European Journal of Preventive Medicine
2025, Vol. 13, No. 5, pp. 115-120
https://doi.org/10.11648/j.ejpm.20251305.15

Research Article

Adoption of Optimal Time Series Models for Forecasting on New and Relapse Tuberculosis Cases in Tanzania

Eunice Silas[1, 2, *], Sanket Pandhare[1], Anael Sam[2], Devotha Nyambo[2], Hamimu Omary Kigumi[3]

[1]Department of Prevention Services, Ministry of Health, Dodoma, Tanzania

[2]Department of Computer Engineering, The Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania

[3]Department of the Health Services, Local Government Training Institute, Dodoma, Tanzania

Abstract

Background: Time-series models forecasting plays key role in predicting TB cases. Despite of its importance some models consist of limitation that decrease its efficiency. To overcome this, adoption of optimal model with highly proficient forecasting is encouraged. *Objective* This study was aimed to adopt an optimal Time Series model for forecasting new and relapse Tuberculosis cases in Tanzania. *Setting:* The study use ARIMA, HWES and LSTM Time series models to find optimal modal that works efficiently on forecasting of TB cases. *Method*s: A cross-sectional study was conducted at Kibong'oto National Infectious Diseases Hospital, Moshi Tanzania from January 2021 to December 2024. Muilt- stage sampling was used to recruit 3911 TB cases registered from January 2015 to December 2020. A Microsoft Excel 2019 was used to create database with total of 2 columns and 72 row. Dataset was divided into training and testing cutoff points of 69% and 31% respectively to obtain optimal time series models as per Xu & Goodacre (2018). Tables and figures were used for interpretation of results. *Results*: A total of 3911 TB cases with annual average of 651.83. The periodic variations and declines were observed. The error metric values MAE, MAPE, and RMSE show ARIMA modal better performance on forecasting the TB cases due highest scores than others modals. *Conclusion:* The ARIMA model offers advanced predictions of TB cases, that help timely planning of prevention and control measures.

Keywords

Adoption, Optimal, Time, Model, Forecasting, Relapse, Tuberculosis

1. Introduction

Tuberculosis is the global infectious diseases that causes morbidity and mortality. It is the 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 [1]. Tanzania is among the 30 high burden countries. According to 2022 reports, Tanzania has reported slightly decreased by 27%, equivalent from 306 in 2015 to

*Corresponding author: silase@nm-aist.ac.tz (Eunice Silas)

Received: 24 August 2025; **Accepted:** 9 September 2025; **Published:** 17 October 2025

222 per 100000 population per year in 2020 and TB notification decreased to 146 from 148 per 100000 population in between 2020 to 2021 [2].

The time-series models forecasting techniques plays an important role in planning and making decision regarding to the TB intervention for purpose of the achieving the End TB strategy by 2030. Despite of its importance some models consist of limitation on its applications thus decrease its efficiency and its effectiveness of delivery of the positive results. Many studies recommend that on selection of the Time series models for forecasting the infectious diseases, understanding its behavior and its capability in producing impact on public health early warning surveillance is needed [3, 4]. Furthermore, the model should have highly proficient in accurate short-term forecasting and has gained substantial traction in the domain of forecasting infectious disease cases, emerging as the predominant methodology for TB prediction worldwide [5, 6].

The recent studies conducted in Iran and China, reported every Time series model used in forecast ions consists of the shortcoming that affect the process which may result to poor detection of the infectious diseases adoption of the optimal Time series modal with low limitations in its function such as Seasonal-Trend decomposition using LOESS (STL) is needed [7].

Therefore, Using Autoregressive Integrated Moving Average (ARIMA), Holt-Winters Exponential Smoothing (HWES) and Long Short-Term Memory (LSTM), This study was aimed to adopt the optimal Time Series model the will efficient and effective on Forecasting on New and Relapse Tuberculosis Cases in Tanzania.

2. Description of the Study Time Series Models

2.1. Autoregressive Integrated Moving Average (ARIMA), Model

An ARIMA model, or Autoregressive Integrated Moving Average model, is a statistical tool used for analyzing and forecasting time series data. It combines autoregressive (AR) and moving average (MA) models, accounting for the dependencies of a time series on its own past values and past forecast errors, while also integrating differencing to handle non-stationary data. Essentially, it models the future values of a time series based on its past values, past forecast errors, and the degree of differencing needed to make the data stationary.

