performance of model on test data. We forecast the incidence of tuberculosis (New_Cases) based on parameters such as hh_Contacts, Prev_Cases, and No_Labs on test data to generalize how well the model fits to new and unseen data. The performance of these models on test data was assessed using standard evaluation metrics like Mean Squared Error (MSE) and R-squared value (Coefficient of Determination). Visual representation plays a crucial role in comparative analyses, enhancing the interpretability of findings. In this study, the Matplotlib package was employed to create bar charts that illustrate a comparative assessment of the evaluation metrics for the implemented machine learning models. Figs. 19 and 20 provides comparison of machine learning models for their predictive performance on test data based on values of Coefficient of determination and Mean Squared Error (MSE) respectively.

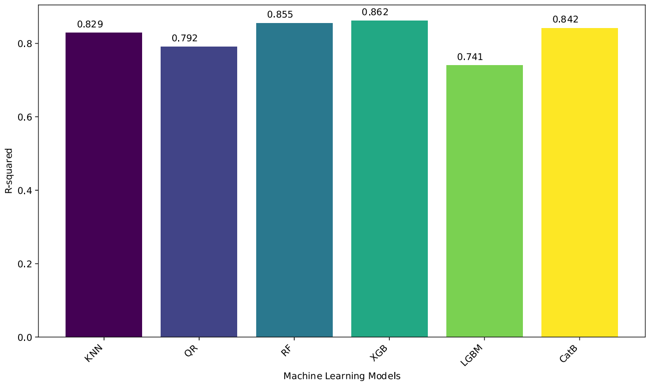

Fig. 19. Comparison based on R-squared.

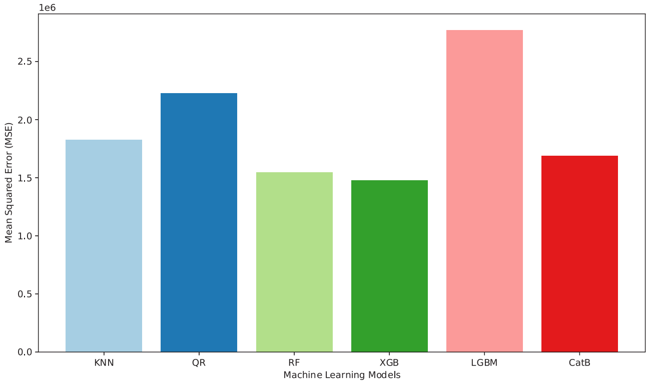

Fig. 20. Comparison based on MSE.

The optimal model is characterized by a high R-squared value and a low Mean Squared Error. After thorough examination of Figs. 19 and 20, our selection for the most effective machine learning model in this study is Extreme Gradient Boosting (XGBoost) model.

I employed the selected XGBoost model to forecast New_Cases values by inputting various Independent variable values. Table 8 presents the resultant predictions alongside the Independent variable values (hh_Contacts, Prev_Cases, No_Labs) and the actual values of the dependent variable (New_Cases). Additionally, within this manuscript, a custom function named ‘suggestion’ is formulated to offer recommendations for mitigating Tuberculosis incidence by addressing pertinent factors. In this context, we focus on the parameters of the XGBoost Model, specifically the Independent variables utilized in constructing the model.

The purpose of constructing the user-defined function ‘suggestion’ is to offer recommendations for mitigating the occurrence of tuberculosis by considering the identified features (independent variables) that influence it. The function outputs are presented below.

Suggestions for reducing New_Cases:
Adjust hh_Contacts to reduce New_Cases (Importance: 0.9330)
Adjust Prev_Cases to reduce New_Cases (Importance: 0.0478)
Adjust No_Labs to reduce New_Cases (Importance: 0.0192)

The visualization of the feature importance is provided in Fig. 21. The user-defined ‘suggestion’ function provides a general recommendation to modify the values of independent variables based on their respective feature importance, aiming to decrease the occurrence of tuberculosis. In the analysis of Fig. 21, notably, the feature ‘hh_Contacts’ demonstrates substantial importance, while ‘Prev_Cases’ and ‘No_Labs’ exhibit relatively lower feature importance. Consequently, to effectively reduce tuberculosis incidence, we recommend focusing on adjusting ‘hh_Contacts’ exclusively. For more precise details regarding the suggested adjustments, an analysis of Fig. 1 was conducted. Given the positive correlation observed between the feature ‘hh_Contacts’ and the

Table 8
Predictions of XGBoost model.

