

FORECASTING MONTHLY TUBERCULOSIS NOTIFICATIONS IN THE PHILIPPINES USING STATISTICAL AND MACHINE LEARNING TIME-SERIES MODELS, 2016-2023

Samuel John E. Parreno

Mathematics Division, UM Digos College, Philippines

Abstract

This study evaluated statistical and machine learning time-series models for forecasting monthly nationwide tuberculosis notifications in the Philippines from January 2016 to December 2023. The dataset contained 96 valid monthly observations after removing one trailing summary row. Total notifications were used as the primary forecasting target, while new, relapse, and other retreatment notifications were used for data verification. A chronological split assigned January 2016 to December 2022 to model training and January 2023 to December 2023 to testing. Seasonal naive, exponential smoothing, seasonal ARIMA, STL-ETS, random forest, XGBoost, and ridge lag regression models were compared using RMSE, MAE, MAPE, and sMAPE. The reported Total column matched the annual workbook summary, although four monthly component checks showed small inconsistencies between Total and the sum of the three registration groups. SARIMA (0,1,2)(0,1,1)[12] with drift achieved the best test performance, with RMSE = 8,123.41, MAE = 7,063.50, MAPE = 13.95%, and sMAPE = 15.32%. Residual diagnostics did not indicate significant autocorrelation in the 2023 test errors. A 24-month projection for 2024-2025 was generated from the final SARIMA model fitted to the full series, with prediction intervals used to express uncertainty. Results support seasonal ARIMA as the most useful model in this dataset, while machine learning lag models require longer series or external predictors for stronger performance.

Keywords:

Tuberculosis Forecasting, Time Series, SARIMA, Machine Learning, Philippines

1. INTRODUCTION

The burden of tuberculosis remains a continuing public health challenge and demands timely use of surveillance data for program planning. Global tuberculosis monitoring emphasizes that case notification data provide a practical record of people diagnosed with tuberculosis and officially reported to health authorities [1], [2]. The End TB Strategy frames surveillance, early diagnosis, patient-centered care, and strengthened health systems as core elements of tuberculosis control [3]. In the Philippines, the National Tuberculosis Control Program provides the operational framework for diagnosis, treatment, recording, reporting, and program implementation across public and private health facilities [4]. In this setting, monthly notifications are not merely administrative counts. They are operational signals that reflect service reach, health-seeking behavior, diagnostic capacity, reporting completeness, and the timing of case registration.

Forecasting is therefore useful for tuberculosis programs because it translates routinely reported time series into anticipatory information. Reliable forecasts can support medicine and laboratory supply planning, staffing, intensified case finding schedules, regional performance reviews, and early identification

of unusual deviations from expected notification patterns. Forecasting does not replace epidemiologic investigation, clinical judgment, or direct measurement of incidence, but it can help programs decide whether observed counts are broadly consistent with prior patterns or whether additional review is warranted. This is especially relevant in tuberculosis because the infection has a large reservoir of latent infection and may respond slowly to changes in transmission, diagnostic access, and social conditions [5].

Monthly tuberculosis notifications can also contain seasonality and abrupt disruptions. Reviews and national notification studies have documented recurring seasonal patterns in tuberculosis notifications in different settings [6]-[8]. At the same time, pandemic-related disruption showed that notification counts may shift sharply when service access, mobility, or health system priorities change. Global analyses described large drops in tuberculosis diagnoses during 2020 and 2021 [9], while a Philippines-specific study linked shelter-in-place policies to reduced tuberculosis notifications and mortality changes during the COVID-19 period [10]. These features make tuberculosis notification forecasting difficult because a model must balance trend, seasonality, and sudden structural changes in a short monthly series.

Time-series forecasting offers a range of tools for this problem. Classical approaches such as ARIMA and exponential smoothing remain central because they model autocorrelation, trend, seasonality, and forecast uncertainty using a relatively small number of observations [11]-[15]. Decomposition-based approaches such as STL can separate seasonal and trend components before forecasting the seasonally adjusted component [16]. Decomposable regression approaches such as Prophet have also been widely discussed for analyst-guided forecasting tasks at scale [17]. Machine learning approaches such as random forests and gradient-boosted trees can learn nonlinear relationships from lagged features, but they depend heavily on the amount of training data, the quality of engineered predictors, and the forecasting design [18], [19]. Forecast comparisons should therefore be made with multiple accuracy measures and with attention to residual diagnostics [20]-[24].

Tuberculosis forecasting studies have used both classical and hybrid approaches. ARIMA and hybrid ARIMA models have been applied to pediatric tuberculosis incidence in Kenya [25]. Hybrid seasonal prediction models, SARIMA-neural network models, empirical mode decomposition hybrids, Elman neural networks, and neural network time-series approaches have also been explored for tuberculosis incidence in China and other settings [26]-[31]. These studies show that no single model family can be assumed to dominate across all tuberculosis datasets. Model choice depends on series length, seasonal structure,

presence of disruption, forecast horizon, and whether external predictors are available.

