scientific reports

OPEN

A comparative analysis of classical and machine learning methods for forecasting TB/HIV co-infection

André Abade[1✉], Lucas Faria Porto[2,6], Alessandro Rolim Scholze[3,6], Daniely Kuntath[2,6], Nathan da Silva Barros[2,6], Thaís Zamboni Berra[4,6], Antonio Carlos Vieira Ramos[5,6], Ricardo Alexandre Arcêncio[4,6] & Josilene Dália Alves[2,6]

TB/HIV coinfection poses a complex public health challenge. Accurate forecasting of future trends is essential for efficient resource allocation and intervention strategy development. This study compares classical statistical and machine learning models to predict TB/HIV coinfection cases stratified by gender and the general populations. We analyzed time series data using exponential smoothing and ARIMA to establish the baseline trend and seasonality. Subsequently, machine learning models (SVR, XGBoost, LSTM, CNN, GRU, CNN-GRU, and CNN-LSTM) were employed to capture the complex dynamics and inherent non-linearities of TB/HIV coinfection data. Performance metrics (MSE, MAE, sMAPE) and the Diebold-Mariano test were used to evaluate the model performance. Results revealed that Deep Learning models, particularly Bidirectional LSTM and CNN-LSTM, significantly outperformed classical methods. This demonstrates the effectiveness of Deep Learning for modeling TB/HIV coinfection time series and generating more accurate forecasts.

Tuberculosis (TB) and the human immunodeficiency virus (HIV) are two infections that impact the progression of each other and represent significant global challenges for health systems [1]. People living with HIV (PLHIV) are 26 times more likely to develop active TB, have a 48% lower chance of cure, are 50% more likely to abandon treatment, and have a 94% higher chance of death from TB compared to those without coinfection [2].

The impact of TB/HIV coinfection is particularly stark when looking at global statistics. Among all incident cases of TB worldwide in 2022, 6.3% were PLHIV, with TB/HIV coinfection being the most significant contributor to mortality, totaling 167,000 lives lost [3]. Brazil plays a significant role in this scenario, responsible for one-third of TB cases reported in the Americas. It is notably the only country to appear on two World Health Organization (WHO) priority lists: one for TB and another for TB/HIV coinfection [2,3]. In 2023, Brazil reported 80,012 new TB cases, with 7,089 involving TB/HIV coinfection [4].

These epidemiological indicators of TB/HIV coinfection constitute a serious public health problem that has concerned health authorities worldwide. In response to this, during the second United Nations high-level meeting on TB in 2023, global commitments were made for the period of 2023 to 2027 by all Member States. These commitments aim for universal access to TB services for at least 90% of people with or at risk of TB [5]. Furthermore, these commitments include investing in technology and innovation as crucial tools for preventing, diagnosing, and treating TB, with special attention to vulnerable or at-risk individuals, such as PLHIV.

Building on these global initiatives, the present study investigates the problem of TB/HIV coinfection from the perspective of predictive time series modeling, exploring a range of techniques from classical statistical models to machine learning-based approaches. Predictive time series modeling utilizes historical data, statistics, and machine learning algorithms to forecast future events or trends [6].

Despite the significant impact of TB/HIV coinfection, a review of the scientific literature conducted in major databases, such as Scopus, PubMed, and Web of Science, found few studies [7–10] that simultaneously evaluated the epidemiological trend of TB/HIV coinfection using time series analysis.

[1]Federal Institute of Education, Science and Technology of Mato Grosso, Department of Computer Science, Campus Barra do Garças, Barra do Garças, Mato Grosso, Brazil. [2]Department of Health Sciences, Barra do Garças, Campus Araguaia, Federal University of Mato Grosso, Cuiabá, Mato Grosso, Brazil. [3]State University of Northern Paraná, Luiz Meneghel Campus, Bandeirantes, Paraná, Brazil. [4]University of São Paulo College of Nursing at Ribeirão Preto, Ribeirão Preto, São Paulo, Brazil. [5]Nursing Department, University of the State of Minas Gerais, Passos, Minas Gerais, Brazil. [6]These authors contributed equally: Lucas Faria Porto, Alessandro Rolim Scholze, Daniely Kuntath, Nathan da Silva Barros, Thaís Zamboni Berra, Antonio Carlos Vieira Ramos, Ricardo Alexandre Arcêncio and Josilene Dália Alves. ✉email: andre.abade@ifmt.edu.br

Previous research, such as the studies by Santos *et al.* (2022) and Osei, Oppong, and Der (2020), has investigated temporal trends in TB/HIV coinfection. However, these studies maintained a generalist perspective, focusing primarily on the disease's past behavior. Recognizing this gap, our study advances the field by presenting an innovative approach that leverages the power of classical statistics and machine learning to predict future scenarios of TB/HIV coinfection. This forward-looking perspective aims to provide more accurate and actionable insights for public health planning and intervention.

Additionally, the methodological framework employed positions itself as a technological tool aligned with the prerogatives of global health sector strategies for 2022-2030[11,12]. It is also noteworthy that the methods of this study can be easily replicated by government authorities, providing an important resource for health practices. This promotes the planning and implementation of strategies for controlling TB/HIV coinfection.

To further this goal, our study aimed to describe the temporal trend of TB/HIV coinfection in a central region of Brazil and generate forecasts using a range of analytical methods, from classical statistical models to machine learning techniques. The study's contributions include a detailed analysis of TB/HIV notifications in a high-incidence region of Brazil and the exploration of predictive models for time series, demonstrating the viability and suitability of these approaches.

These models provide a foundation for optimizing healthcare systems, enabling accurate predictions that guide efficient resource allocation and proactive responses to emerging health needs. Integrating predictive technologies promotes innovations in healthcare, paving the way for new diagnostic, treatment, and prevention approaches. Incorporating these models into global health strategies improves healthcare service effectiveness and drives a data-oriented approach, essential for achieving the ambitious goals set for 2030 [13].

Methods
Study type and scenario
This is an ecological time-series study, applying predictive modeling, conducted in the state of Mato Grosso, located in the Central-West region of Brazil, characterized by approximate coordinates of 14° to 17° South latitude and 53° to 58° West longitude.

The object of the study is TB/HIV co-infection, which represents a critical area of investigation due to the synergistic interaction between the two diseases, which exacerbates clinical outcomes and complicates treatment protocols. Individuals with HIV are at a significantly higher risk of developing active TB due to their compromised immune systems. This dual burden not only increases morbidity and mortality rates, but also poses substantial challenges for public health systems, particularly in resource-limited settings. Understanding the dynamics of TB/HIV coinfection is essential for developing effective strategies to manage and mitigate the impact of these co-occurring epidemics[4].

Regions with high incidence rates of TB and HIV offer a relevant context for studying coinfection, due to the complex interaction between these two diseases. Mato Grosso, especially in areas of socioeconomic vulnerability, may have significant rates of these conditions, making it an important study site.

The state of Mato Grosso is a major migration corridor and has significant population mobility due to seasonal work in agriculture and other industries. This can affect transmission patterns and the dynamics of TB/HIV coinfection. Additionally, Mato Grosso is characterized by significant population diversity, including dense urban areas, rural communities, and indigenous peoples. This diversity allows for the analysis of the co-infection of TB/HIV in different social, economic, and cultural contexts.

Study population and dataset
The study population consisted of all TB/HIV cases in the state of Mato Grosso that were notified in the Notifiable Diseases Information System (SINAN) between the years 2012 and 2023. The time frame from 2012 to 2023 was selected to encompass a sufficient period for observing meaningful trends and patterns in TB/HIV coinfection incidence, allowing for robust time series analysis. This period also coincides with the availability of comprehensive and high-quality data from the health authorities. Only individuals diagnosed with TB/HIV according to the International Classification of Diseases (ICD-10), considering the code B20.0 for TB/HIV coinfection, were included in the study.

The SINAN is a Brazilian information system maintained by the Department of Informatics of the Unified Health System (DATASUS)[14] responsible for recording and processing information on mandatory reportable diseases across the country[14]. SINAN collects detailed data on various diseases, including TB, covering demographic information (age, sex, race/color, education, occupation), clinical details (signs and symptoms, comorbidities, laboratory results, type of diagnosis), epidemiological data (residence, notification date, symptom onset date), and treatment information (therapeutic regimen, case evolution, treatment outcomes). This database is essential for epidemiological surveillance, allowing the monitoring of disease incidence and prevalence, identifying outbreaks, tracking case progression, and evaluating the effectiveness of public health interventions. Population data were extracted from the 2022 Demographic Census, available on the website of the Brazilian Institute of Geography and Statistics (IBGE)[15].