hh_Contacts	Prev_Cases	No_Labs	New_Cases	Predicted
18 000.0	843	593	6676	6112.723145
590.0	46	16	233	188.471542
1900.0	520	84	1203	557.502686
180.0	6	7	57	84.627579
1500.0	45	46	405	830.118652
720.0	7	1	155	129.411469
17 000.0	242	220	4462	3486.389648
5200.0	307	143	2018	1474.339478
730.0	25	114	305	342.530182
270.0	7	13	126	101.647568
6300.0	422	188	2515	2455.607910
5300.0	301	219	1589	1530.781982
2500.0	281	28	993	1077.725586
11 000.0	258	310	2920	3463.915527
310.0	16	11	237	104.209106
150.0	15	10	68	118.916321
140.0	18	7	68	118.916321
1800.0	105	38	1081	729.664307
19 000.0	2527	249	5874	5098.303223
7400.0	701	89	2173	2961.008301
24 000.0	300	164	5061	5137.084473
330.0	20	15	248	134.795303
4400.0	53	27	1084	1069.607422
24 000.0	382	87	5082	5059.948730
6000.0	37	44	1209	1803.658936
120.0	8	1	51	87.545189
360.0	19	30	316	300.128662
2500.0	120	60	955	1170.333984
31 000.0	672	105	7025	8419.661133
7900.0	528	90	2051	2948.931885
17 000.0	2182	460	10 407	4861.993164
84 000.0	1311	147	10 871	14 654.570312
1000.0	31	30	366	633.293152
59 000.0	309	187	10 131	10 918.321289
750.0	7	9	173	141.998291
9400.0	329	84	1777	2351.060791
280.0	0	5	128	71.079674
29 000.0	419	250	9669	10 926.706055

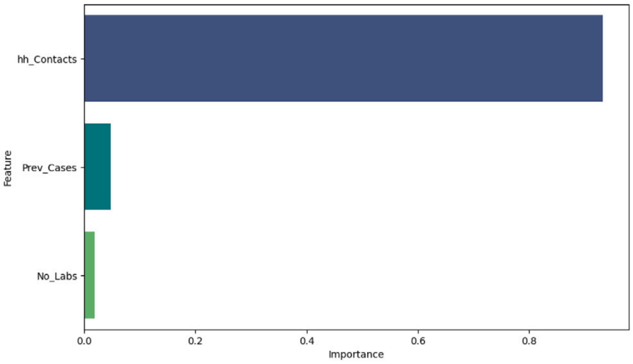

Fig. 21. Feature importance.

target variable 'New_Cases', we advocate for decreasing 'hh_Contacts' to diminish the incidence of tuberculosis.

5. Conclusions and future work

In this investigation, an extensive examination of predictive machine learning models was undertaken to conduct a comparative analysis aimed at identifying the most effective model for forecasting Tuberculosis incidence. Assessment metrics such as Mean Squared Error (MSE) and R-squared value (Coefficient of Determination) were consistently applied across all implemented machine learning models to evaluate their predictive capabilities. The selected model, XGBoost, demonstrated superior predictive accuracy and offered recommendations to mitigate Tuberculosis incidence based on data-derived insights and patterns. This prediction model helps identify high-risk areas, enabling targeted interventions and effective prevention strategies in the context of antibiotic-resistant TB. This data-driven approach has facilitated the development of a user-defined function that provides actionable suggestions for reducing Tuberculosis incidence. One key recommendation is to reduce household contacts to decrease new TB cases, which can be effectively achieved through quarantine measures. In the context of the 'WHO Global Tuberculosis Programme' led by the World Health Organization (WHO), this study proves valuable and informative. It contributes towards the overarching goal of achieving a world free of TB, with zero fatalities, disease prevalence, and suffering attributable to Tuberculosis.

In our prospective endeavors, we anticipate advancing the comprehensiveness and intricacy of predictive machine learning by integrating additional pertinent parameters. The augmentation is designed to facilitate an in-depth examination of Tuberculosis (TB) incidence through the application of time series analysis in conjunction with other regional and socio-economic factors. This refined methodology is poised to contribute to the development of more sophisticated models, thereby enhancing the efficacy of preventive measures aimed at mitigating the transmission of TB. Furthermore, future research initiatives could explore the integration of advanced data sources, such as genomics and environmental factors, to further refine predictive models and strengthen our understanding of the complex dynamics underlying TB transmission. Additionally, the incorporation of innovative technologies, such as real-time monitoring systems and artificial intelligence algorithms, could offer new avenues for the timely identification and containment of TB outbreaks, thereby bolstering public health efforts. Despite these advancements, our study has certain limitations. One key limitation is the relatively small number of significant predictors currently available for the predictive model. This constraint limits the model's ability to capture the full range of factors influencing TB incidence. Moreover, there are fewer practical applications in real life due to the limited scope of available data. To address these limitations, future research should focus on identifying and incorporating more variables that are strongly connected to the spread of TB cases. With a richer dataset, we can develop more accurate predictive models, leading to more informed and effective decisions for reducing the occurrence of new TB cases. By acknowledging these limitations, we highlight the importance of continuous data enhancement and methodological improvements to better support TB prevention and control efforts.

