Communication in Physical Sciences, 2026, 13(5):706-717

Prediction of Infectious Diseases using Machine Learning: A Case Study of Tuberculosis in Nigeria

Rosemary Chika Nweze, Confidence Ifeoma Odoh, and Nneka Maryann Okafor
Received: 14 December 2025/Accepted: 28 April 2026 /Published: 14 May 2026
https://dx.doi.org/10.4314/cps.v13i5.3

Abstract: *Tuberculosis (TB) remains one of the leading causes of infectious disease mortality globally and continues to pose a major public health challenge in Nigeria, a country classified among the high TB-burden nations. Early prediction of TB incidence is essential for effective disease surveillance, timely intervention, and efficient allocation of healthcare resources. In this study, machine learning (ML) models were developed to predict tuberculosis trends in Nigeria using historical TB notification data obtained from the Nigerian National Tuberculosis and Leprosy Control Programme (NTBLCP) and other demographic, socio-economic, healthcare, and environmental datasets covering the period from 2010 to 2022. Key predictive variables included population density, HIV prevalence, poverty index, healthcare accessibility, historical TB incidence, temperature, and humidity. Four machine learning algorithms, namely Random Forest, Support Vector Machine (SVM), Logistic Regression, and Neural Network models were trained and evaluated using accuracy, sensitivity, precision, F1-score, and Receiver Operating Characteristic (ROC) curve analysis. The Random Forest model achieved the best predictive performance with an accuracy of 96%, sensitivity of 98%, precision of 93%, and F1-score of 95%. Logistic Regression and Neural Network models each achieved an accuracy of 86%, sensitivity of 88%, precision of 76%, and F1-score of 82%, while the SVM model recorded an accuracy of 86%, sensitivity of 91%, precision of 71%, and F1-score of 80%.The findings demonstrate that ensemble machine learning approaches, particularly Random Forest, are highly effective for tuberculosis prediction due to their ability to capture complex relationships among epidemiological and socio-environmental risk factors. This study highlights the potential of machine learning-based predictive systems in strengthening TB surveillance, improving early disease detection, and supporting evidence-based public health decision-making in Nigeria.*

Keywords: *Machine learning, tuberculosis, predictive model, infectious diseases, epidemiology.*

Rosemary Chika Nweze
Computer Science Department, Faculty of Natural and Applied Sciences, State University of Medical & Applied Sciences (SUMAS), Igbo-Eno, Enugu State, Nigeria.
Email: *rosemary.nweze@sumas.edu.ng*
https://orcid.org/0009-0007-7134-8302

Confidence Ifeoma Odoh
Computer Science Department, Faculty of Natural and Applied Sciences, State University of Medical & Applied Sciences (SUMAS), Igbo-Eno, Enugu State, Nigeria.
Email: *confidence.odoh@sumas.edu.ng*

Nneka Maryann Okafor
Federal Radio Corporation, Enugu, Nigeria.
Email: *nekafrcn@gmail.com*

1.0 Introduction

Tuberculosis (TB) remains one of the leading causes of morbidity and mortality globally. It continues to constitute a major public health challenge in Nigeria despite the availability of effective treatment and control strategies. . According to the World Health Organization, millions of new TB infections and deaths are recorded annually, with developing countries

bearing the highest disease burden due to poverty, weak healthcare systems, and inadequate diagnostic facilities. Nigeria consistently ranks among the top high-burden TB countries worldwide, accounting for a significant proportion of undiagnosed and untreated cases (World Health Organization, 2023; World Health Organization, 2024). The spread and persistence of TB are strongly associated with socio-economic and environmental determinants such as poverty, overcrowding, malnutrition, HIV co-infection, poor sanitation, and limited access to healthcare services. (Abubakar et al., 2022).

"Conventional tuberculosis surveillance systems in Nigeria primarily depend on passive case detection, manual reporting procedures, and periodic epidemiological assessments, which are often affected by delayed reporting, under diagnosis, and incomplete data (NTBLCP, 2022). These challenges limit the ability of public health authorities to anticipate TB trends and respond proactively. Consequently, there is an increasing need for intelligent predictive systems capable of analyzing large and complex datasets to support early warning mechanisms, disease forecasting, and evidence-based public health interventions. There are also gaps in TB-specific applications within Nigeria, particularly studies that integrate socio-economic, environmental, and healthcare access variables. This study seeks to address these gaps by developing and evaluating ML-based models tailored to the Nigerian TB context.

In recent years, advancements in artificial intelligence and machine learning (ML) have transformed healthcare analytics by enabling the automated analysis of large-scale datasets and the identification of complex non-linear relationships among disease predictors. (Rajkomar et al., 2019; Chien et al., 2020). Unlike traditional epidemiological approaches, machine learning models can continuously learn from historical patterns, improve predictive performance, and handle multidimensional datasets with high accuracy. Machine learning has been widely applied in public health to predict diseases such as malaria, dengue, COVID-19, cholera, and tuberculosis (Liu et al., 2023; Guan et al., 2019). This study applies four machine learning models—Random Forest, Support Vector Machine (SVM), Logistic Regression, and Neural Network models—to predict tuberculosis trends in Nigeria using historical, demographic, socio-economic, healthcare, and environmental datasets.