2.2. Long Short-Term Memory (LSTM) Models

Long Short-Term Memory (LSTM) models are a type of recurrent neural network (RNN) specifically designed to handle the vanishing gradient problem, making them effective for processing sequential data and capturing long-range dependencies. They achieve this through internal memory cells and gating mechanisms that regulate the flow of information.

An LSTM unit is typically composed of a cell and three gates: an input gate, an output gate [8] and a forget gate [9]. The cell remembers values over arbitrary time intervals, and the gates regulate the flow of information into and out of the cell. Forget gates decide what information to discard from the previous state, by mapping the previous state and the current input to a value between 0 and 1.

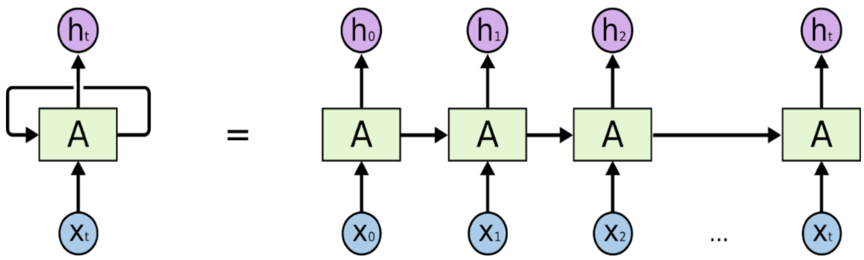

Figure 1. *Representation of the unfolding of LSTM time-steps (Li et al., 2020).*

In Figure 1 first, it takes the X_0 from the sequence of input and then outputs h_0 which together with X_1 is the input for the next step. Also, h_1 from the next is the input with X_2 for the next step, and so on. This way, it keeps remembering the context while training the model [10].

2.3. Holt-Winters Exponential Smoothing (HWES)

Holt-Winters Exponential Smoothing (HWES) is a time series forecasting method that extends simple exponential smoothing to capture trend and seasonality in data. It's particularly useful when dealing with data that exhibits both trends (consistent upward or downward movements) and seasonal patterns (repeating cycles) It consists of levels, trends, and seasonality [11]. The technique used was based on a decomposition approach, the model was fitted into the training and testing datasets. The Holt-Winters forecasting algorithm allows users to smooth a time series and use that data to forecast areas of interest. Exponential smoothing assigns exponentially decreasing weights and values against historical data to decrease the value of the weight for the older data. In other words, more recent historical data is assigned more weight in forecasting than the older results.

3. Materials and Methods

3.1. Study Design

The study is cross- sectional study design that employed quantitative approaches. The quantitative research method was used to collected data on the new and relapse TB (bacteriologically confirmed or clinical diagnosis) the forecasted from the existing data after running the Autoregressive Integrated Moving Average (ARIMA), Holt-Winters Exponential Smoothing (HWES) and Long Short-Term Memory (LSTM) Time Series Models undependably for purpose of evaluating the effectiveness and efficiency of each model.

3.2. Study Setting

The study was conducted in Moshi municipality, Kilimanjaro region Northern part of Tanzania.. According to 2022 National Population Census the region has an estimated population of 1,861,9 [12]. In 2023 Ministry of Health, National TB and Leprosy Program (NTLP) reported Kilimanjaro region to have TB case notification of 3694 and bacteriological confirmation of 113.7 [13]. The region was selected due to the presence of the Kibong'oto Infectious Diseases Hospital (KIDH) which is the national hospital and center for excellence for Infectious Diseases including Tuberculosis diagnosis and treatment. The hospital is used as the national referral hospital for Tuberculosis especially Multidrug resistance (MDR) TB thus the hospital consist of the all Time Series models used in this study for forecasting the new and relapse TB and high number of the TB patients for required sample.

3.3. Study Population and Sampling Strategy

The dependent variable is the new and relapse TB cases, which are newly registered cases in the TB Information System (TBIS) (DHS2) of Kibong'oto Infectious Disease Hospital KIDH) Moshi, Kilimanjaro region, Tanzania from January 2023 to December 2023. A muilt-stage sampling that involves Simple randomly sampling and Purposive sampling was used in this study to obtain the participants. Simple randomly sampling was used to select the TB patients from the data Based of the January 2015 to 2020 while the purposive sampling was used to selected the system administrators who were used during the in - depth interview. Using the precision formulae of calculating sample size A total of the 72 TB patients who were either bacteriological or clinical TB confirmed were included in the study.