The objective of this study was to evaluate statistical and machine learning time-series models for forecasting monthly nationwide tuberculosis notifications in the Philippines from 2016 to 2023. The analysis used Total notifications as the primary forecasting target, verified the relationship between Total and the registration-group components, compared several forecasting approaches on a held-out 2023 test period, assessed residual behavior for model adequacy, and generated a final 24-month projection using the best-performing model. The analysis was designed as a reproducible forecasting case study for tuberculosis program planning and surveillance support.

2. MATERIALS AND METHODS

2.1 DATASET AND PREPROCESSING

The study used a monthly Excel dataset of nationwide tuberculosis notifications in the Philippines covering January 2016 to December 2023. The workbook contained the columns Time, New, Relapse, Other Retreatments, and Total. The Time column was converted to monthly dates and sorted in ascending order. One trailing summary row with a blank date was removed before time-series modeling. The cleaned analytic dataset contained 96 monthly observations from January 2016 to December 2023. The reported Total series was used as the primary outcome because it corresponds to the overall monthly notification count. The New, Relapse, and Other Retreatments columns were retained for validation and descriptive context.

Data validation proceeded in two steps. First, the time index was checked for monthly continuity from January 2016 to December 2023. Second, the reported Total was compared with the sum of New, Relapse, and Other Retreatments for each month. The workbook summary total was also compared with the sum of the 96 monthly Total values. The monthly time index was complete, and the sum of the monthly Total values was 3,100,935, matching the workbook's trailing summary value. Four monthly rows showed differences between the reported Total and the sum of the three registration groups, with a maximum absolute difference of 220 notifications. Because the Total column matched the workbook summary and represented the stated primary outcome, the reported Total values were retained as authoritative for forecasting.

A chronological train-test split was used. The training period consisted of 84 monthly observations from January 2016 to December 2022. The test period consisted of 12 monthly observations from January 2023 to December 2023. This split preserved the ordering of the time series and evaluated each model on a full annual seasonal cycle. The test set was not used to fit the candidate models. This design is appropriate for forecasting because randomly splitting months would leak future information into the training data and would not represent an operational forecasting situation.

2.2 STATIONARITY ASSESSMENT AND CANDIDATE MODELS

Stationarity was assessed to guide differencing decisions for the SARIMA model. The Augmented Dickey-Fuller test applied

to the level Total series produced a p-value of 0.576, which did not support stationarity. First differencing alone also did not produce clear stationarity at the 5% level, with $p = 0.165$. After applying both first differencing and seasonal differencing at lag 12, the p-value was 0.038. These results supported the use of a model with one nonseasonal difference and one seasonal difference for the SARIMA component. Other model families were fitted directly to the training series or to log-transformed lag features, depending on their design.

Seven candidate forecasting approaches were evaluated. The first was a seasonal naive benchmark in which each 2023 monthly forecast equaled the observed value from the corresponding month in 2022. This benchmark is important because any useful seasonal model should outperform a simple repetition of the previous seasonal cycle. The second was an exponential smoothing model with damped additive trend and multiplicative monthly seasonality. Exponential smoothing is suitable for level, trend, and seasonal patterns and often performs well in short forecasting tasks [14], [15]. The third was a SARIMA model with monthly seasonality. A limited seasonal grid search on the training period identified SARIMA (0,1,2) (0,1,1) [12] with drift as the lowest-AIC candidate among the fitted SARIMA specifications. This model was then used for the test forecast and for the final full-sample projection.

The fourth model was STL-ETS. Seasonal-trend decomposition using loess was applied with a 12-month seasonal period, and an ETS model was fitted to the seasonally adjusted component [16]. The fifth model was random forest regression using engineered lag features. The sixth model was XGBoost regression, also using lag features. The seventh model was ridge lag regression as a regularized linear autoregressive machine learning comparator. For the three lag-regression models, the response variable was $\log(\text{Total})$. Predictors included lags 1 to 12 of $\log(\text{Total})$, a trend index, month number, sine and cosine monthly seasonal terms, and rolling means over the previous 3, 6, and 12 months. Forecasts for the 12-month test period were generated recursively, so predicted values were used as lag inputs once the forecast horizon moved beyond the available training observations. This recursive approach reflects the real forecasting task more closely than using observed future lags.

The Prophet model was considered because it is a common decomposable time-series approach for trend and seasonality [17]. It was not included in the final fitted comparison because the analysis environment did not include the Prophet package. Long short-term memory networks were also not fitted because the cleaned series contained only 96 observations, and a deep recurrent model would be difficult to train and validate robustly on such a short monthly series. The decision favored transparent and reproducible models whose estimation requirements were more consistent with the available data.