The aim of this study is to develop and evaluate predictive machine learning models capable of improving tuberculosis surveillance, forecasting disease trends, and supporting data-driven public health decision-making in Nigeria.. "The significance of this study lies in its potential contribution to strengthening tuberculosis surveillance and control strategies in Nigeria through the application of intelligent predictive systems. The findings may assist healthcare policymakers, epidemiologists, and public health agencies in identifying high-risk regions, optimizing resource allocation, improving early intervention strategies, and enhancing evidence-based decision-making for TB management."

1.1 Literature Review

Several studies have explored the application of machine learning and predictive analytics in infectious disease surveillance and epidemiological forecasting. These studies demonstrate the growing importance of artificial intelligence in improving disease monitoring, prediction accuracy, and public health response systems. Recent advances in machine learning have significantly improved the prediction and surveillance of infectious diseases, including tuberculosis. Previous studies emphasize the integration of diverse data sources, advanced algorithms, and explainable models to enhance predictive accuracy and public health relevance.

Abubakar et al. (2022) examined the tuberculosis burden in Nigeria, highlighting persistent gaps in case detection and reporting. Their work emphasized the need for data-driven tools to complement traditional surveillance systems. Although the study did not directly apply machine learning, it provided critical epidemiological insights and contextual factors such as HIV co-infection, poverty, and health system constraints that are essential predictors in ML-based TB models.

Adepoju (2022) analyzed Nigeria's TB control efforts and identified under diagnosis, delayed reporting, and weak health infrastructure as major barriers to effective disease management. The study suggested that innovative digital health and predictive analytics solutions, including machine learning, could improve early detection and guide targeted interventions in high-risk populations.

Liu et al. (2023) applied multiple machine learning models including Random Forest and Gradient Boosting to predict infectious disease incidence using socio-economic and environmental data. Their results demonstrated that ensemble models consistently outperformed traditional regression approaches. The study further showed that incorporating social determinants of health significantly enhanced model performance. The finding is highly relevant to TB prediction in Nigeria./ Collectively, these studies demonstrate that ensemble and hybrid machine learning approaches often outperform traditional statistical methods in infectious disease prediction due to their ability to capture complex relationships among epidemiological variables.

Song and Yoon (2024) investigated the use of social media sentiment analysis for infectious disease prediction. By integrating public sentiment data with epidemiological records, their models achieved improved predictive performance compared to models relying solely on health data. While the study focused on general infectious diseases rather than TB specifically, it highlighted the potential of non-traditional data sources to strengthen disease forecasting frameworks.

Amshi et al. (2024) developed a machine learning driven outbreak prediction model for cholera in Nigeria using advanced preprocessing techniques such as SMOTE for data balancing and dimensionality reduction. Their Extreme Gradient Boosting model achieved very high predictive accuracy. Although disease-specific, the methodological framework demonstrates how advanced ML pipelines can be adapted for TB prediction using Nigerian public health data.

World Health Organization (2023, 2024) reports emphasized the growing role of digital technologies and artificial intelligence in strengthening TB surveillance and monitoring. The reports highlighted the importance of predictive analytics for identifying hotspots, forecasting disease trends, and optimizing resource allocation particularly in high-burden countries like Nigeria. Although previous studies have demonstrated the effectiveness of machine learning in infectious disease prediction, limited research has specifically focused on tuberculosis forecasting in Nigeria using integrated socio-economic, environmental, and healthcare datasets. This study therefore contributes to existing knowledge by developing comparative machine learning models tailored to the Nigerian tuberculosis surveillance context."

20 Materials and Method

2.1 Dataset

The datasets used in this study comprised epidemiological, demographic, socio-economic, healthcare accessibility, and environmental variables associated with tuberculosis incidence in Nigeria. TB notification data were obtained from the Nigerian National Tuberculosis and Leprosy Control Programme (NTBLCP) and World Health Organization reports covering multiple years from 2010 to 2022. The dataset consisted

of yearly and quarterly tuberculosis notification records collected between 2010 and 2022 across multiple regions in Nigeria. Population and socio-economic indicators were sourced from the National Bureau of Statistics (NBS) and the World Bank. Environmental and climatic data were obtained from publicly available meteorological datasetsThe integration of epidemiological, socio-economic, and environmental datasets enabled the development of a comprehensive predictive framework capable of capturing multiple determinants of tuberculosis transmission."