Data analysis
The variables used in this study include the date of case notification, number of cases, sex, AIDS-related conditions, and the HIV test. Regarding the variables, we included gender-specific stratifications (male and female), and the general population to capture potential differences in TB/HIV coinfection trends across different demographic groups. This stratification helps in understanding the impact of gender on the incidence and progression of the disease, which is crucial for developing targeted public health strategies. Before starting the actual data analysis, it was crucial to understand the variables available in SINAN; this understanding allowed for the removal

of redundancies, ensuring the integrity and quality of the data. After pre-processing with the data properly treated, the monthly TB/HIV incidence rate was computed based on the number of cases and the population, stratified by sex (male and female) considering the period from 2012 to 2023. The following formula was used to calculate the incidence rate of TB/HIV coinfection:

$$\frac{Number\ of\ new\ confirmed\ TB/HIV\ cases\ in\ residents\ of\ Mato\ Grosso\ state}{Total\ number\ of\ residents\ in\ MatoGrosso\ state\ during\ the\ same\ time\ period} \times 100.000\ peoples \quad (1)$$

The resulting incidence expresses the number of new tuberculosis cases per a given number of people at risk during a specific period[16]. This measure is useful to compare incidence rates between different geographical areas, population groups, or over time.

Subsequently, time series were constructed using the computed incidence rate, from the perspective of the total population and stratified by sex. Univariate time series are sets of observations of a single variable collected over time, usually at regular intervals[17]. The main goal of univariate time series analysis is to understand the underlying patterns that govern the variable's variation over time and, based on this, make future predictions.

The STL (Seasonal and Trend Decomposition Using LOESS) method was applied for the decomposition of time series. This is a popular approach for decomposing a time series and its trend, seasonality, and residual (or error) components[18]. This strategy is reported in the literature as successful in identifying patterns and variations in temporal data. Additionally, we calculated the anomaly points in the time series using the method proposed by Ahmad and Purdy (2016), known as Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD)[19].

In this perspective, identifying the trend in a time series is essential to understand the overall behavior of the data. This allows for predicting possible future movements and understanding whether the series is stationary, ascending, or descending over time. Thus, in this study, we used a combination of statistical tests, including the Augmented Dickey-Fuller (ADF)[20], Dickey-Fuller GLS (DF-GLS)[21], KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Stationarity Test[22], Phillips-Perron Test (Z-tau)[23], and Seasonality Test[24] to assess the produced time series. We consider this strategy to be a methodologically robust practice that offers a more comprehensive and accurate understanding of the underlying characteristics of the data.

Each of these tests provides unique and complementary insights, essential for accurate and reliable time series analysis. First, both the ADF and the DF-GLS are crucial for detecting the presence of unit roots, which indicate nonstationarity. However, while the ADF is the standard and widely used test, the DF-GLS offers greater sensitivity in cases where the time series present slow deterministic trends. Combining these two tests allows for a more comprehensive analysis, increasing confidence in the correct detection of the series' trend type.

Additionally, the KPSS offers a complementary perspective by testing the null hypothesis of stationarity. While the ADF and DF-GLS test the null hypothesis of non-stationarity, the KPSS assumes the opposite. The Phillips-Perron Test, in turn, adds robustness to the analysis by adjusting the test statistic for autocorrelation and heteroscedasticity, common issues in time series. This means that even when the basic assumptions of the ADF or DF-GLS tests are violated, the Phillips-Perron test can provide valid results, thus ensuring the analysis is reliable under a variety of conditions. Lastly, the seasonality test is crucial for identifying patterns that repeat at regular intervals, a critical aspect in many time series, especially those related to economic, climatic, and behavioral phenomena. Ignoring seasonality can lead to misinterpretations of the underlying trends and volatilities in the data.

The combination of these tests provides a more balanced and unbiased view of the nature of the time series data, allowing for the identification of trends, patterns, and fundamental statistical characteristics as accurately as possible. This multidimensional approach not only increases the precision of analyses but also strengthens the foundation for informed decision-making and the formulation of more assertive predictions.

Forecasting modeling

Time series predictive modeling is a branch of statistics and data analysis that focuses on using historical information to predict future events or trends[25]. In our study, we explore approaches ranging from classical statistics to machine learning, using a diverse array of analytical methods, such as exponential smoothing (Simple Exponential Smoothing[26], Double Exponential Smoothing[27], Triple Exponential Smoothing (Holt-Winters)[28]), autoregressive models (ARIMA)[29], and machine learning-based models (Support Vector Regression (SVR)[30], Extreme Gradient Boosting (XGBoost)[31], Long Short-Term Memory (LSTM)[32], Gated Recurrent Unit (GRU)[33], Convolutional Neural Network (CNN)[34], to make predictions about TB/HIV coinfection and increase the robustness of the results. If multiple independent methods point to the same trend or forecast, there is implicitly an increase in the reliability of the results achieved.

Thus, by using a combination of methods, especially those that vary in complexity and approach, one can reduce the risk of overfitting to a single set of data characteristics and foster different temporal horizons in modeling. Additionally, the use of a wide range of techniques opens the door to innovation and the exploration of new analytical approaches, which can be particularly important in a constantly evolving research field like the epidemiology of infectious diseases.

In the implementation of the models, we used the parameter optimization strategy by applying the Grid Search technique[35], aiming to improve performance considering the parameter space and the complexity of each model used. Initially, we implemented a search with a broader and coarser grid, followed by more refined searches in promising areas of the parameter space for each model. In Table 1, we present the models used with their appropriate classification of type, function or architecture and the search space for the grid search optimization method.

Type	Algorithm model	Function/Architecture	Parameters search space
Classic Statistics	SimpleExpS-moothing	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	smoothing level [0.1 ∼ 0.99]
	DoubleExpS-moothing	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}$	smoothing level [0.1 ∼ 0.99] smoothing trend [0.1 ∼ 0.99]
	Holt-Winters	$\ell_t = \alpha(y_t - s_{t-L}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}$ $s_t = \gamma(y_t - \ell_t) + (1-\gamma)s_{t-L}$	smoothing level [0.1 ∼ 0.99] smoothing slope [0.1 ∼ 0.99] smoothing seasonal [0.1 ∼ 0.99]
	ARIMA	$\phi(B)(y_t - \mu) = \theta(B)\varepsilon_t$ $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ $\theta(B)\varepsilon_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q} \nabla^d y_t = (1-B)^d y_t$	P [0 ∼ 5] Q [0 ∼ 5] D [0 ∼ 5]
Machine Learning	SVR	$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b$	kernel='rbf' gamma [0.1 ∼ 0.9] C [1 ∼ 10] epsilon [0.1 ∼ 0.09]
	XGBoost	$\mathrm{Obj}(\Theta) = \sum_{i=1}^{n}\mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K}\Omega(f_k)$	n. estimators [100, 250, 500, 1000] max depth [3, 5, 10, 15] learning rate [0.0001, 0.001, 0.01, 0.1]
	LSTM	$LSTM(60)-> DropOut(0.25)-> LSTM(120)-> DropOut(0.25)->Dense(20)-> Dense(1)$	epochs [100 ∼ 500] batch size = [1 ∼ 5] optimizer = [Adam] learing rate [0.001 ∼ 0.01] units per layer = [50 ∼ 150] dropout rate = [0.1 ∼ 0.5]
	Bidirectional LSTM	$Bidirectional(LSTM(64))-> DropOut(0.2)-> Bidirectional(LSTM(64))->\ DropOut(0.2)-> Dense(1)$	epochs [100 ∼ 500] batch size = [1 ∼ 5] optimizer = [Adam] learing rate [0.001 ∼ 0.01] units per layer = [16, 32, 64, 128] dropout rate = [0.1 ∼ 0.5]
	GRU	$GRU(64)-> DropOut(0.2)-> GRU(64)-> DropOut(0.2)-> Dense(1)$	epochs [100 ∼ 500] batch size = [1 ∼ 5] optimizer = [Adam] learing rate [0.001 ∼ 0.01] units per layer = [16, 32, 64, 128, 256] dropout rate = [0.1 ∼ 0.5]
	CNN	$CONV(64, kernel_size = 2)-> RELU-> MaxPooling->CONV(64, kernel_size = 2)-> RELU-> MaxPooling-> Flatten()-> Dense(256)-> RELU-> Dense(1)$	epochs [100 ∼ 500] batch size = [1 ∼ 5] optimizer = [Adam] learing rate [0.001 ∼ 0.01] units per layer = [16, 32, 64, 128] dropout rate = [0.1 ∼ 0.5]
	CNN + LSTM	$CONV(64, kernel_size = 2)-> RELU-> MaxPooling-> LSTM(50)-> Flatten()-> Dense(50)-> RELU-> Dense(1)$	epochs [100 ∼ 500] batch size = [1 ∼ 5] optimizer = [Adam] learing rate [0.001 ∼ 0.01] units per layer (CNN) = [16, 32, 64, 128] units per layer (LSTM) = [50 ∼ 150] dropout rate = [0.1 ∼ 0.5]
	CNN + GRU	$CONV(64, kernel_size = 2)-> RELU-> MaxPooling->TimeDistributed(Flatten)-> GRU(256)-> RELU-> Dense(256, kernel_l 2)-> RELU-> Dense(1)$	epochs [100 ∼ 500] batch size = [1 ∼ 5] optimizer = [Adam] learing rate [0.001 ∼ 0.01] units per layer (CNN) = [16, 32, 64, 128] units per layer (GRU) = [18, 32, 64, 128, 256] dropout rate = [0.1 ∼ 0.5]

Table 1. Overview of the models evaluated in the study, detailing the specific algorithms used, their mathematical functions or architectures, and the parameter ranges explored during the GridSearch optimization process.