2.3 EVALUATION METRICS, DIAGNOSTICS, AND FINAL FORECAST

Model performance on the 2023 test period was measured using root mean squared error, mean absolute error, mean absolute percentage error, and symmetric mean absolute percentage error. RMSE gives greater weight to large errors, MAE reports the average absolute error in notification counts, MAPE expresses average absolute error as a percentage of

observed values, and sMAPE provides a scale-free percentage measure that is less asymmetric than conventional MAPE when forecasts and observations differ substantially [20]. These metrics were interpreted together rather than in isolation because each measure emphasizes different properties of the forecast errors.

Residual diagnostics were summarized for the 2023 test residuals, defined as observed values minus predicted values. For each model, the mean error, residual standard deviation, Ljung-Box Q test p-value at lag 6, and Shapiro-Wilk p-value were computed. The Ljung-Box test was used as a check for remaining autocorrelation in the residual series [21]. Because the test period contained only 12 residuals, the diagnostics were interpreted cautiously and used as supplementary evidence rather than definitive proof of adequacy. The final model was chosen primarily by out-of-sample accuracy, with residual diagnostics used to check whether the best model left an obvious serial pattern in the test errors.

The final forecast was generated by refitting the selected SARIMA specification to all 96 monthly observations from January 2016 to December 2023. A 24-month forecast was then produced for January 2024 to December 2025. Prediction intervals were taken from the SARIMA state-space forecast distribution and were reported as 95% intervals. These intervals represent model-based uncertainty under the fitted specification and should be interpreted as conditional on the historical data and model assumptions.

3. RESULTS AND DISCUSSION

Table.1. Dataset Summary and Train-Test Split

Item	Value
Cleaned monthly observations	96
Study period	January 2016 to December 2023
Primary outcome	Monthly Total TB notifications
Training period	January 2016 to December 2022
Training observations	84
Test period	January 2023 to December 2023
Test observations	12
Sum of monthly Total	3,100,935
Workbook trailing summary Total	3,100,935
Component-check mismatches	4 months, maximum absolute difference = 220

The Table.1 summarizes the cleaned dataset and the train-test split. The analytic series contained 96 monthly observations and no missing monthly dates. The training set covered 87.5% of the series, while the 2023 test set covered the remaining 12.5%. Across the full period, Total notifications summed to 3,100,935. The mean monthly Total was 32,301.41 notifications, with a minimum of 10,905 in April 2020 and a maximum of 59,274 in March 2023. These descriptive values show the combined influence of long-term growth, pandemic-period disruption, and post-disruption recovery in the notification series.

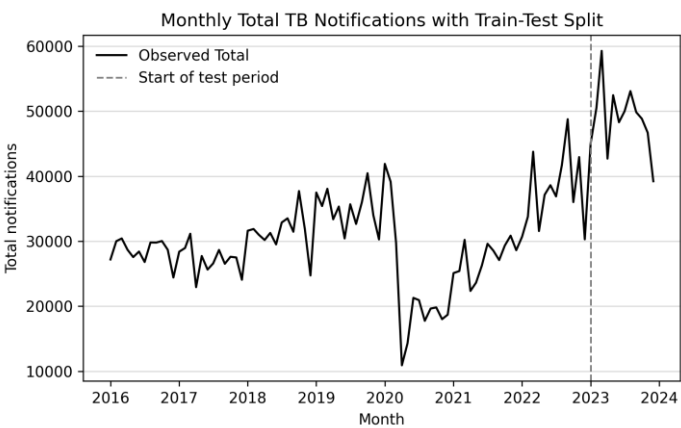


Fig.1. Monthly Total TB notifications with train-test split

The Fig.1 shows the monthly Total notification series with the test-period boundary. Notifications fluctuated between approximately 23,000 and 42,000 before the pandemic disruption, fell sharply in 2020, and then recovered during 2021 and 2022. The 2023 test period began at a higher level than most of the pre-pandemic series and included the highest monthly count in the dataset. This pattern created a difficult forecasting problem because models trained only through 2022 had to extrapolate into a period of stronger notification volume. The full annual totals further illustrate the disruption and recovery: Total notifications declined from 419,085 in 2019 to 272,059 in 2020, a decrease of 35.08%, and then increased to 585,842 in 2023, which was 39.79% higher than 2019 and 71.45% higher than 2016.

Table.2. Model Specifications

Model	Specification
Seasonal Naive	Monthly seasonal naive benchmark with $m = 12$.
ETS	Damped additive trend exponential smoothing with multiplicative monthly seasonality.
SARIMA	SARIMA (0,1,2)(0,1,1)[12] with drift, selected by AIC from a limited seasonal grid.
STL-ETS	STL decomposition with 12-month seasonality and damped-trend ETS on the adjusted component.
Random Forest	Regression on log-lag features, month seasonality, trend, and rolling means; recursive forecasts.
XGBoost	Regularized gradient-boosted tree regression using the same lag and seasonal features.
Ridge Lag Regression	Regularized linear regression on log-lag, seasonal, trend, and rolling-mean features.