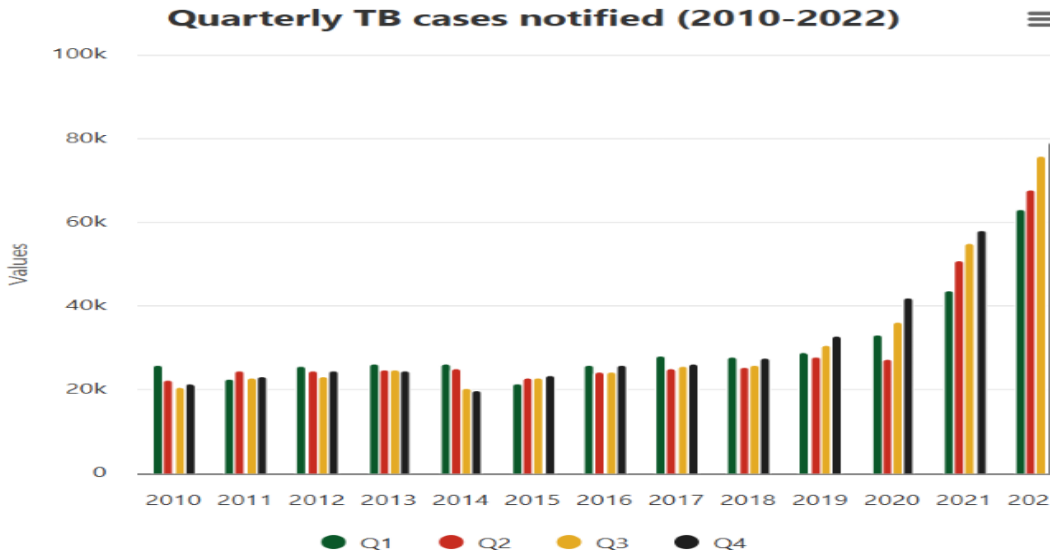


Fig. 1: Sample of quarterly TB cases notified from NTBLCP (2010-2022)
Source: https://ntblcp.org.ng/data-centre/#top

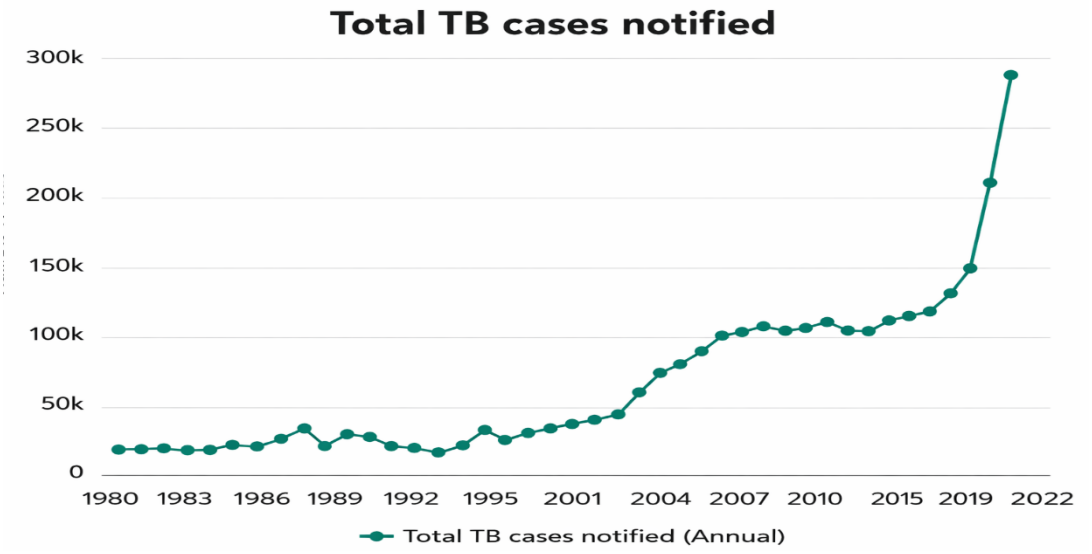


Fig. 2: Annual tuberculosis cases reported in Nigeria from 1980 to 2022

2.2 Data Preprocessing

To improve data quality and model performance, the collected datasets were

preprocessed prior to model development. Data preprocessing involved handling missing values, removing duplicate records, correcting inconsistencies, encoding categorical variables, normalizing numerical features, and generating lag variables to capture temporal dependencies in tuberculosis incidence. Feature scaling and normalization were performed using standard normalization techniques to ensure uniformity among variables with different measurement scales.

2.3 Feature Extraction

Feature extraction was performed to identify and retain the most informative variables influencing tuberculosis incidence while reducing data dimensionality and improving model efficiency. The process involves identifying the most relevant variables and transforming them into a format suitable for ML algorithms. Feature extraction was carried out to identify the most relevant variables influencing TB incidence. Key features included historical TB cases, population density, HIV prevalence, poverty index, healthcare accessibility, temperature, and humidity. The feature extraction techniques used in this research include

a. Principal Component Analysis (PCA): PCA simplifies complex datasets by focusing on the most important features. It transforms data into a new coordinate system, preserving the most important variations while reducing dimensionality.
b. Linear Discriminant Analysis (LDA): This is similar to PCA, but focuses on maximizing the separation between different classes.

2.4 Machine Learning Algorithms

Machine learning algorithms used in TB prediction models include decision trees, random forests, logistic regression, support vector machines (SVM), and deep learning techniques such as neural networks. These models were selected based on their ability to capture complex relationships between epidemiological, demographic, socio-economic, and environmental data. Also ensemble learning methods which combine the predictions of these multiple models helps to improve accuracy and reduce the risk of model overfitting.