3.4. Data Collection Tools and Procedures

The data collected from the database of care2x Integrated Hospital Information System and District Health Information System (DHIS2) All selected new and relapse (bacteriologically or clinically confirmed) TB cases of January 2015 to December 2020 from the database were recorded and used to create time series database using Microsoft Excel 2019. The created time series contain a total of 2 columns and 72 rows. The aggregated data have two-timed specified seasonality for the forecasting model to perform better. To maintain the Reliability and Validity of the collected data the triangulation techniques whereby the data from multiple sources were obtained.

3.5. Data Processing, Analysis and Interpretation

All collected quantitively data entered in Microsoft Excel 2019. The null values were found after viewing a dataset in a Googlecolab editor and dropped using dropna function from python programming language. All dataset was divided into training and testing datasets as per Xu & Goodacre (2018) Tables and figures were used to illustrates the results from the analysis. Xu & Goodacre (2018) cutoff points will be used as the refrence for evaluation of the efficiency and effectiveness of the time series models to obtain the optimal Time series model. According to the Xu & Goodacre (2018) 70% training, 30% test or 80% training, and 20% test split of the dataset is generally accepted in determining the training and test datasets to obtain the efficiency and effectiveness of the model. [14] For this study 69% of the dataset was used for training and 31% for testing in order to avoid model overfitting.

3.6. Ethical Considerations

The ethical approval of this study was sought from Kibong'oto Infectious Diseases Hospital, the Nelson Mandela Institution of Science and Technology, and the Centre for Educational Development in Health, Ethical Research Committee. (KNCHREC). The permission to conduct the study was obtained from the management KIDH To maintain Confidentiality the aggregated data taken from the database were not contain personal details such as patient name, identification number, address, and contact information.

4. Results

4.1. Number of the TB Cases Registered

A total of 3911 new and relapse TB cases were reported to KIDH from January 2015 to December 2020 with an annual average of 651.83 TB cases (See Table 1). There were irregular variations of TB cases however by the end of 2017 there was a sudden increase which is followed by a sudden decline in early 2018. There was a periodic variation from 2018 to 2019 while 2020 shows some irregular variations. (See Figure 1)

Table 1. *Kibong'oto Infectious Disease Hospital monthly new and relapse TB cases from January 2015 to December 2020.*

Month\Year	2015	2016	2017	2018	2019	2020
January	51	55	30	56	57	56
February	42	84	46	41	61	55
March	57	56	38	34	51	48
April	43	43	36	41	43	58
May	53	87	46	53	42	39
June	50	62	51	54	53	45
July	54	56	49	55	46	59
August	61	71	62	65	66	40
September	57	45	58	52	70	53
October	64	49	88	66	63	29
November	67	52	67	68	67	56
December	69	39	58	67	68	38

4.2. Evaluation of the Selected Time Series Models to Adopt Optimal Model for Forecasting New and Relapse TB Cases

To find the best model more than one error metric values MAE, MAPE, and RMSE was applied are shown in Table 2. Based on all criteria the interpretation score the result was better in ARIMA model while the scores for than LSTM and Holt Winter model was low. From the results interpretation the performance of ARIMA on forecasting the New and relapse TB case was shown to be high compared to the LSTM and Holt Winter model. Therefore the results adopt the ARIM as the optimal model for n forecasting the New and relapse TB compared to other models (see Figure 2).