CRediT authorship contribution statement

Hamna Mariyam K.B.: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Sayooj Aby Jose:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Anuwat Jirawattanapanit:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Investigation, Funding acquisition. **Karuna Mathew:** Writing – review & editing, Visualization, Validation, Software, Methodology, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to express their deepest gratitude to Phuket Rajabhat University, Thailand, for their invaluable support and contributions. Their commitment and collaboration have been instrumental to this work.

Appendix A. Supplementary materials

The following supplementary materials are available at our GitHub repository:

- TuberculosisData

The repository contains raw data in 18 CSV files, Data Dictionary and the Input Dataset.

References

Abdualgalil, Bilal, Abraham, Sajimon, Ismael, Waleed, George, Dais, 2022. Modeling and forecasting tuberculosis cases using machine learning and deep learning approaches: A comparative study. http://dx.doi.org/10.1007/978-981-19-2600-6_11.

Orjuela-Canñón Alvaro, D., Jutinico Andrs, L., Carlos, Awad, Erika, Vergara, Angélica, Palencia, 2022. Machine learning in the loop for tuberculosis diagnosis support. Front. Public Health (ISSN: 2296-2565) 10, http://dx.doi.org/10.3389/fpubh.2022.876949, URL https://www.frontiersin.org/articles/10.3389/fpubh.2022.876949.

Ang, Cheng, 2001. Teaching mathematical modelling in Singapore schools. Math. Educ. 6.

Anggriani, Nursanti, Panigoro, Hasan S., Rahmi, Emli, Peter, Olumuyiwa James, Jose, Sayooj Aby, 2023. A Predator-Prey Model with additive Allee Effect and intraspecific competition on predator involving Atangana-Baleanu-Caputo derivative. Results Phys. 106489.

Badawy, M, Ramadan, N, Hefny, H.A., 2023. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. J. Electr. Syst. Inf. Technol. 10, 40. http://dx.doi.org/10.1186/s43067-023-00108-y.

Bokonda, Loola, Khadija, Ouazzani Touhami, Souissi, Nissrine, 2020. Predictive analysis using machine learning: Review of trends and methods. http://dx.doi.org/10.1109/ISAECT50560.2020.9523703.

Breiman, Leo, 2001. Random forests. Mach. Learn. 45, 5–32.

Chen, Tianqi, Guestrin, Carlos, 2016. XGBoost: A scalable tree boosting system. pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Chubb, M.C, Jacobsen, K.H., 2010. Mathematical modeling and the epidemiological research process. Eur. J. Epidemiol. 25, 1319. http://dx.doi.org/10.1007/s10654-009-9397-9.

Dndar, Sefa, Gokkurt, Burcin, Soylu, Yasin, 2012. Mathematical modelling at a glance: A theoretical study. Procedia - Soc. Behav. Sci. 46, http://dx.doi.org/10.1016/j.sbspro.2012.06.086.

Hancock, J.T, Khoshgoftaar, T.M., 2020. CatBoost for big data: an interdisciplinary review. J. Big Data 7, 94. http://dx.doi.org/10.1186/s40537-020-00369-8.

Hrizi, Olfa, Gasmi, Karim, Ltaifa, Ibtihel Ben, Alshammari, Hamoud, Karamti, Hanen, Krichen, Moez, Ammar, Lassaad Ben, Mahmood, Mahmood A., 2022. Tuberculosis disease diagnosis based on an optimized machine learning model. J. Healthc. Eng. 2022, 8950243. http://dx.doi.org/10.1155/2022/8950243, 13 pages.

Jose, Aby, Sayooj, et al., 2022. Mathematical modeling on transmission and optimal control strategies of corruption dynamics. Nonlinear Dynam. 109 (4), 3169–3187.

Jose, Aby, Sayooj, et al., 2023a. Mathematical modeling of chickenpox in Phuket: Efficacy of precautionary measures and bifurcation analysis. Biomed. Signal Process. Control 84, 104714.