Random Forest: An ensemble learning method known for high accuracy and robustness in handling complex datasets.

Neural Networks: Effective in capturing non-linear relationships and interactions among multiple predictors.

Support Vector Machines (SVM): Suitable for classification and regression tasks and effective in high-dimensional spaces.

Logistic Regression: A baseline model valued for interpretability and transparency.

2.5 Model Training

The dataset was split into training (70–80%), validation (10–15%), and testing (10–15%) subsets. Data splitting encourages generalization. Generalization refers to a model's ability to adapt effectively to new or unseen data, such as test datasets. Training datasets are the data used to train the model. The purpose of training is to identify the best-performing model. During this process, the machine learning algorithm learns patterns and relationships in the data by optimizing its parameters and evaluating its performance. Validation dataset is used to check the overall accuracy of the model and the test data is used to evaluate the performance of the model. The newly developed model is tested with new dataset to determine its ability to generalize well. Hyper parameters are then tuned to further improve the model's overall performance.

2.6 Model Evaluation

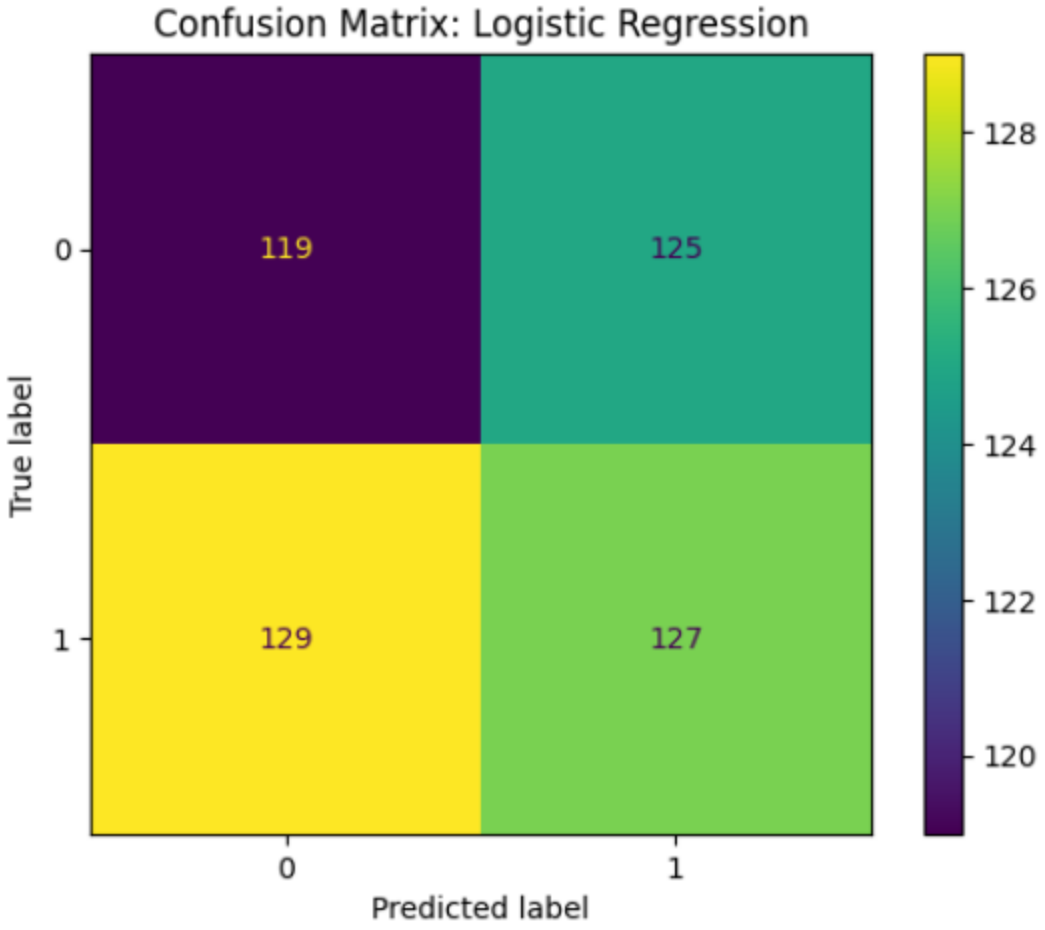
Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic curve (AUC). These metrics assess the models'

ability to correctly predict TB incidence while minimizing false positives and false negatives.

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.84	0.93	0.88	82
1	0.88	0.76	0.81	58
accuracy			0.86	140
macro avg	0.86	0.84	0.85	140
weighted avg	0.86	0.86	0.86	140

Fig. 3: Classification report for Logistic Regression model.



"Fig. 4: Confusion matrix for the Logistic Regression model."