Performance metrics

To evaluate the performance of the proposed models, we adopted metrics centered on evaluating prediction errors and quality of fit. The selected metrics include Mean Squared Error (MSE)[36], Mean Absolute Error (MAE)[37], Mean Absolute Percentage Error (MAPE)[38], Mean Absolute Scaled Error (MASE)[39], Mean Squared Logarithmic Error (MSLE)[40], Root Mean Squared Error (RMSE)[37], and Symmetric Mean Absolute Percentage Error (sMAPE)[41]. For a more refined analysis of the model fit, we also used the Akaike Information Criterion (AIC)[42] and the Bayesian Information Criterion (BIC)[43], which provide a comparative evaluation of the effectiveness of the model relative to its complexity and fit to the observed data.

During the training process, we adjusted the internal parameters of the models through various iterations, carefully monitoring the performance at each step based on these metrics. Parameters that result in the best fit, as indicated by the lowest value in error metrics and the best balance as determined by AIC and BIC, are selected

for the final model. This approach allows us to optimize the accuracy of the predictions and ensure that the model is well-fitted to the data, minimizing unnecessary complexity.

However, it is important to note that AIC and BIC are common criteria for traditional statistical models, such as exponential smoothing methods and autoregressive moving average models. However, as metrics based on the complexity relationship of the model structure, they are less relevant for evaluating machine learning-based approaches. In machine learning, especially in deep neural networks, the number of parameters can be extremely high, and the relationship between parameters and model complexity is not as straightforward.

For comparing the prediction accuracy of various models, we applied the Diebold-Mariano (DM) test[44]. This statistical test assesses whether there is a significant difference in forecast errors between two competing models. We calculated prediction errors for each model, computed the difference between these errors, and employed the DM test to evaluate the significance of these differences. Negative DM values indicate that the model in the row outperforms the model in the column. We analyzed results using a significance level of 0.05 for the p-value cutoff.

Ethical aspects

The study was approved by the Research Ethics Committee of the Federal University of the State of Mato Grosso, under opinion number 5,509,469, following the ethical recommendations of the National Health Council, by Resolution 466/12. All methods were carried out according to the guidelines and standards outlined in this resolution. The study did not require the subject's consent, as secondary data were used.

Results

From 2012 to 2023, 1,637 cases of TB/HIV coinfection were identified, resulting in an incidence of 3.99 cases per 100,000 inhabitants. Of this total, an incidence of 5.71 cases per 100,000 male inhabitants and 2.21 cases per 100,000 female inhabitants were observed. The analysis of incidence rates revealed that, in the state of Mato Grosso, during the studied period, the majority of TB/HIV coinfection cases prevailed among male individuals.

In Figure 1, the time series under the calculated monthly incidence rate over the investigated period are presented, stratified by sex and the general population. The anomalies relating to each stratification were also indicated in Figure 1. In the general population, no significant anomalies were identified, while in the male group an anomaly point was identified in September 2017 and in the female group in September 2014. Regarding the incidence rate trend, the total population presents fluctuations throughout the period, maintaining clear stationarity. However, the male incidence rate visually demonstrates a slightly increasing trend over time, especially with a peak around 2022, while the female incidence rate presents the lowest values on the scale over time, being visibly stable in the period.
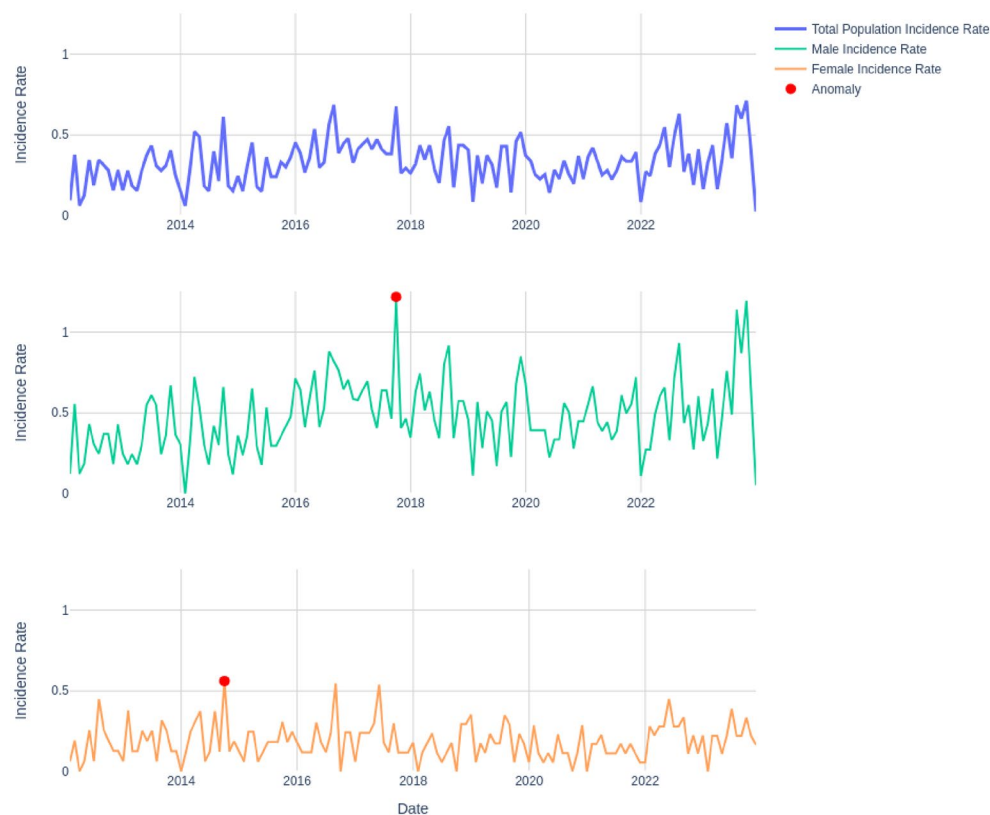
Figure 1. Time series of monthly incidence rates of TB/HIV co-infection, stratified by male, female sex and total population of the State of Mato Grosso, Brazil (2012-2023).

Statistics for Trend Identification		Female	Male	Total Population
ADF	Test Statistic	−11.778	−8.701	−9.462
	P-value	0.000	0.000	0.000
	Trend	Constant	Constant	Constant
	Critical Values	−3.48 (1%) −2.88 (5%) −2.58 (10%)	−3.48 (1%) −2.88 (5%) −2.58 (10%)	−3.48 (1%) −2.88 (5%) −2.58 (10%)
	Null Hypothesis	The process contains a unit root	The process contains a unit root	The process contains a unit root
	Alternative Hypothesis	The process is weakly stationary	The process is weakly stationary	The process is weakly stationary
DF-GLS	Test Statistic	−7.870	−4.809	−4.818
	P-value	0.000	0.000	0.000
	Trend	Constant	Constant	Constant
	Critical Values	−2.70 (1%) −2.08 (5%) −1.77 (10%)	−2.70 (1%) −2.08 (5%) −1.77 (10%)	−2.70 (1%) −2.08 (5%) −1.77 (10%)
	Null Hypothesis	The process contains a unit root	The process contains a unit root	The process contains a unit root
	Alternative Hypothesis	The process is weakly stationary	The process is weakly stationary	The process is weakly stationary
KPSS	Test Statistic	0.081	0.591	0.452
	P-value	0.687	0.024	0.053
	Trend	Constant	Constant	Constant
	Critical Values	0.74 (1%) 0.46 (5%) 0.35 (10%)	0.74 (1%) 0.46 (5%) 0.35 (10%)	0.74 (1%) 0.46 (5%) 0.35 (10%)
	Null Hypothesis	The process is weakly stationary	The process is weakly stationary	The process is weakly stationary
	Alternative Hypothesis	The process contains a unit root	The process contains a unit root	The process contains a unit root
Phillips-Perron Test	Test Statistic	−11.792	−9.310	−9.908
	P-value	0.000	0.000	0.000
	Trend	Constant	Constant	Constant
	Critical Values	−3.48 (1%) −2.88 (5%) −2.58 (10%)	−3.48 (1%) −2.88 (5%) −2.58 (10%)	−3.48 (1%) −2.88 (5%) −2.58 (10%)
	Null Hypothesis	The process contains a unit root	The process contains a unit root	The process contains a unit root
	Alternative Hypothesis	The process is weakly stationary	The process is weakly stationary	The process is weakly stationary
Seasonality Test	P-value	0.475	0.009	0.017
	Period	12	12	12
	Result	The time series does not have significant seasonality	The time series has significant seasonality	The time series has significant seasonality

Table 2. Identification of the trend in the time series to understand the general behavior of the data, determining ancestry, descent and stationarity.