The Table.2 lists the model specifications. The specification set deliberately included a simple seasonal benchmark, established statistical models, a decomposition model, and lag-based machine learning models. This design made it possible to determine whether the additional flexibility of machine learning improved practical forecasting performance in a short monthly tuberculosis notification series. The lag-based models used only information that would have been available at the forecast origin and therefore avoided using observed 2023 values as predictors.

Table.3. Forecast Accuracy Metrics on the 2023 Test Set

Model	RMSE	MAE	MAPE (%)	sMAPE (%)
SARIMA	8,123.41	7,063.50	13.95	15.32
STL-ETS	10,326.60	9,855.19	20.07	22.51
ETS	11,293.33	10,679.89	21.54	24.39
Seasonal Naive	12,036.03	11,146.25	22.78	26.23
Random Forest	14,921.67	13,907.74	27.71	32.71
XGBoost	15,511.75	14,631.75	29.31	34.84
Ridge Lag Regression	19,380.34	18,624.67	37.83	47.40

The Table.3 presents the 2023 test-set accuracy metrics. SARIMA was the best-performing model across all four metrics, with RMSE = 8,123.41, MAE = 7,063.50, MAPE = 13.95%, and sMAPE = 15.32%. STL-ETS ranked second, with RMSE = 10,326.60 and MAPE = 20.07%. ETS and the seasonal naive benchmark had higher errors, with MAPE values of 21.54% and 22.78%, respectively. The machine learning lag models performed worse in this split: random forest had MAPE = 27.71%, XGBoost had MAPE = 29.31%, and ridge lag regression had MAPE = 37.83%. The result does not imply that machine learning models are unsuitable for tuberculosis forecasting in general. It indicates that, in this particular short monthly univariate series, SARIMA better captured the seasonal differenced structure needed for 12-step ahead forecasting.

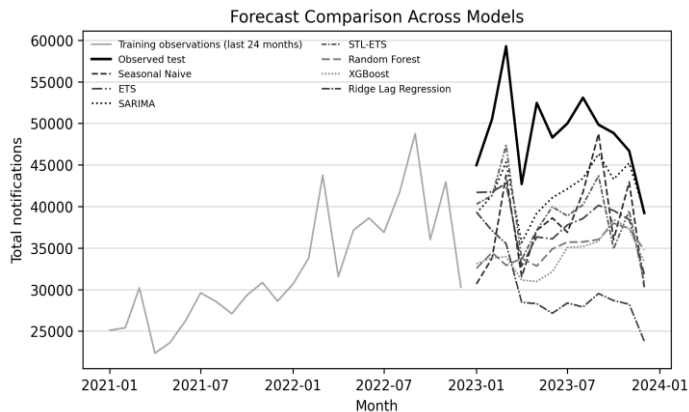


Fig.2. Forecast comparison plot across models

The Fig.2 compares the test-period forecasts across models. The SARIMA and STL-ETS forecasts rose into early 2023 more than the seasonal naive and tree-based models, which helped reduce errors during the high-notification months. The lag-based machine learning models tended to underpredict much of 2023 because they were trained on a series in which the post-2020 rebound was still evolving. Tree-based methods are also constrained in extrapolation when future values exceed many of the patterns seen in training. This behavior explains why random forest and XGBoost, although flexible in nonlinear settings, did not outperform SARIMA in this dataset.

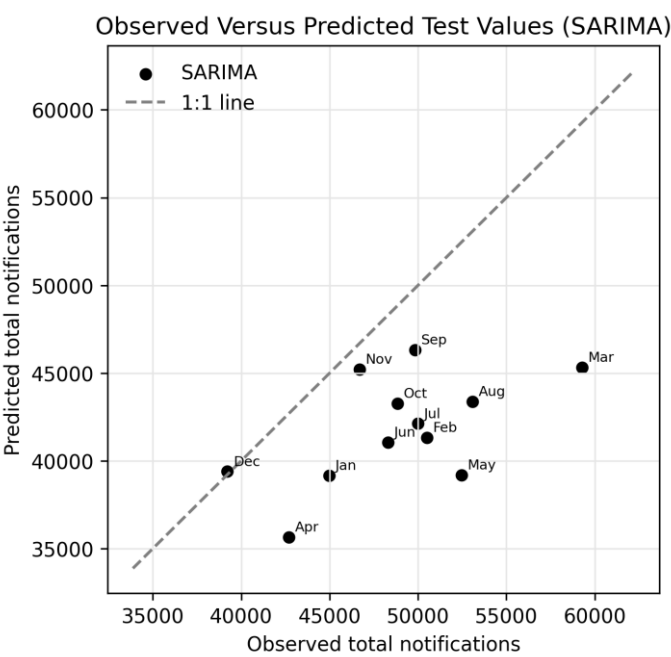


Fig.3. Observed versus predicted values for the test period

The Fig.3 plots observed versus predicted values for the best-performing model during the 2023 test period. Points close to the diagonal line indicate better agreement between observed and predicted values. The SARIMA predictions generally followed the relative ordering of the 2023 months, but they remained lower than the observed values during the highest notification months. This is consistent with the positive mean error shown in Table 5 and reflects the fact that the 2023 notification level was stronger than the trajectory estimated from 2016 to 2022. Forecasting performance was therefore acceptable for a short operational series but not exact enough to support uncritical month-by-month planning without interval-based uncertainty.