Neural Network Classification Report:

	precision	recall	f1-score	support
0	0.85	0.93	0.89	82
1	0.88	0.78	0.83	58
accuracy			0.86	140
macro avg	0.87	0.85	0.86	140
weighted avg	0.87	0.86	0.86	140

Fig 5: Classification report for Neural Network model

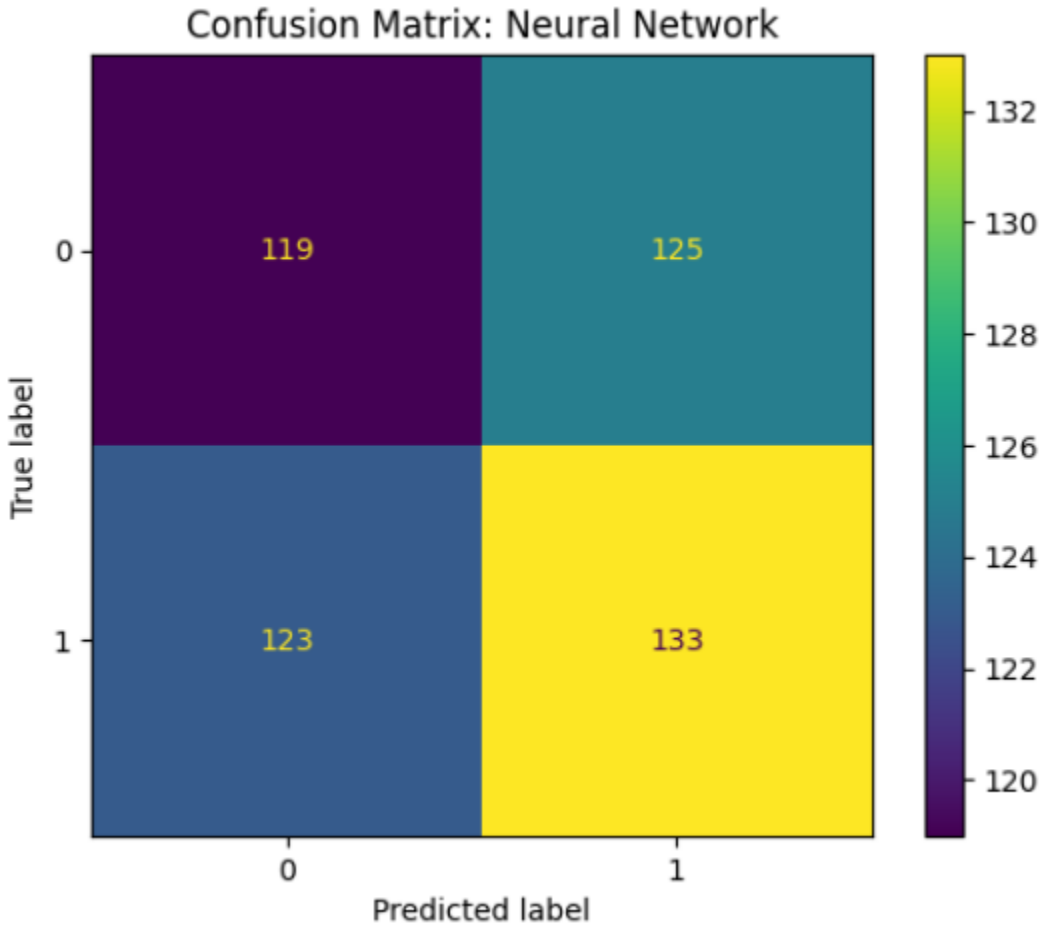


Fig 6: Confusion matrix result for Neural Network

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	82
1	0.98	0.93	0.96	58
accuracy			0.96	140
macro avg	0.97	0.96	0.96	140
weighted avg	0.96	0.96	0.96	140