In Table 2, the results of the ADF, DF-GLS, KPSS, Phillips-Perron, and Seasonality tests are presented to identify trends in the time series, with the aim of determining whether the data are stationary or contain a unit root, which implies the presence of a stochastic or deterministic trend.

The ADF, DF-GLS, and Phillips-Perron tests computed for individuals of female and male sex, as well as the total population, show highly negative statistics with a p-value <= 0.001, rejecting the null hypothesis that the data contain a unit root in favor of the alternative hypothesis that the data are stationary. This means that the time series of these incidence rates have constant mean and variance over time. The KPSS test, which has a different approach where the null hypothesis is that the series is stationary, showed that for women (p-value of 0.687) and the total population (p-value of 0.053), the null hypothesis is not rejected, suggesting that the series are stationary. However, for male individuals, the p-value of 0.024 rejects the null hypothesis, indicating that the series may not be stationary. Furthermore, the test conducted to verify the presence of significant seasonal patterns in the time series showed that for female individuals, there is no relevant seasonality (p-value of 0.475),

while for male individuals and the total population, there is evidence of significant seasonality (p-values of 0.009 and 0.017, respectively). These results are important for understanding the overall behavior of the data and can be crucial in modeling forecasts or detecting trends and cycles.

In Table 3, the results obtained with the models detailed in the Methods Section are presented. All different models were trained using the parameters obtained through the GridSearch optimization strategy that were made available as supplementary information through Table Table S1. Each model was evaluated using a data-split strategy into training and validation sets, with approximately 80% of the observations designated for training and approximately 20% for validation. The validation set was fixed at the final 24 observations for each stratification by sex and the total population to ensure a uniform interval among the series, facilitating comparisons and analyses. The performance of each model was evaluated using the metrics detailed in the Methods section, specifically in the subsection titled Performance metrics. Furthermore, the results of the Diebold-Mariano (DM) test, assessing significant differences in predictive performance, are available as supplementary information in this study's Table Table S2.

Time Series	Models	Metrics							
		AIC	BIC	MSE	MAE	MSLE	RMSE	MASE	sMAPE
Female	Simple Exponential Smoothing (SES)	−510.30	−504.74	0.019	0.118	0.013	0.138	1.060	64.74%
	Double Exponential Smoothing (DES)	−481.06	−469.94	0.011	0.086	0.008	0.106	0.771	42.91%
	Triple Exponential Smoothing (Holt-Winters)	−461.64	−417.17	0.014	0.098	0.009	0.119	0.876	51.54%
	Autoregressive Integrated Moving Average (ARIMA)	−167.77	−140.41	0.012	0.088	0.008	0.111	0.789	45.95%
	Support Vector Regression (SVR)	*	*	0.012	0.088	0.008	0.110	0.809	44.65%
	Extreme Gradient Boosting (XGBoost)	*	*	0.011	0.084	0.008	0.107	0.748	42.90%
	LSTM	*	*	0.013	0.096	0.008	0.116	0.705	36.47%
	LSTM Bidirectional	*	*	**0.009**	**0.067**	**0.006**	**0.094**	**0.492**	**26.81%**
	CNN	*	*	0.012	0.084	0.007	0.109	0.614	32.18%
	CNN + LSTM	*	*	0.009	0.070	0.006	0.097	0.514	26.97%
	GRU	*	*	0.016	0.103	0.009	0.125	0.759	39.19%
	GRU + CNN	*	*	0.018	0.110	0.010	0.133	0.806	41.14%
Male	Simple Exponential Smoothing (SES)	−393.00	−387.42	0.087	0.226	0.033	0.295	0.769	43.88%
	Double Exponential Smoothing (DES)	−369.66	−358.51	0.074	0.215	0.030	0.272	0.732	41.09%
	Triple Exponential Smoothing (Holt-Winters)	−363.56	−318.96	0.075	0.214	0.029	0.273	0.729	41.49%
	Autoregressive Integrated Moving Average (ARIMA)	1.63	28.73	0.073	0.203	0.030	0.271	0.693	38.26%
	Support Vector Regression (SVR)	*	*	0.090	0.246	0.035	0.300	0.789	46.36%
	Extreme Gradient Boosting (XGBoost)	*	*	0.087	0.240	0.035	0.296	0.816	46.03%
	LSTM	*	*	0.006	0.061	0.003	0.076	0.369	21.19%
	LSTM Bidirecional	*	*	**0.003**	**0.044**	**0.002**	**0.057**	**0.268**	**16.61%**
	CNN	*	*	0.012	0.080	0.007	0.109	0.487	27.56%
	CNN + LSTM	*	*	0.013	0.096	0.007	0.115	0.584	30.47%
	GRU	*	*	0.011	0.081	0.006	0.103	0.492	26.17%
	GRU + CNN	*	*	0.020	0.109	0.011	0.143	0.663	33.95%
Total population	Simple Exponential Smoothing (SES)	−496.62	−491.04	0.039	0.157	0.020	0.197	0.821	45.20%
	Double Exponential Smoothing (DES)	−487.12	−475.97	0.031	0.138	0.016	0.175	0.719	38.79%
	Triple Exponential Smoothing (Holt-Winters)	−468.00	−423.40	0.033	0.145	0.017	0.181	0.759	42.13%
	Autoregressive Integrated Moving Average (ARIMA)	−157.29	−143.48	0.033	0.144	0.017	0.182	0.752	40.84%
	Support Vector Regression (SVR)	*	*	0.037	0.164	0.019	0.192	0.812	46.06%
	Extreme Gradient Boosting (XGBoost)	*	*	0.029	0.134	0.015	0.170	0.702	37.83%
	LSTM	*	*	0.014	0.087	0.007	0.117	0.456	28.26%
	LSTM Bidirectional	*	*	0.010	0.075	0.006	0.102	0.393	25.69%
	CNN	*	*	0.029	0.140	0.015	0.169	0.730	39.75%
	CNN + LSTM	*	*	**0.006**	**0.058**	**0.003**	**0.075**	**0.301**	**19.66%**
	GRU	*	*	0.017	0.107	0.009	0.130	0.561	33.13%
	GRU + CNN	*	*	0.032	0.142	0.016	0.179	0.744	40.41%

Table 3. Summary of the performance metrics of the different models evaluated in the study to predict the incidence rates of TB/HIV co-infection, stratified by male, female sex and total population.

For the data stratification with female individuals, the best overall performing model was the Bidirectional LSTM, with the lowest values for MSE (0.009), MAE (0.067), MSLE (0.006), RMSE (0.094), MASE (0.492), and sMAPE (26.81%), indicating a lower prediction error compared to the other models evaluated. When comparing the classic statistical models, although the AIC and BIC of the Simple Exponential Smoothing (SES) method, with values of −510.30 and −504.74 respectively, are the lowest indicating a better balance between fit and model complexity, the other applied metrics consolidate the Double Exponential Smoothing (DES) as having the best performance. Figure 2A visualizes the performance of each model according to the applied training strategy and tracks their evolution according to the presented metrics.

Figure 3A visualizes the performance of the best model (Bidirectional LSTM) with data stratified by female individuals during the period from 2012 to 2023. This visualization allows for a more detailed observation of the predictive capacity of the model concerning the present perspective of observed data. Figure 4A presents the time series with future projections for the period from 2024 to 2025, according to the predictive capacity of the best model.

For the dataset with male individuals, the Bidirectional LSTM again stands out, displaying the lowest values in most error metrics, including a significantly low sMAPE (16.61%). When comparing only the classic statistical models, ARIMA showed the lowest sMAPE (38.26%), indicating the highest accuracy in forecasts among models in this category. ARIMA also demonstrates competitive values in other error metrics such as MSE (0.073), MAE (0.203), MSLE (0.030), RMSE (0.271), and MASE (0.693), indicating good overall performance in prediction accuracy. However, the ARIMA model showed significantly high AIC (1.63) and BIC (28.73) values compared to other models. This atypical discrepancy, especially considering its good performance in other error metrics, suggests caution, indicating potentially unnecessary complexity that may not generalize well to unseen data.