Table.5. Residual Diagnostic Summary

Model	Mean Error	Residual SD	Ljung-Box Q(6) p	Shapiro-Wilk p
SARIMA	7,033.98	4,244.29	0.290	0.933
STL-ETS	9,855.19	3,221.62	0.025	0.956
ETS	10,679.89	3,834.67	0.509	0.741
Seasonal Naive	11,146.25	4,743.67	0.410	0.168
Random Forest	13,907.74	5,646.97	0.974	0.888
XGBoost	14,631.75	5,379.40	0.856	0.987
Ridge Lag Regression	18,624.67	5,597.34	0.588	0.218

The Table.5 summarizes residual diagnostics for the test errors. SARIMA had a mean error of 7,033.98 notifications and a residual standard deviation of 4,244.29. The Ljung-Box Q(6) p-value for SARIMA residuals was 0.290, which did not indicate statistically significant residual autocorrelation at the 5% level. The Shapiro-Wilk p-value was 0.933, which did not indicate strong departure from normality among the 12 test residuals. STL-ETS had a Ljung-Box p-value of 0.025, suggesting remaining serial pattern in the test errors, while the other models had non-

significant Ljung-Box p-values but larger forecast errors. The diagnostics support SARIMA as the most balanced option because it combined the lowest forecast error with no clear residual autocorrelation signal.

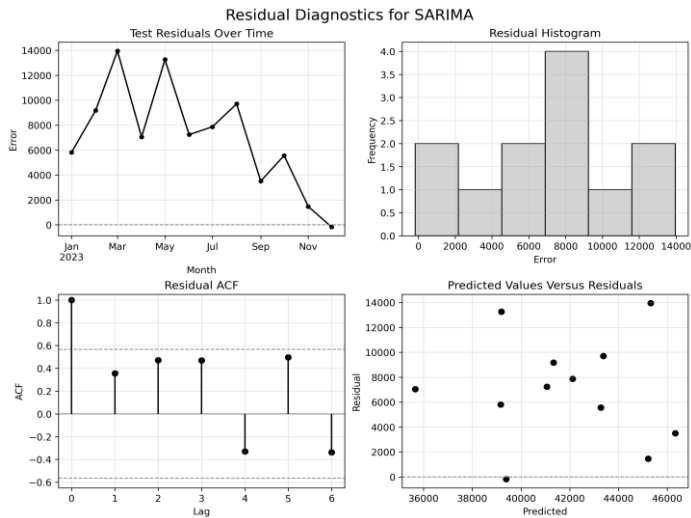


Fig.4. Residual diagnostic plot for the best-performing model

The Fig.4 provides a visual diagnostic view of the SARIMA residuals. The residual time plot shows mostly positive errors, confirming that the model underpredicted several 2023 months, but the residual autocorrelation function did not show a strong systematic pattern beyond the approximate confidence bounds. The histogram and residual-versus-predicted plot do not suggest a severe outlier-driven result. These diagnostics are important because a model with low error but highly patterned residuals would be less attractive for operational forecasting. In this study, the residual behavior of SARIMA was acceptable given the short held-out period.

Table.4. Final Selected Model Forecast Values

Month	Forecast	Lower 95% PI	Upper 95% PI
Jan 2024	50,503	39,656	61,350
Feb 2024	53,705	40,372	67,037
Mar 2024	59,918	44,751	75,085
Apr 2024	47,842	31,040	64,644
May 2024	54,099	35,808	72,391
Jun 2024	53,989	34,321	73,658
Jul 2024	55,580	34,625	76,536
Aug 2024	57,886	35,718	80,054
Sep 2024	58,824	35,507	82,141
Oct 2024	56,742	32,330	81,155
Nov 2024	57,514	32,053	82,974
Dec 2024	51,293	24,826	77,760
Jan 2025	62,099	33,290	90,909
Feb 2025	65,400	34,840	95,960
Mar 2025	71,802	39,649	103,955
Apr 2025	59,914	26,244	93,585
May 2025	66,360	31,237	101,483

Jun 2025	66,439	29,922	102,956
Jul 2025	68,219	30,358	106,079
Aug 2025	70,713	31,555	109,870
Sep 2025	71,839	31,426	112,252
Oct 2025	69,946	28,316	111,577
Nov 2025	70,906	28,092	113,720
Dec 2025	64,874	20,909	108,839

The Table.4 reports the final selected model forecast values for January 2024 to December 2025 after refitting SARIMA to the full 2016–2023 series. The projected monthly means ranged from 47,842 in April 2024 to 71,839 in September 2025. The forecast preserved a seasonal pattern with higher values around the first and third quarters and lower values toward December. The intervals widened across the 24-month horizon, which is expected because uncertainty accumulates as forecasts extend further beyond the observed series.