Fig. 7: Classification report for Random Forest

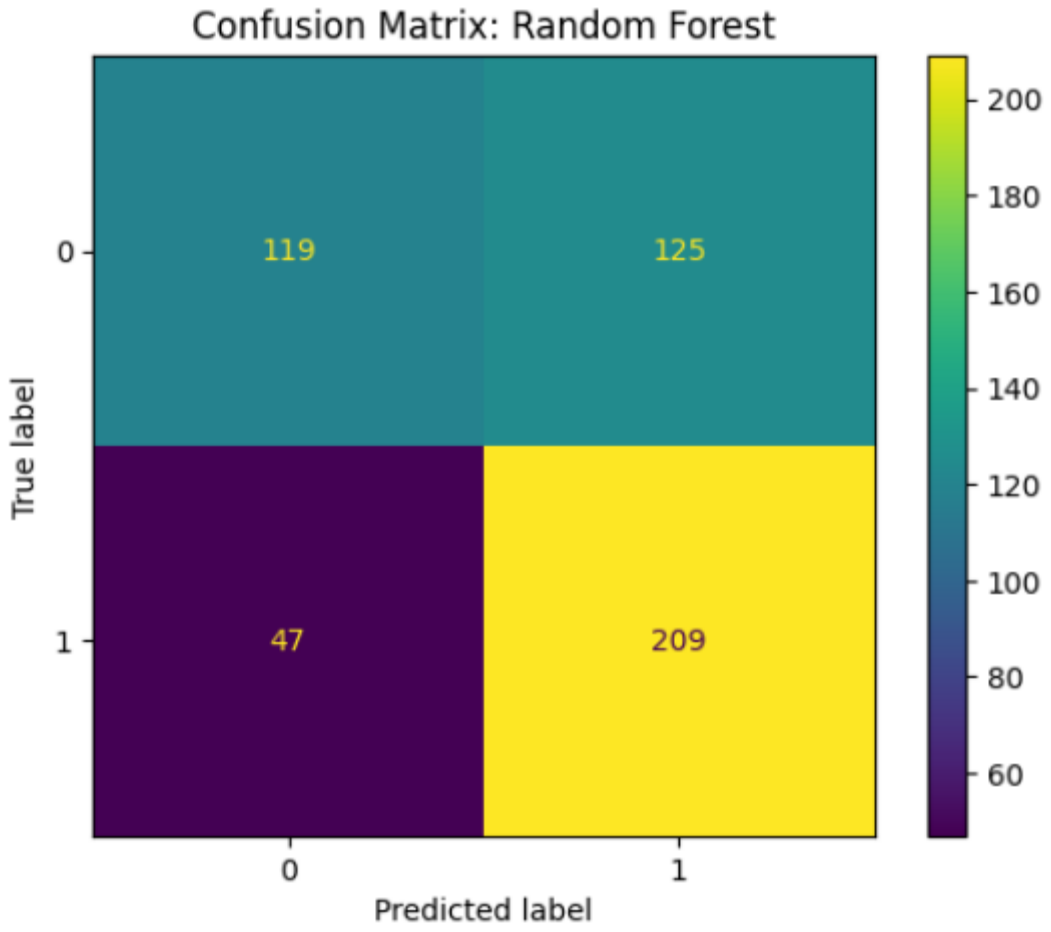


Fig 8: Confusion matrix for Random Forest

Support Vector Machine Classification Report:

	precision	recall	f1-score	support
0	0.81	0.95	0.88	82
1	0.91	0.69	0.78	58
accuracy			0.84	140
macro avg	0.86	0.82	0.83	140
weighted avg	0.85	0.84	0.84	140