Figure 2B visualizes the performance of all models used with data from male individuals. Figure 3B highlights the performance of the best model (Bidirectional LSTM) during the period from 2012 to 2023. Concluding this stratification, Figure 4B presents the time series with future projections for the period 2024 to 2025, using the best-performing model.

For the total population, the CNN + LSTM model followed by the Bidirectional LSTM model showed the best performance, with the lowest errors in all considered metrics, indicating high accuracy in forecasts for this time series. As for models based on classic statistics, Double Exponential Smoothing (DES) showed one of the lowest values for AIC (−487.12) and BIC (−475.97), suggesting that this model is more efficient in terms of balancing fit to data and complexity of the evaluated models. This factor is evidenced by the lowest values in almost all additional error metrics employed in the model evaluation.
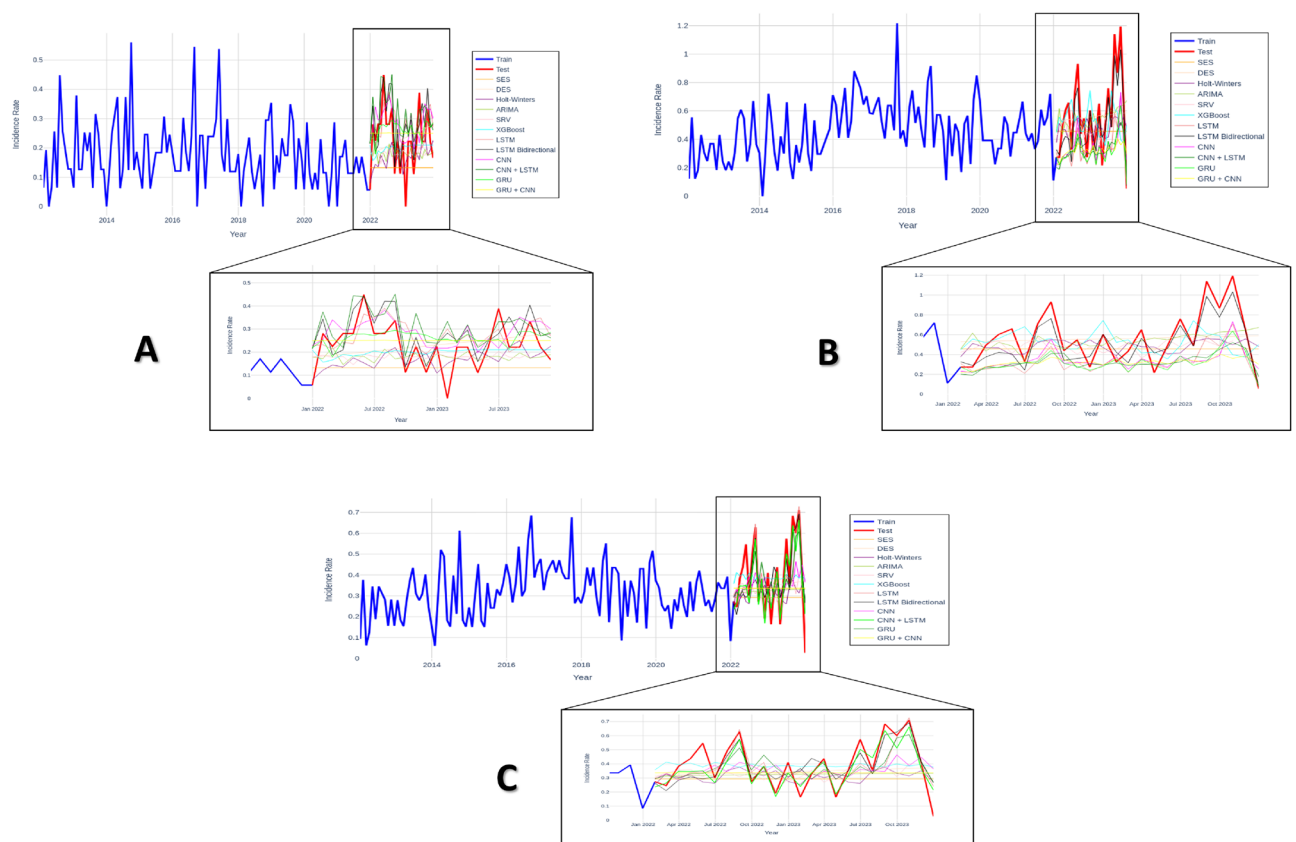

Figure 2. Visualization of the training of the models evaluated for predicted versus actual time series of TB/HIV co-infection for the period from 2012 to 2023, with panels zoom the predictions for the period of the last 24 months, stratified by A - Female, B - Male and C - Total Population.
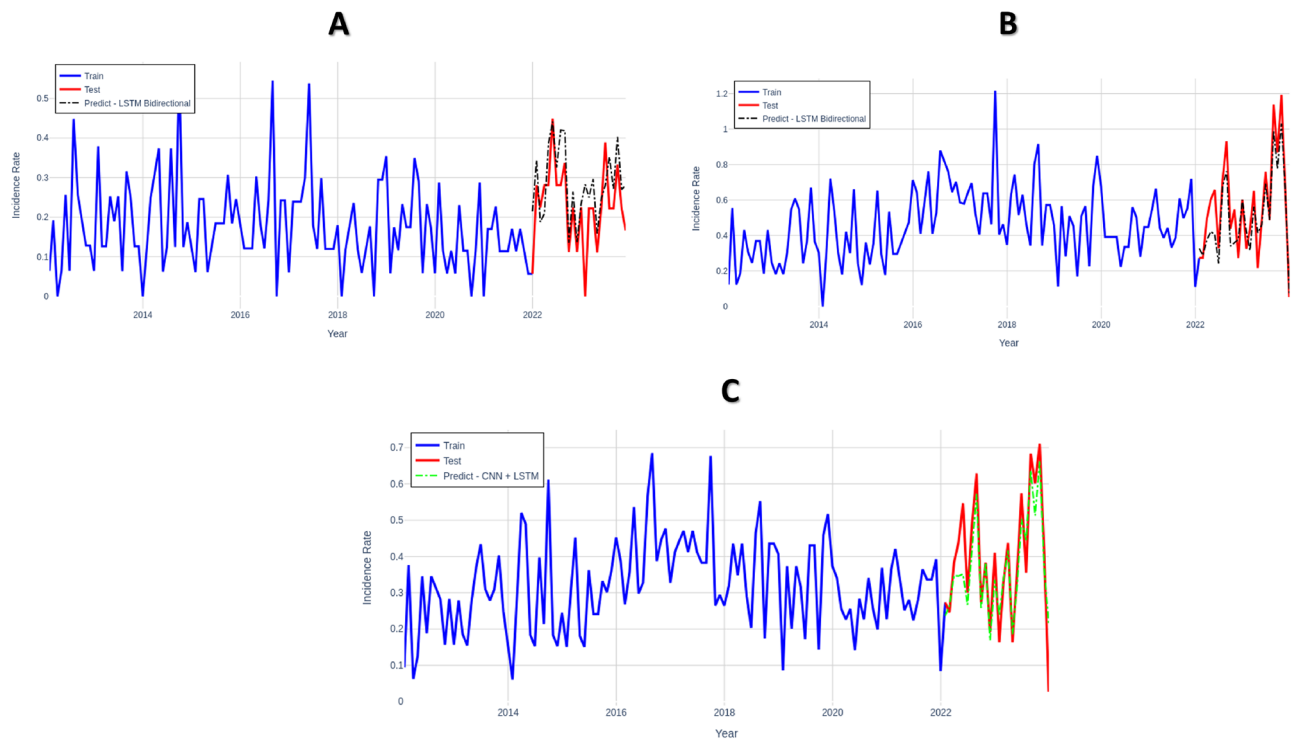
Figure 3. Visualization of the training and prediction of the best model for predicted versus actual time series of TB/HIV co-infection for the period 2012 to 2023, stratified by A - Female, B - Male and C - Total Population.
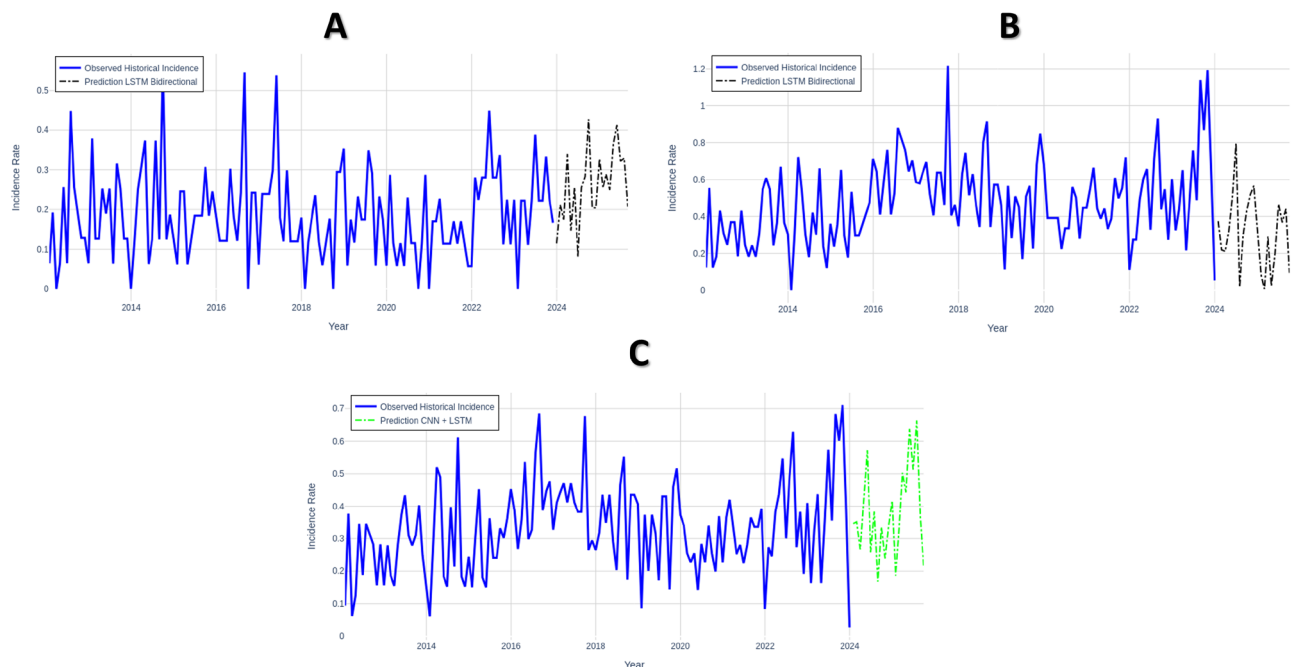
Figure 4. Presentation of the time series and their respective future projections for the period from 2024 to 2025, based on the predictive capability of the standout models. A - Female Segment, B - Male Segment, and C - Total Population.