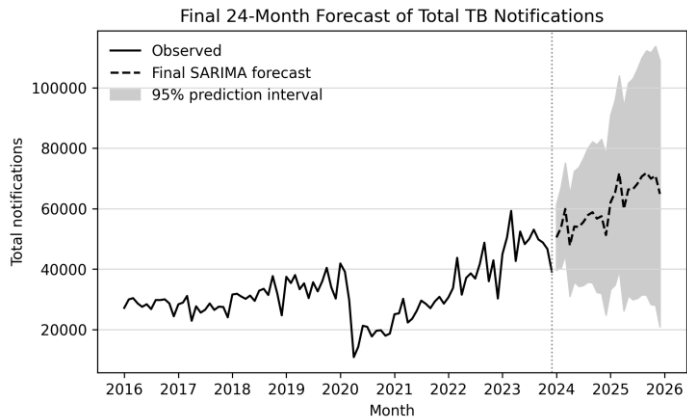


Fig.5. Final forecast plot with prediction intervals

The Fig.5 displays the final 24-month forecast with 95% prediction intervals. The interval band is wide, especially in the second year, indicating that the model should be used as a planning guide rather than as a deterministic prediction. The projected path suggests that if recent reporting patterns continued under the fitted SARIMA structure, monthly notifications could remain above most pre-pandemic levels. For program planning, the central forecast can be used for scenario preparation, while the lower and upper intervals provide a more cautious range for resource planning.

The overall evidence favors SARIMA for this dataset. Its advantage is plausible because the monthly series contains a clear seasonal component, a large disruption, and limited observations for training complex models. Statistical models that explicitly difference the series and model seasonal autocorrelation can perform well when the main information lies in past values and seasonal structure. Machine learning models often become more competitive when more observations, richer covariates, spatial detail, intervention indicators, climate variables, search behavior signals, or health service indicators are available. In the present univariate setting, their additional flexibility did not translate into better out-of-sample performance.

The findings are consistent with the broader tuberculosis forecasting literature, where ARIMA, SARIMA, and hybrid approaches are repeatedly used because tuberculosis data are

often available as surveillance time series with monthly or quarterly aggregation [25]-[31]. Hybrid and neural approaches may improve accuracy in some contexts, but their benefit depends on sufficient data and careful validation. For Philippine national TB notifications from 2016 to 2023, a parsimonious SARIMA model provided the most useful balance of accuracy, interpretability, residual behavior, and forecast uncertainty.

Several limitations should be considered. First, the analysis used national monthly aggregates, so it cannot capture regional heterogeneity, local outbreaks, urban-rural differences, laboratory access, or private-sector reporting variation. Second, the series was short for machine learning and did not include external predictors such as screening campaigns, policy changes, testing capacity, mobility, holidays, or health facility reporting completeness. Third, 2020 represented a major system disruption, and the model treated it as part of the historical pattern rather than as a formally modeled intervention. Fourth, notification counts are not the same as true incidence because they depend on diagnosis and reporting. Future work can extend the analysis using regional panels, covariate-augmented models, intervention indicators, rolling-origin cross-validation, and probabilistic forecast evaluation.

The comparison with prior tuberculosis forecasting studies reinforces the need to treat model selection as data-specific. The present SARIMA result is consistent with studies where seasonal ARIMA structures captured recurring tuberculosis notification patterns effectively, but it differs from settings where hybrid models improved performance after nonlinear residual modeling. For example, hybrid tuberculosis forecasting work in China used ARIMA-neural or decomposition-based combinations to improve incidence prediction [27], [28], [29]. Zheng et al. used time-series and Elman neural network methods for tuberculosis incidence prediction in Kashgar, China [30], while Orjuela-Cañón et al. evaluated neural network models for tuberculosis incidence forecasting [31]. These studies support the methodological value of machine learning, but they also show why adequate sample size and validation design matter. In the Philippine national series analyzed here, the training set had only 84 monthly observations before the 12-month test period, so the lag-based machine learning models had limited historical information from which to learn nonlinear post-disruption dynamics.