Fig 9: Classification report for Support Vector Machine

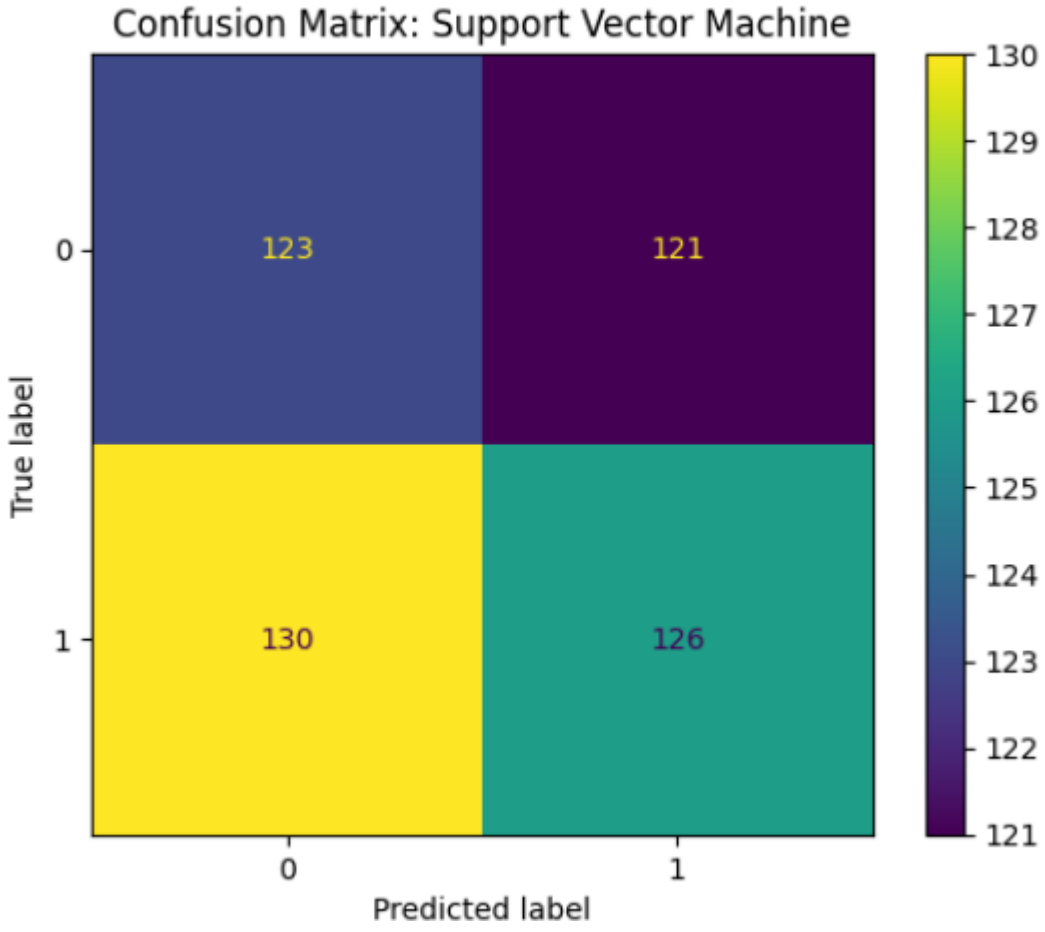


Fig 10: Confusion matrix for Support Vector Machine

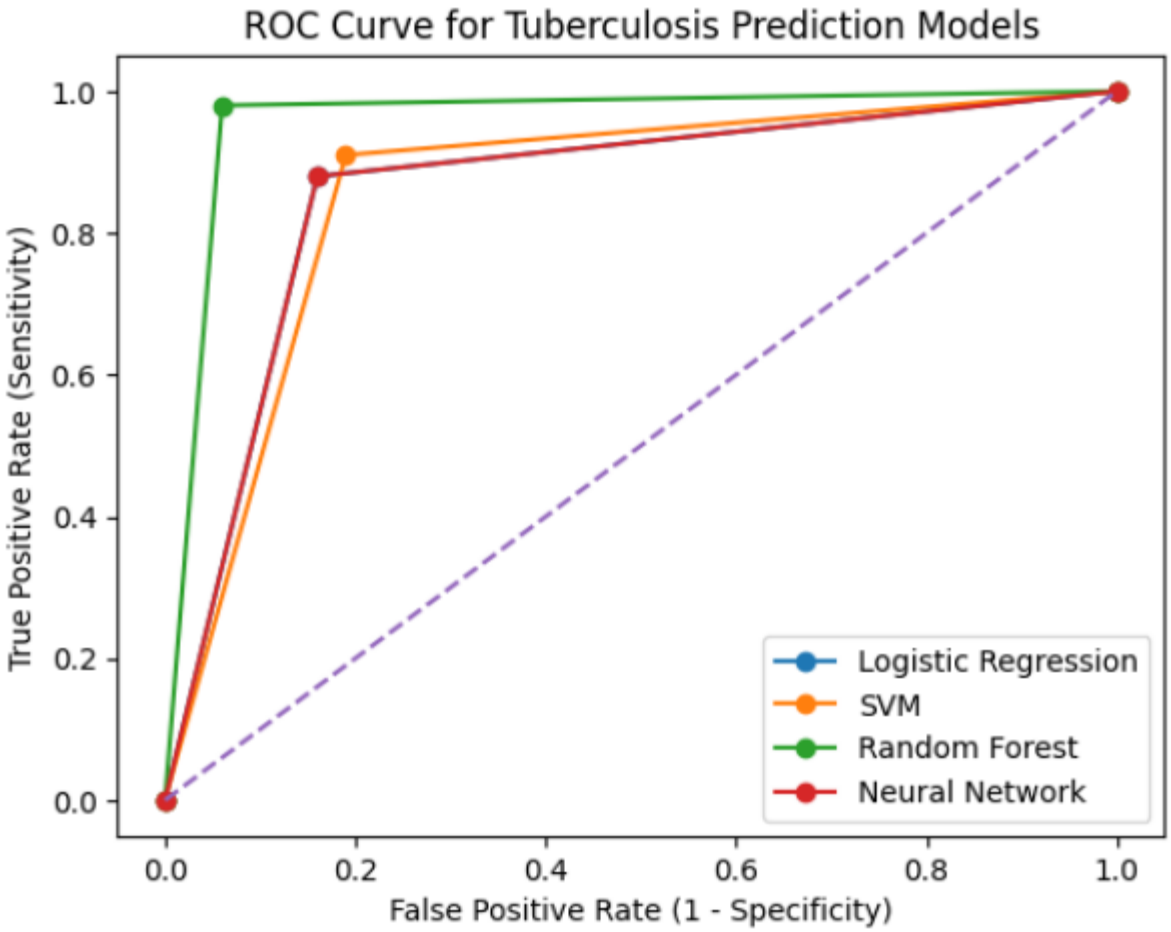


Fig 11: Receiver Operating Characteristic (ROC) curve for Tuberculosis prediction.

3.0 Results and Discussion

This study evaluated and compared the predictive performance of four machine learning algorithms—Random Forest, Support Vector Machine (SVM), Neural Network, and Logistic Regression—for tuberculosis prediction in Nigeria.

Although all the models trained achieved good results, the random forest model outperformed them in predictive accuracy, robustness and sensitivity.

Table 1: Performance evaluation metrics of machine learning models for tuberculosis prediction."

Algorithm	Accuracy	Sensitivity (Recall)	Precision	F1-Score
Logistic Regression	0.86	0.88	0.76	0.82
Neural Network	0.86	0.88	0.76	0.82
Random Forest	0.96	0.98	0.93	0.95
Support Vector Machine	0.86	0.91	0.71	0.80

The performance of four machine learning algorithms Logistic Regression, Support Vector Machine (SVM), Random Forest, and Neural Network was evaluated for tuberculosis prediction using accuracy, sensitivity, precision, and F1-score.