Figure 2C presents the performance of all models used with the complete data, termed the total population. The performance of the best model for this data set is presented through Figure 3C. Finally, Figure 4C presents the time series with future projections for the period from 2024 to 2025 using the CNN + LSTM model, which showed the best performance.

The Bidirectional LSTM consistently performs best across the three time series, followed closely by the CNN + LSTM model. This demonstrates the effectiveness of deep learning models in capturing complex patterns in these datasets.

Discussion

This study aimed to explore prediction models for the incidence of TB/HIV coinfection, estimating the epidemic trend of cases from a past, present, and future perspective, thereby providing a reference for prevention and control for public policy agencies. To this end, an analysis of TB/HIV notification data from 2012 to 2023 in the state of Mato Grosso, Brazil, was conducted. The monthly incidence rate for the reported cases in this range was calculated, and the temporal trend of these data was evaluated, stratified by male, female, and total population.

In this context, we compare cases of TB/HIV in the state of Mato Grosso with other regions of the country. The Central-West region, where Mato Grosso is located, ranked second in the number of TB/HIV coinfection notifications in 2022, with a proportion of 10.0%. The South region ranks first with 12.6% of notifications. The state of Mato Grosso, with 9.3%, also stands out for being among the ten states with the highest number of cases of coinfection, even surpassing the national average of 8.4%[3].

This epidemiological scenario highlights that TB/HIV cases in Mato Grosso are significantly relevant compared to other regions of the country. According to Humayun et al.(2022), there is a predominance of coinfection in male individuals, corroborating with the results found in our investigation that indicate a higher incidence in this population group.

The evidence reinforces the need for special attention to male individuals, especially in regions with extensive rural areas, such as the state of Mato Grosso[45]. Developing and implementing specific prevention and control measures can improve TB/HIV coinfection rates in this group. These include targeted health campaigns for men, mobile health clinics, workplace health initiatives, community-based interventions, improved health literacy, enhanced screening and diagnostic services, and integrated health services for concurrent TB and HIV care.

To further understand the dynamics of TB/HIV incidence rates, the time series trends for the period from 2012 to 2023 were evaluated using the ADF, DF-GLS, KPSS, and Phillips-Perron tests. The choice of multiple tests is justified by their complementarity and the specific characteristics of the analyzed datasets. The identified weak stationarity implies that the mean and autocovariance are stable over time, although the complete distribution of time series values does not need to be constant. The autocovariance should be influenced only by the time difference (lag) between two observations, regardless of the specific moments when the observations occur[46].

These evaluations revealed that the value distribution for the time series, stratified by female, male, and total population, demonstrated a stationary trend, with significant seasonality observed in the male and total population groups. The absence of a decreasing trend in notifications suggests that current strategies and interventions to combat TB/HIV are insufficient to reduce the incidence of these conditions. This highlights the need for reviewing and intensifying prevention, diagnosis, and treatment measures.

Seasonality underscores the particular challenges in controlling TB/HIV coinfection, such as the stigma associated with both diseases, difficulties in accessing and adhering to treatment, and the necessity for integrated health approaches that consider the interactions between TB, HIV, and other social determinants of health.

The observed trends are further complicated by external factors, most notably the COVID-19 pandemic, which had significant impacts on TB notifications worldwide[2]. These impacts were due to a combination of factors including the reorganization of health systems, changes in people's behavior in seeking medical care, and disruptions in health services. Thus, when observing the period from 2020 to 2022 regarding TB/HIV notifications in the state of Mato Grosso, it was possible to verify the same downward trend.

It should be emphasized that the decline in TB/HIV coinfection notifications may not indicate a lack of cases, but rather that these cases possibly went unreported due to difficulties in accessing health services. During the pandemic, late diagnoses and treatment interruptions were common, which may have contributed to complications arising from TB/HIV coinfection. The situation experienced during the pandemic further highlighted the need to fulfill commitments related to TB and HIV, especially through the expansion of ethical and person-centered care, with equity and access to health and social rights[4].

In light of these challenges, predictive models play a crucial role in controlling and managing TB, one of the oldest and most persistent infectious diseases affecting humanity. The insights provided by these models guide the development of policies and public health programs, steering decisions on research priorities, health infrastructure development, and education and communication strategies. They are essential for the ongoing monitoring of the effectiveness of public health policies and TB control programs, allowing for quick adjustments in response to real-time feedback, and are effectively considered a global strategy to end TB, providing a basis for evidence-based decisions in disease control.

In this perspective, this study explored two prevalent approaches to constructing predictive models for infectious diseases. The first group of models comes from classical statistics, such as exponential smoothing (SES, DES, Holt-Winters) and autoregressive integrated moving average (ARIMA); the second comprises machine learning-based prediction models like Support Vector Regression (SVR), Decision Trees (XGBoost), and artificial neural network models (LSTM, CNN, and GRU).

Although classical statistical predictive models have been relatively successful in predicting infectious diseases, they struggle to extract nonlinear relationships in a time series[47]. In contrast, one of the main advantages of machine learning over classical statistical methods is its ability to capture and model the complexity and nonlinearity of data without the need to explicitly specify the form of the relationship between input and output variables[48]. However, it is important to note that the literature already tacitly indicates that the characteristics of the data are prevalent in determining the assertiveness of these methods, regardless of the approach[49-51].

Thus, in the context of the data evaluated in this study, it was conclusive that deep learning models, specifically Bidirectional LSTM and the custom CNN + LSTM model, demonstrated superiority in predicting these stratified datasets, evidenced by lower error metrics, suggesting greater efficiency in capturing complex patterns in the data compared to simpler methods like SES, DES, and Holt-Winters. Another point of note is the performance of the ARIMA model[52,53], which has shown a peculiar prediction capacity in a wide variety of scenarios and only performed better in the male time series when compared only to classical statistical models. This result reaffirms that the success of a model's prediction is intrinsically related to the characteristics and behavior of the data, which can be very beneficial, as simpler models like SES and DES can provide assertive guidance for decision-making processes regarding TB/HIV coinfection.

However, attention must be given to the assertiveness of these simpler models. Despite lower AIC and BIC[54] indicating a more efficient model in terms of simplicity and information use, they do not necessarily translate into greater accuracy in predictions, as demonstrated by higher sMAPE values.

A lower sMAPE is preferable, indicating a smaller percentage error between predictions and actual values, reflecting more accurate predictions. A lower sMAPE is preferable, indicating a smaller percentage error between predictions and actual values, reflecting more accurate predictions. For instance, lower sMAPE values are crucial because sMAPE is a measure of accuracy that is less sensitive to extreme values and more interpretable in percentage terms, making it easier to understand the model's predictive performance. By comparing sMAPE values across different models, we can identify which models perform better in terms of prediction accuracy.