The seasonal structure also deserves attention. Studies of tuberculosis seasonality have shown that the timing and amplitude of tuberculosis notifications vary across settings, climates, and reporting systems. Li et al. reported seasonal variation in active tuberculosis notifications in China [7], while Fares reviewed biological and social explanations for seasonal tuberculosis patterns [6]. The Philippine series did not merely show smooth seasonality; it also contained the 2020 notification collapse and a strong 2022-2023 rebound. These features explain why the seasonal naive benchmark, although useful as a minimum comparator, was not sufficient. The model had to account for seasonal recurrence and changing level at the same time. SARIMA addressed this through nonseasonal and seasonal differencing, a strategy aligned with standard time-series modeling guidance [13].

From a forecasting-practice perspective, the results emphasize the importance of benchmark comparison, rolling or holdout validation, and multiple error measures. Large forecasting

competitions and methodological reviews have repeatedly shown that simpler statistical models can remain competitive, particularly when the data are short, noisy, or mostly univariate [22], [23]. Machine learning can add value when richer predictors and longer histories are available, but flexible models can underperform when they cannot extrapolate beyond the range of training patterns. The XGBoost and random forest results in this study illustrate this point: both models used nonlinear lag features, yet both underpredicted the elevated 2023 test period. This does not weaken the role of soft computing in disease surveillance. Rather, it clarifies that soft computing models should be matched with sufficient data, meaningful predictors, and validation procedures appropriate to the forecast horizon.

4. CONCLUSION

This study compared statistical and machine learning time-series models for forecasting monthly nationwide tuberculosis notifications in the Philippines from 2016 to 2023. After cleaning the workbook and removing one trailing summary row, the analysis used 96 monthly observations and evaluated models on a full 2023 holdout period. Among the fitted models, SARIMA (0,1,2)(0,1,1)[12] with drift achieved the lowest test errors across RMSE, MAE, MAPE, and sMAPE. Residual diagnostics did not indicate significant autocorrelation in the SARIMA test errors, although the model tended to underpredict the high 2023 notification level.

The final SARIMA projection for 2024-2025 suggests that monthly notifications may remain above most pre-pandemic levels if recent patterns continue, but the prediction intervals are wide and should guide cautious planning. For this dataset, SARIMA is the most useful forecasting approach because it combines accuracy, interpretability, and interval estimation. Machine learning lag models did not outperform the statistical models in this short univariate series, but they remain promising for future work with longer histories and richer predictors.

5. DATA AVAILABILITY

The data analyzed in this study were provided as an Excel workbook containing monthly nationwide tuberculosis notifications for the Philippines from 2016 to 2023. The cleaned analytic dataset, model outputs, and figures generated for this manuscript can be reproduced from the uploaded workbook and the documented preprocessing and modeling procedures. Public sharing of the raw data should follow the permissions and restrictions attached to the source dataset.

6. ETHICAL CONSIDERATIONS

This study used aggregated monthly surveillance counts and did not involve individual patient records, identifiable personal information, clinical charts, biological specimens, or direct contact with human participants. The analysis was conducted for secondary statistical modeling and program planning support. Because the data were aggregated at the national monthly level, the risk of individual identification was negligible.