Among the evaluated models, the Random Forest classifier demonstrated the best overall predictive performance. The model achieved an accuracy of 96%, indicating excellent classification capability. Its sensitivity of 98% shows a strong ability to correctly identify tuberculosis-positive cases, while the precision value of 93% indicates a relatively low false-positive rate. Furthermore, the F1-score of 95% reflects an effective balance between sensitivity and precision. Both Logistic Regression and Neural Network models demonstrated moderate predictive performance, each achieving an accuracy of 86%, sensitivity of 88%, precision of 76%, and an F1-score of 82%. Although these models were effective in identifying tuberculosis-positive cases, their lower precision values suggest reduced reliability compared to the Random Forest model. The Support Vector Machine model achieved an accuracy of 86% and a sensitivity of 91%, indicating relatively strong capability in identifying tuberculosis-positive cases. However, its lower precision value of 71% suggests a higher occurrence of false-positive predictions, which reduced its overall F1-score to 80%.

Receiver Operating Characteristic (ROC) curve analysis further confirmed the superior discriminatory performance of the Random Forest model, which exhibited the largest Area Under the Curve (AUC). This indicates a stronger ability to distinguish between tuberculosis-positive and tuberculosis-negative cases compared to the other evaluated models. The findings of this study demonstrate significant differences in the predictive capabilities of the evaluated machine learning algorithms for tuberculosis surveillance in Nigeria. Models with lower sensitivity values may fail to identify infected individuals, thereby increasing the risk of undetected From a public health perspective, such false negatives are particularly problematic, as undetected TB cases contribute to ongoing community transmission and delayed treatment initiation.

The Neural Network model showed marginal improvement in precision and recall, suggesting some capacity to model non-linear relationships among demographic, clinical, and environmental variables associated with TB transmission. However, its overall performance remained limited, highlighting the challenges of applying neural architectures to epidemiological datasets with constrained sample sizes and heterogeneous features.

In contrast, the Random Forest model demonstrated substantially superior performance, achieving the highest recall (0.82) and F1-score (0.71). High recall is epidemiologically significant because it reflects the model's ability to correctly identify the majority of TB-positive individuals. This characteristic is essential for disease surveillance systems, where early detection and case finding directly influence transmission control, treatment outcomes, and resource allocation.

The higher Area under the Curve (AUC) achieved by the Random Forest model further indicates strong discriminatory power between TB-positive and TB-negative cases. This suggests that ensemble-based approaches are better suited for capturing the complex interactions among socio-economic, environmental, and health-related predictors that drive TB incidence in Nigeria.

4.0 Conclusion

This study evaluated the performance of Random Forest, Support Vector Machine (SVM), Logistic Regression, and Neural Network models for predicting tuberculosis incidence in Nigeria using epidemiological, socio-economic, healthcare, and environmental data. The results showed that the Random

Forest model achieved the best overall performance with the highest accuracy, sensitivity, precision, and F1-score.

The findings demonstrate that machine learning models can effectively capture complex relationships among tuberculosis risk factors and improve disease prediction. The study further highlights the potential of machine learning-based systems in strengthening TB surveillance, supporting early detection, improving resource allocation, and enhancing public health decision-making in Nigeria. However, the study was limited by data availability and possible inconsistencies in surveillance records. Future research should incorporate larger and real-time datasets as well as advanced hybrid or deep learning approaches to further improve predictive performance.

5.0 References

Abubakar, I., & et al. (2022). Tuberculosis in Nigeria: Burden, challenges, and prospects. *International Journal of Tuberculosis and Lung Disease, 26*(10), 897–905. https://doi.org/10.5588/ijtld.22.0145

Adepoju, P. (2022). Nigeria's battle with tuberculosis. *The Lancet Respiratory Medicine, 10*(4), e25–e26. https://doi.org/10.1016/S2213-2600(22)00045-6

Guan, Q., He, S., & Zhang, Y. (2019). Machine learning models for predicting infectious disease outbreaks. *Computers, Environment and Urban Systems, 74*, 128–139. https://doi.org/10.1016/j.compenvurbsys.2018.12.002

Liu, Y., & et al. (2023). Application of machine learning models in infectious disease prediction. *BMC Infectious Diseases, 23*, Article 112. https://doi.org/10.1186/s12879-023-08045-7

National Tuberculosis and Leprosy Control Programme. (2022). *Annual TB report, Nigeria*.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

World Health Organization. (2023). *Global tuberculosis report 2023*. https://www.who.int/publications/i/item/9789240083851

World Health Organization. (2024). *Global tuberculosis report 2024*. https://www.who.int/publications/i/item/9789240095311

Declaration

Consent for publication

Not Applicable

Availability of data and materials

The publisher has the right to make the data public

Ethical Considerations

Not applicable

Competing interest

The authors report no conflict or competing interest

Funding:

The authors declared no xternal esource of funding

Authors' Contribution

Rosemary Chika Nweze conceptualized the study, designed the machine learning models, conducted data analysis, and drafted the manuscript. Confidence Ifeoma Odoh contributed to data collection, preprocessing, literature review, and model evaluation. Nneka Maryann Okafor participated in result interpretation, manuscript editing, and critical revision of the study. All authors reviewed and approved the final manuscript.