Regarding the performance of the SVR and XGBoost models based on the machine learning approach and parameter optimization using the Grid Search technique, it can be argued that their results might be conditioned by a suboptimal configuration, which can lead to inferior performance[31]. Another relevant issue is that although these two models are successful, they might not be the most suitable for dealing with the TB/HIV dataset with intrinsic temporal characteristics. There's also the possibility of the curse of dimensionality or overfitting, coupled with a deficiency in capturing seasonal patterns and trends. Thus, models like SVR and XGBoost might not be as effective in modeling patterns without extensive feature engineering[55,56].

On the other hand, deep learning models, designed specifically to manage large volumes of data, exhibit a remarkable ability to generalize. This is largely due to their ability to efficiently extract and organize feature hierarchies. Among these, models such as Long Short-Term Memory (LSTM) prove particularly effective in capturing temporal sequences and identifying long-range dependencies present in the data[6].

Moreover, despite their sensitivity to configuration, deep learning models tend to be more robust compared to other methods. This robustness stems from their unique ability to learn complex data representations, allowing more flexible adaptation to different types of patterns and structures inherent in the datasets they are trained on.

Our results based on time series and predicting the future behavior of TB/HIV coinfection suggest that, in Mato Grosso, meeting the targets proposed for 2030 by the UN SDGs and the End TB Strategy seems increasingly unlikely. According to Silva *et al.* (2021), if TB is not controlled and the current death rate continues, 31.8 million people will die from TB between 2020 and 2050, leading to an economic loss of 17.5 billion dollars.

The epidemiological scenario projected in this study shows that the incidence of TB/HIV will not be reduced unless decisive measures are taken by policymakers and health professionals. According to WHO, recommendations to advance TB/HIV coinfection control include TB screening for all PLHIV at the time of diagnosis and all follow-up visits, as well as routine HIV testing for all TB patients. It is crucial that PLHIV with active TB receive both TB treatment and antiretroviral therapy (ART). Furthermore, PLHIV without active TB should receive TB preventive treatment to reduce the risk of developing the disease[58].

Brazil has invested in public policies that span the health sector with intra and intersectoral scope, aiming to accelerate efforts to eliminate diseases such as TB. In 2023, the country's Ministry of Health reaffirmed its commitment to eliminating TB and announced the goal of eliminating the disease by 2030, advancing the target initially proposed in the National Plan (10 cases per 100,000 inhabitants and less than 230 deaths per year) by five years[4].

Among the implemented policies, the Healthy Brazil Program - Unite to Care, established on 7 February 2024, stands out as a government policy aligned with the SDGs, aimed at controlling TB, HIV, and other socially determined diseases. The priority is for those affected by these diseases to undergo proper treatment, with reduced costs and better results in the network of health professionals and services[59]. Added to this is the resumption of investment in innovation, science, and technology, including exclusive funding for TB research[4]. These actions may represent an advancement in terms of addressing TB/HIV coinfection in the country, especially in the state of Mato Grosso, the setting of this study, where future predictions for the incidence of this coinfection are not encouraging.

This study stands out for its utilization of a wide array of predictive models, including deep learning approaches such as Bidirectional LSTM and CNN + LSTM, showcases the study's strength in employing cutting-edge data analysis techniques to address complex epidemiological challenges. These models demonstrated superior performance in capturing the intricate patterns within the data, suggesting their potential utility in guiding more effective TB/HIV prevention and control strategies.

However, the study is not without limitations. One significant concern is the accuracy of the dataset used, which may be compromised by incomplete record-keeping or underreporting of TB/HIV cases. Such discrepancies could skew the analysis and affect the reliability of the predictive models. Potential inaccuracies arise from various sources, including data entry errors, inconsistencies in reporting practices, and limitations in diagnostic capabilities, particularly in resource-limited settings.

To mitigate the problem of data inaccuracies in future studies, several strategies can be implemented. Enhancing data collection processes through standardized reporting protocols and comprehensive training for healthcare workers is essential. Employing robust data validation and cleaning techniques, along with integrating data from multiple sources such as electronic health records and community health surveys, can improve data accuracy.

Additionally, using advanced analytical methods like machine learning algorithms to handle missing data and establish continuous monitoring and feedback systems will help identify and correct emerging issues promptly.

Another limitation, in addition to data precision, is the use of univariate time series analysis, which ignores the impact of socioeconomic variables and other external factors that significantly influence TB incidence rates. Incorporating these multifaceted factors could provide a more holistic understanding of the drivers behind the trends in TB/HIV coinfection and enhance the predictive accuracy of the models.

Furthermore, deep learning models, such as Bidirectional LSTM and CNN + LSTM, are sensitive to their configuration. Parameter optimization was performed using Grid Search; however, suboptimal configurations might still affect performance. These models also require substantial computational resources and time for training. Our models accounted for seasonality and trends in the data, but abrupt changes in external conditions (e.g., public health interventions, pandemics) could alter these patterns, reducing the predictive accuracy of the models.

Conclusion

In conclusion, this study has demonstrated the superiority of deep learning models, specifically Bidirectional LSTM and custom CNN + LSTM, over classical statistical models in capturing complex patterns in TB/HIV coinfection data. This superiority is evidenced by lower error metrics. These metrics indicate that these models are more effective in predicting the incidence of TB/HIV coinfection. The comparison of various models in the study highlights the importance of considering the specific characteristics of the data when selecting the appropriate predictive model.

The superior performance of deep learning models can lead to more reliable predictions of TB/HIV co-infection trends. This enables health authorities to allocate resources more efficiently and plan specific strategic actions in advance. Additionally, it allows for the review of the efficiency of existing strategies. It also facilitates the implementation and evaluation of new control actions aimed at reducing the number of cases.

However, our study has some significant limitations. Data quality and integrity can directly affect model accuracy. Univariate time series analysis that ignores the impact of socioeconomic variables. The sensitivity of models to correct configuration and the substantial consumption of computational resources and training time. Finally, our models account for seasonality and trends in the data, but abrupt changes in external conditions (e.g., public health interventions, pandemics) can alter these patterns, reducing predictive accuracy.

For future research, it will be essential to incorporate socioeconomic data that characterize the significant social determinants relevant to TB. This will enable the exploration of predictive modeling of multivariate time series. Such an approach could lead to even more accurate and comprehensive models. Further studies could also investigate the application of these advanced models in different regions and settings to validate their generalizability and robustness in varying contexts.

Data availability

The epidemiological data, comprising all TB/HIV cases in the state of Mato Grosso reported to the Notifiable Diseases Information System (SINAN), corroborate the findings of this investigation. It is important to note that SINAN is a Brazilian information system responsible for recording and processing information about diseases that are mandatory to report across the country, maintained by the Department of Informatics of the Unified Health System (DATASUS). Data are publicly accessible at "https://datasus.saude.gov.br/acesso-a-informacao/casos-de-tuberculose-desde-2001-sinan/", according to the transparency and openness policy of data and information aimed at stimulating research, innovation, scientific production, business generation and the economic and social development of Brazil. Furthermore, the population data were extracted from the 2022 Demographic Census, available on the website of the Brazilian Institute of Geography and Statistics (IBGE) at "https://censo 2022.ibge.gov.br/panorama/", which also adheres to the Brazilian Government's policy of transparency and openness of data and information.

References

1. Gaspar, R. S., Nunes, N., Nunes, M. & Rodrigues, V. P. Temporal analysis of reported cases of tuberculosis and of tuberculosis-hiv co-infection in brazil between 2002 and 2012. *J. Bras. Pneumol.* **42**, 416–422. https://doi.org/10.1590/S1806-37562016000000054 (2016).
2. World Health Organization. Global tuberculosis report 2023. https://iris.who.int/bitstream/handle/10665/373828/9789240083 851-eng.pdf?sequence=1 (2023).
3. Secretaria de Vigiãancia em Saúde e Ambiente - Ministério da Saúde. Boletim epidemiológico - tuberculose 2023. https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/especiais/2023/boletim-epidemiologico-de-tuber culose-numero-especial-mar.2023 (2023).
4. Ministério da Saúde. *Boletim Epidemiológico - Tuberculose 2024*, vol. Número Especial (Departamento de HIV/Aids, Tuberculose, Hepatites Virais e Infecçõµes Sexualmente Transmissíveis, Coordenação Geral de Vigilâcncia da Tuberculose, Micoses, 2024).
5. General Assembly of the United Nations. *Political declaration of the high-level meeting on the fight against tuberculosis: draft resolution/submitted by the president of the general assembly* (General Assembly of the United Nations, New York, 2023).
6. Han, Z., Zhao, J., Leung, H., Ma, K. & Wang, W. A review of deep learning models for time series prediction. *IEEE Sens. J.* **21**, 7833–7848. https://doi.org/10.1109/JSEN.2019.2923982 (2019).
7. Pimpin, L. *et al.* Tuberculosis and hiv co-infection in european union and european economic area countries. *European Respiratory Journal* **38**, 1382–1392, https://doi.org/10.1183/09031936.00198410 (2011). https://erj.ersjournals.com/content/38/6/1382.full.pdf.
8. Lima, M. d. S. *et al.* Mortality related to tuberculosis-hiv/aids co-infection in brazil, 2000-2011: epidemiological patterns and time trends. *Cadernos de Saude Publica* **32**, e00026715, https://doi.org/10.1590/0102-311X00026715 (2016).
9. Osei, E., Oppong, S. & Der, J. Trends of tuberculosis case detection, mortality and co-infection with hiv in ghana: A retrospective cohort study. *PLoS ONE* **15**, 1–17. https://doi.org/10.1371/journal.pone.0234878 (2020).