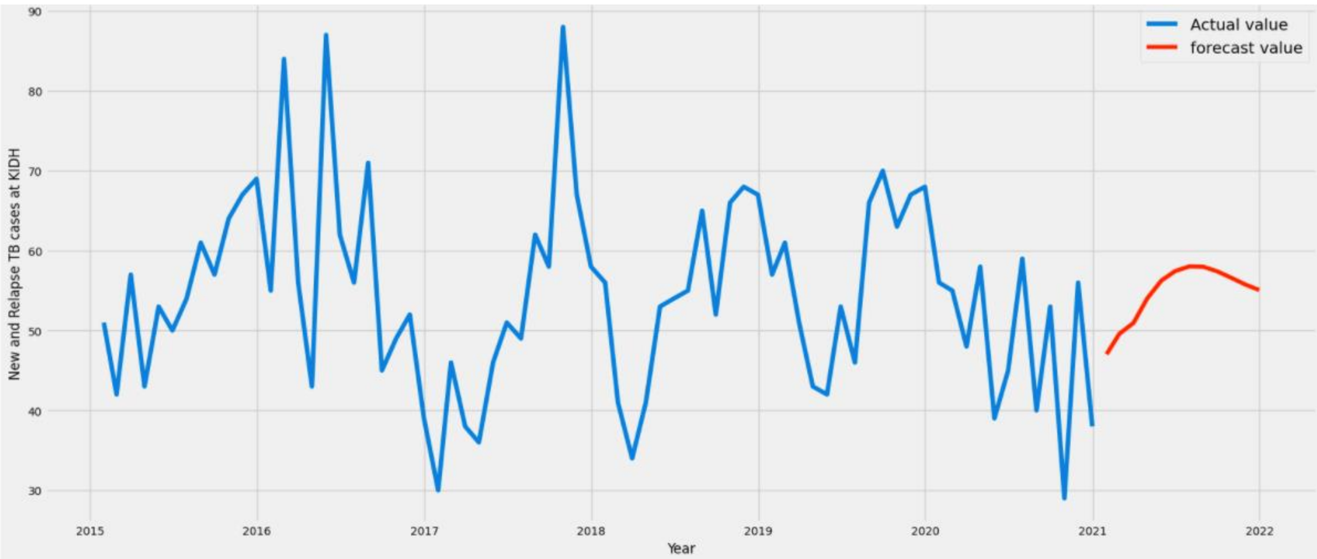

Figure 2. *The actual value of TB cases from January 2015 to December 2020 and the forecasted value from January to December 2021.*

Table 2. Evaluation of the optimal of the selected Time Seres Model on Forecasting New and Relapse TB cases using the criteria 11.734, 0.186 and 15.84.

Model	MAE	MAPE	RMSE
Holt-winter's	11.734	0.186	15.84
ARIMA (4,0,1)	8.79	0.163	10.375
LSTM	7.0079	0.1513	2.9576

5. Discussion

The results shown the irregular variations of TB cases occurrence of the New and relapse TB cases. By the end of 2017 there was a sudden increase which is followed by a sudden decline in early 2018. There was a periodic variation from 2018 to 2019 while 2020 shows some irregular variations. The results from the study was similar to the results conducted by Alshaikh A. Shokeralla in Sudan on Hybrid Time Series-Regression Model for Tuberculosis Forecasting in Resource-Limited Settings Using the ARMA time series model the study results shown the variations of the occurrence of the TB cases 2018 to 2022 [15, 16]. Another study conducted by Mohd Ariff Ab Rashid et al. 2023 in In Malaysia also shown the variation od the occurrence of the TB cases from January from January 2013 to June 201820, [17] This implies that early forecasting of the TB cases is needed for purpose of the planning and making decision on resources for diagnosis and treatment.

On the adoption of Time series modal for Forecasting the new and relapse TB models the study shown the ARMA Time series modal was the optimal model for forecasting the TB cases due to high scoring of the MAE, MAPE, and RMSE that intemperate its high performance. ARMA Save as important model for early warning of the occurrence of the TB cases. The TB dataset in ARMA shown the seasonality, and residual behavior which is early sign for forecasting the occurrence of TB. The findings from various studies was similar to the study conducted in china by Wang H, Li Y. 2021 that compare the effectiveness and efficient of the ARM A and LSTM model Time series modal on the forecasting the occurrence of the influenza. The results from the study was shown the ARMA f performed well forecasting the disease compared to the LSTM. [16]. Another study conducted by Alshaikh A. Shokeralla (2025) in Sudan on Hybrid Time Series-Regression Model for Tuberculosis Forecasting shown the ARMA Time series modal is better model to be sues for forecasting the TB cases in the limited resource setting compared to the other models [15]. This implies that the ARMA time series model is the universal model for forecasting the TB cases compared to other models.

6. Conclusions

The ARIMA Time series model is the best forecasting model for new TB cases in Tanzania. This model offers advanced predictions of new and relapse TB cases, providing valuable guidance for timely planning of prevention and control measures. Having advanced knowledge into the spatial distribution of new TB cases in Tanzania for the upcoming months would significantly assist in precisely directing control measures across the nation. Thus the study recommend the adoption of the ARIMA rather than other models for improving the treatment and diagnosis of the TB cases.