REFERENCES

- [1] World Health Organization, "Global Tuberculosis Report 2024", Available at: <https://www.who.int/publications/i/item/9789240101531>, Accessed in 2024.
- [2] World Health Organization, "Consolidated Guidance on Tuberculosis Data Generation and Use: Module 1: Tuberculosis Surveillance", Available at: <https://www.who.int/publications/i/item/9789240075290>, Accessed in 2024.
- [3] World Health Organization, "The End TB Strategy", Available at: <https://www.who.int/publications/i/item/WHO-HTM-TB-2015.19>, Accessed in 2015.
- [4] Department of Health, "National Tuberculosis Control Program Manual of Procedures", Available at: https://itis.doh.gov.ph/assets/img/downloads/mop/NTP_MOP_6th_Edition.pdf, Accessed in 2019.
- [5] R.M.G.J. Houben and P.J. Dodd, "The Global Burden of Latent Tuberculosis Infection: A Re-Estimation using Mathematical Modelling", *PLoS Medicine*, Vol. 13, No. 10, pp. 1-11, 2016.
- [6] A. Fares, "Seasonality of Tuberculosis", *Journal of Global Infectious Diseases*, Vol. 3, No. 1, pp. 46-55, 2011.
- [7] X.X. Li, L.X. Wang and H. Zhang, "Seasonal Variations in Notification of Active Tuberculosis Cases in China, 2005-2012", *PLoS ONE*, Vol. 8, No. 7, pp. 1-11, 2013.
- [8] Z. Gashu, S. Jerene and G. Ensermu, "Seasonal Patterns of Tuberculosis Case Notification in the Tropics of Africa: A Six-Year Trend Analysis in Ethiopia", *PLoS ONE*, Vol. 13, No. 11, pp. 1-15, 2018.
- [9] J.R. Ledesma, J.R. Ma and M. Vongpradith, "Global, Regional and National-Level Impacts of the COVID-19 Pandemic on Tuberculosis Diagnoses, 2020-2021", *Microorganisms*, Vol. 11, No. 9, pp. 1-7, 2023.
- [10] R. Crowder, D.M. Geocaniga-Gaviola and R.A. Fabella, "Impact of Shelter-in-Place Orders on TB Case Notifications and Mortality in the Philippines during the COVID-19 Pandemic", *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, Vol. 25, pp. 1-13, 2021.
- [11] G.E.P. Box, G.M. Jenkins, G.C. Reinsel and G.M. Ljung, "Time Series Analysis: Forecasting and Control", 2015.
- [12] R.J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice", 2021.
- [13] R.J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The Forecast Package for R", *Journal of Statistical Software*, Vol. 27, No. 3, pp. 1-22, 2008.
- [14] E.S. Gardner, "Exponential Smoothing: The State of the Art", *Journal of Forecasting*, Vol. 4, No. 1, pp. 1-28, 1985.
- [15] E.S. Gardner, "Exponential Smoothing: The state of the Art - Part II", *International Journal of Forecasting*, Vol. 22, No. 4, pp. 637-666, 2006.
- [16] R.B. Cleveland, W.S. Cleveland, J.E. McRae and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure based on Loess", *Journal of Official Statistics*, Vol. 6, No. 1, pp. 3-33, 1990.
- [17] S.J. Taylor and B. Letham, "Forecasting at Scale", *The American Statistician*, Vol. 72, No. 1, pp. 37-45, 2018.
- [18] L. Breiman, "Random forests", *Machine Learning*, Vol. 45, pp. 5-32, 2001.
- [19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, Vol. 8, pp. 785-794, 2016.
- [20] R.J. Hyndman and A.B. Koehler, "Another Look at Measures of Forecast Accuracy", *International Journal of Forecasting*, Vol. 22, No. 4, pp. 679-688, 2006.
- [21] G.M. Ljung and G.E.P. Box, "On a Measure of Lack of Fit in Time Series Models", *Biometrika*, Vol. 65, No. 2, pp. 297-303, 1978.
- [22] S. Makridakis, E. Spiliotis and V. Assimakopoulos, "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods", *International Journal of Forecasting*, Vol. 36, No. 1, pp. 54-74, 2020.
- [23] F. Petropoulos, D. Apiletti and V. Assimakopoulos, "Forecasting: Theory and Practice", *International Journal of Forecasting*, Vol. 38, No. 3, pp. 705-871, 2022.
- [24] N.K. Ahmed, A.F. Atiya, N.E. Gayar and H. El-Shishiny, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting", *Econometric Reviews*, Vol. 29, No. 6, pp. 594-621, 2010.
- [25] S. Siamba, A. Otieno and J. Koech, "Application of ARIMA and hybrid ARIMA Models in Predicting and Forecasting Tuberculosis Incidences among Children in Homa Bay and Turkana Counties, Kenya", *PLOS Digital Health*, Vol. 2, No. 2, pp. 1-17, 2022.
- [26] S. Cao, F. Wang, W.W.S. Tam, L.A. Tse, J.H. Kim, J. Liu and Z. Lu, "A Hybrid Seasonal Prediction Model for Tuberculosis Incidence in China", *BMC Medical Informatics and Decision Making*, Vol. 13, No. 56, pp. 1-17, 2013.
- [27] Z. Li, Z. Wang, H. Song, Q. Liu, B. He and P. Shi, "Application of a Hybrid Model in Predicting the Incidence of Tuberculosis in a Chinese Population", *Infection and Drug Resistance*, Vol. 12, pp. 1011-1020, 2019.
- [28] Y. Wang, C. Xu, Z. Wang, S. Zhang, Y. Zhu and J. Yuan, "Temporal Trends Analysis of Tuberculosis Morbidity in Mainland China from 1997 to 2025 using a New SARIMA-NARNNX Hybrid Model", *BMJ Open*, Vol. 9, pp. 1-13, 2019.
- [29] Y. Wang, C. Xu, Y. Li, W. Wu, L. Gui, J. Ren and S. Yao, "An Advanced Data-Driven Hybrid Model of SARIMA-NNAR for Tuberculosis Incidence Time Series Forecasting in Qinghai Province, China", *Infection and Drug Resistance*, Vol. 13, pp. 867-880, 2020.
- [30] Y. Zheng, X. Zhang, X. Wang, K. Wang and Y. Cui, "Predictive Study of Tuberculosis Incidence by Time Series Method and Elman Neural Network in Kashgar, China", *BMJ Open*, Vol. 11, No. 1, pp. 1-17, 2021.
- [31] A.D. Orjuela-Canon, A.L. Jutinico, M.E. Duarte Gonzalez, C.E. Awad Garcia, E. Vergara and M.A. Palencia, "Time Series Forecasting for Tuberculosis Incidence Employing Neural Network Models", *Heliyon*, Vol. 8, No. 7, pp. 1-19, 2022.