10. Siqueira Santos, L. F. *et al.* Tuberculosis/hiv co-infection in northeastern brazil: Prevalence trends, spatial distribution, and associ-ated factors. *The Journal of Infection in Developing Countries* **16**, 1490–1499, https://doi.org/10.3855/jidc.16570 (2022).
11. World Health Organization. The end tb strategy: global strategy and targets for tuberculosis prevention, care and control after 2015. Geneva, Switzerland (2015). Accessed 21 Mar 2024.
12. United Nations. General assembly resolution a/res/70/1. transforming our world, the 2030 agenda for sustainable development. [cited 2016 Feb 10] (2015). Available from: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.
13. Organization, W. H. Global health sector strategies on hiv, viral hepatitis and sexually transmitted infections for the period 2022–2030. https://iris.who.int/bitstream/handle/10665/360348/9789240053779-eng.pdf?sequence=1 (2022).
14. Departamento de Informática do SUS (DATASUS). Sistema de informação de agravos de notificação (sinan). DataSUS - Minstçãrio da Saúde (2024).
15. Instituto Brasileiro de Geografia e Estatística (IBGE). Censo demográfico 2022. Instituto Brasileiro de Geografia e Estatística (2022).
16. Dye, C. *et al.* Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. *JAMA* **282**, 677–686 (1999).
17. Newbold, P. & Granger, C. W. Experience with forecasting univariate time series and the combination of forecasts. *J. R. Stat. Soc.: Ser. A (General)* **137**, 131–146 (1974).
18. Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. Stl: A seasonal-trend decomposition. *J. Off. Stat* **6**, 3–73 (1990).
19. Ahmad, S. & Purdy, S. *Real-time anomaly detection for streaming analytics* **1607**, 02480 (2016).
20. Dickey, D. A. & Fuller, W. A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **74**, 427–431 (1979).
21. Stock, J. H. & Watson, M. W. A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica: J. Econ. Soc.* 783–820 (1993).
22. Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *J. Econ.* **54**, 159–178 (1992).
23. Phillips, P. C. & Perron, P. Testing for a unit root in time series regression. *biometrika* **75**, 335–346 (1988).
24. Hamilton, J. D. *Time Series Analysis* (Princeton University Press, 1994).
25. Brockwell, P. J. & Davis, R. A. *Introduction to Time Series and Forecasting* (Springer, 2002).
26. Brown, R. G. *Smoothing, Forecasting and Prediction of Discrete Time Series* (Prentice-Hall, 1963).
27. Holt, C. C. *Forecasting trends and seasonal by exponentially weighted averages* (ONR Research Memorandum, Carnegie Institute of Technology, 1957).
28. Chatfield, C. The holt-winters forecasting procedure. *J. Roy. Stat. Soc.: Ser. C (Appl. Stat.)* **27**, 264–279 (1978).
29. Box, G. E. & Pierce, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Am. Stat. Assoc.* **65**, 1509–1526 (1970).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
31. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Confer-ence on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).
32. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
33. Cho, K. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 (Association for Computational Linguistics, 2014).
34. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
35. Belete, D. M. & Huchaiah, M. D. Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results. *Int. J. Comput. Appl.* **44**, 875–886 (2022).
36. Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H. & Lee, T.-C. *Regression Analysis: Theory, Application, and Computation* (John Wiley & Sons, 1985).
37. Chai, T. & Draxler, R. R. Root mean square error (rmse) or mean absolute error (mae)?-arguments against avoiding rmse in the literature. *Geosci. Model Dev.* **7**, 1247–1250 (2014).
38. De Myttenaere, A., Golden, B., Le Grand, B. & Rossi, F. Mean absolute percentage error for regression models. *Neurocomputing* **192**, 38–48 (2016).
39. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**, 679–688 (2006).
40. Rao, J. N. K., Jiang, J. & Das, K. Mean squared error of empirical predictor. *Ann. Stat.* **32**, 818–840. https://doi.org/10.1214/009053604000000201 (2004).
41. Makridakis, S. Accuracy measures: theoretical and practical concerns. *Int. J. Forecast.* **9**, 527–529 (1993).
42. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974).
43. Schwarz, G. Estimating the dimension of a model. *Annals Stat.* 461–464 (1978).
44. Diebold, F. X. & Mariano, R. S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **20**, 134–144 (1995).
45. Alves, J. D. *et al.* Bayesian spatio-temporal models for mapping tb mortality risk and its relationship with social inequities in a region from brazilian legal amazon. *Trans. R. Soc. Trop. Med. Hyg.* https://doi.org/10.1093/trstmh/traa008 *(2020).*
46. Abadir, K. M. & Talmain, G. Autocovariance functions of series and of their transforms. *J. Econ.* **124**, 227–252 (2005).
47. Wang, G. *et al.* Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting hiv incidence in guangxi, china. *Epidemiol. Infect.* **147**, https://doi.org/10.1017/S095026881900075X (2019).
48. Zhao, D. *et al.* The research of arima, gm(1,1), and lstm models for prediction of tb cases in china. *PLoS ONE* **17**, https://doi.org/10.1371/journal.pone.0262734 (2022).
49. Velicer, W. & Harrop, J. The reliability and accuracy of time series model identification. *Eval. Rev.* **7**, 551–560. https://doi.org/10.1177/0193841X8300700408 (1983).
50. Hinich, M. Testing for dependence in the input to a linear time series model. *J. Nonparametric Stat.* **6**, 205–221. https://doi.org/10.1080/10485259608832672 (1996).
51. Patton, A. J. A review of copula models for economic time series. *J. Multivar. Anal.* **110**, 4–18. https://doi.org/10.1016/j.jmva.2012.02.021 (2012).
52. Conejo, A., Plazas, M., Espínola, R. & Molina, A. B. Day-ahead electricity price forecasting using the wavelet transform and arima models. *IEEE Trans. Power Syst.* **20**, 1035–1042. https://doi.org/10.1109/TPWRS.2005.846054 (2005).
53. Khashei, M., Bijari, M. & Ardali, G. R. Hybridization of autoregressive integrated moving average (arima) with probabilistic neural networks (pnns). *Comput. Ind. Eng.* **63**, 37–45. https://doi.org/10.1016/j.cie.2012.01.017 (2012).
54. Yang, Y. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika* **92**, 937–950 (2005).
55. Rieger, C. & Zwicknagl, B. Deterministic error analysis of support vector regression and related regularized kernel methods. *J. Mach. Learn. Res.* **10**, 2115–2132. https://doi.org/10.5555/1577069.1755856 (2009).
56. Zhang, P., Jia, Y. & Shang, Y. Research and application of xgboost in imbalanced data. *Int. J. Distrib. Sens. Netw.* **18**, https://doi.org/10.1177/15501329221106935 (2022).

57. Silva, S., Arinaminpathy, N., Atun, R., Goosby, E. & Reid, M. Economic impact of tuberculosis mortality in 120 countries and the cost of not achieving the sustainable development goals tuberculosis targets: A full-income analysis. *Lancet Glob. Health* **9**, e1372–e1379. https://doi.org/10.1016/S2214-109X(21)00299-0 (2021).
58. HIV and Tuberculosis. https://www.who.int/westernpacific/health-topics/hiv-aids/hiv-and-tuberculosis (2024). Accessed 21 Mar 2024.
59. Decree No. 11,908, 06 fevereiro de 2024. https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/decreto/D11908.htm (2024). Accessed on 21 de março de 2024.

Acknowledgements

Author contributions

A.A. conceived the experiment(s), A.A. and J.D.A. conducted the experiment(s), D.K. and N.d.S.B. worked on data extraction and normalization, L.F.P., A.R.S., T.Z.B., A.C.V.R., and R.A.A. analyzed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-69580-4.

Correspondence and requests for materials should be addressed to A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.