Jose, Aby, Sayooj, et al., 2023b. Mathematical modeling on co-infection: transmission dynamics of Zika virus and Dengue fever. Nonlinear Dynam. 111 (5), 4879–4914.

Jose, Sayooj Aby, Yaagoub, Zakaria, Joseph, Dianavinnarasi, Ramachandran, Raja, Jirawattanapanit, Anuwat, 2024. Computational dynamics of a fractional order model of chickenpox spread in Phuket province. Biomed. Signal Process. Control 91, 105994.

Ke, G, Meng, Q, Finley, T, Wang, T, Chen, W, Ma, W, Ye, Q, Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc., 31463154, [Online]. Available: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

Kumar, A, Wadhvani, R, Rasool, A, Gupta, M., 2023. Quantile regression in machine learning: A survey. In: 2023 Third International Conference on Secure Cyber Computing and Communication. ICSCCC, Jalandhar, India, pp. 750–755. http://dx.doi.org/10.1109/ICSCCC58608.2023.10176807.

Machado, M.R, Karray, S, de Sousa, I.T., 2019. LightGBM: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In: 2019 14th International Conference on Computer Science & Education. ICCSE, Toronto, ON, Canada, pp. 1111–1116. http://dx.doi.org/10.1109/ICCSE.2019.8845529.

Nalunjogi, Joanitah, Mucching-Toscano, Sergio, Sibomana, Jean.Pierre, Centis, Rosella, D'Ambrosio, Lia, Alffenaar, Jan-Willem, Denholm, Justin, Blanc, Franois-Xavier, Borisov, Sergey, Danila, Edvardas, Duarte, Raquel, Garca-Garca, Jos-Mara, Goletti, Delia, Ong, Catherine W.M., Rendon, Adrian, Thomas, Tania A., Tiberi, Simon, van den Boom, Martin, Sotgiu, Giovanni, Migliori, Giovanni Battista, 2023. Impact of COVID-19 on diagnosis of tuberculosis, multidrug-resistant tuberculosis, and on mortality in 11 countries in Europe, Northern America, and Australia. A Global Tuberculosis Network study. Int. J. Infect. Dis. (ISSN: 1201-9712) 130 (Supplement 1), S25–S29. http://dx.doi.org/10.1016/j.ijid.2023.02.025.

Nithya, B, Ilango, V., 2017. Predictive analytics in health care using machine learning tools and techniques. In: 2017 International Conference on Intelligent Computing and Control Systems. ICICCS, Madurai, India, pp. 492–499. http://dx.doi.org/10.1109/ICCONS.2017.8250771.

Ozturk Kiyak, E., Ghasemkhani, B, Birant, D., 2023. High-level K-nearest neighbors (HLKNN) A supervised machine learning model for classification analysis. Electronics 12, 3828. http://dx.doi.org/10.3390/electronics12183828.

Peter, Galbraith, Clatworthy, N., 1990. Beyond standard ModelsMeeting the challenge of modelling. Educ. Stud. Math. 21, 137–163. http://dx.doi.org/10.1007/BF00304899.

Sandhu, Gursimrat K., 2011. Tuberculosis: current situation, challenges and overview of its control programs in India. J. Glob. Infect. Dis. 3 (2), 143–150. http://dx.doi.org/10.4103/0974-777X.81691.

Shamil, E, Jose, Sayooj Aby, Panigoro, Hasan S., Jirawattanapanit, Anuwat, Omede, B.I., Yaagoub, Zakaria, 2014. Understanding COVID-19 propagation: A comprehensive mathematical model with Caputo fractional derivatives for Thailand. Front. Appl. Math. Stat. 10, 1374721.

Tang, Na, et al., 2023. Machine learning prediction model of tuberculosis incidence based on meteorological factors and air pollutants. Int. J. Environ. Res. Public Health 20 (5), 3910. http://dx.doi.org/10.3390/ijerph20053910.

Tiwari, Akshita, Maji, Srabanti, 2019. Machine learning techniques for tuberculosis prediction. In: International Conference on Advances in Engineering Science Management & Technology (ICAESMT) - 2019. Uttaranchal University, Dehradun, India.

World Health Organization, Global Tuberculosis Programme, data source: https://www.who.int/teams/global-tuberculosis-programme/data.

World Health Organization, Global tuberculosis report 2020, https://www.who.int/publications/i/item/9789240013131.

World Health Organization, Global tuberculosis report 2023, https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023.