Abbreviations

AIDS	Acquired Immunol Deficiency Syndrome
ARIMA	Autoregressive Integrated Moving Average
COVID	Corona Virus Diseases
HIV	Human Immune Virus
HWES	Holt-Winters Exponential Smoothing
LSTM	Long Short-Term Memory

Acknowledgments

We acknowledge the funders and who contribute the efforts on accomplished this research.

Author Contributions

Eunice Silas: designed the study, collected data, analyzed and wrote the manuscript

Sanket Pandhare: Conceptualize the study

Anael Sam: Conceptualize the study

Devotha Nyambo: Conceptualize the study

Hamimu Omary Kigumi: Conceptualize the study

Funding

The author(s) received funds from Tanzania Ministry of Education, Science, Technology and Vocational Training (MoEVET).

Data Availability Statement

The availability of data from this study is open access to published journal with copy rights from the primary authors of the study.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] WHO report(2023). Global Tuberculosis's report 2023.

[2] Ministry of Health, National TB and Leprosy Program, 2022.

[3] Musa EO, Satti AM. ARIMA Modeling of Tuberculosis Incidence in Sudan. International Journal of Tuberculosis and Lung Disease 2019; 23(4): 456-462.

[4] Abdalla K, et al. Malaria Incidence Forecasting Using ARIMA Models in Sudan. Malaria Journal 2020; 19: 45.

[5] Cleveland RB, Cleveland WS, McRae JE, Terpenning ISTL. A Seasonal-Trend Decomposition Procedure Based on Loess. Journal of Official Statistics 2018; 6(1): 3-73.

[6] Kim D, Choi S. A Hybrid STL-Regression Model for Tuberculosis Forecasting. Journal of Public Health Informatics 2020; 12(2): e1234.

[7] Gomez, K. A. and Gomez, A. A. (1984) Statistical Procedures for Agricultural Research. 2nd Edition, John Wiley and Sons, New York, 680 p.

[8] Hochreiter, Sepp; Schmidhuber, Jürgen (1996-12-03). "LSTM can solve hard long time lag problems". Proceedings of the 9th International Conference on Neural Information Processing Systems. NIPS'96. Cambridge, MA, USA: MIT Press: 473-479.

[9] Felix A. Gers; Jürgen Schmidhuber; Fred Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". Neural Computation. 12 (10): 2451 2471.

[10] Li, Z.-Q., Pan, H.-Q., Liu, Q., Song, H., & Wang, J.-M. (2020). Comparing the performance of time series models with or without meteorological factors in predicting incident pulmonary tuberculosis in eastern China. *Infectious Diseases of Poverty*, 9(1), 1-11.

[11] Dhamodharavadhani, R Rathipriya. Advances in Big Data and Cloud Computing: Proceedings of ICBDCC18, 229-239, 2019.

[12] Tanzania Bureau of Statistics(2022). The Population and Housing Census 2022.

[13] Ministry of Health, National Tuberculosis and Leprosy Programme Annual Report (2023). https://ntlp.go.tz/site/assets/files/1157/ntlp_annual_report_2023_final_1.pdf

[14] Alshaikh A. Shokeralla. A Hybrid Time Series-Regression Model for Tuberculosis Forecasting in Resource-Limited Settings. International Journal of Statistics in Medical Research, 2025, 14, 299-307.

[15] Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, 2(3), 249-262.

[16] Alshaikh A. Shokeralla(2025), "A Comparative Analysis of NNAR and LSTM Models for Short-Term COVID-19 Forecasting in Saudi Arabia", IJSCE, Vol. 15 No. 2 (2025): Volume-15 Issue-2, May 2025.

[17] Mohd Ariff Ab Rashid 1, Rafdzah Ahmad Zaki 1, Wan Rozita WanMahiyuddin 2, Abqariyah Yahya Forecasting New Tuberculosis Cases in Malaysia: A Time-Series Study Using the Autoregressive Integrated Moving Average (ARIMA) Model, 2023 Sep 4; 15(9): e44676. https://doi.org/10.7759/cureus.44676

Biography

Eunice Silas is a Master's student at the Nelson Mandela African Institute of Science and Technology, specialize in Information Systems and Network Security. Also, she is an expert in health digital systems, focus on immunization digital systems.

Research Field

Eunice Silas: Health Information system and Digital Health

Sanket Pandhare: Data Center Management and Cloud computing

Anael Sam: Cybersecurity and Health Information System

Devotha Nyambo: Machine Learning

Hamimu Omary Kigumi: Public Health Researches in Communicable and Non communicable diseases