A Comparative Analysis of Predictive Models for Hospital Average Length of Stay

Ana Karla Dias de Souza[1], **João Fausto Lorenzato de Oliveira**[1], **Cleyton Mário de Oliveira Rodrigues**[1]

[1]Polytechnic School of Pernambuco (PPGEC)
University of Pernambuco (UPE) - Recife, Brazil

{karla.dias,fausto.lorenzato,cleyton.rodrigues}@upe.br

Abstract. *The average length of stay (ALS) is a key indicator for assessing hospital efficiency and bed management. Monthly data from the Complexo Hospitalar Universitário da Universidade de Pernambuco (CHU/UPE) were used to evaluate five predictive modeling approaches for ALS: a naïve baseline, ARIMA, LightGBM, XGBoost, and LSTM. Model evaluation was conducted using walk-forward validation, employing Mean Absolute Error, Root Mean Squared Error, and Mean Absolute Percentage Error as performance metrics. Statistical comparison was performed using the Model Confidence Set (MCS) test with a 10% significance level. The results indicate that no model showed superior performance, suggesting that the most appropriate approach depends on the temporal characteristics of each time series.*

1. Introduction

Health systems face the constant challenge of balancing the supply of services, which is limited by available resources, with growing public demand. In this context, different indicators are used to evaluate efficiency and support decision-making in hospital management [de Souza et al. 2016]. Among them, the average length of stay (ALS) stands out as an important metric for monitoring bed utilization, since hospital beds are one of the most critical resources and their occupancy is dynamic, varying over time [Ventura et al. 2025]. According to the study conducted by Cunha et al. [Cunha et al. 2023], high ALS values are associated with higher costs and may affect the quality of care. However, the simplistic strategy of increasing capacity without considering variability tends to be inefficient, since different factors can influence the behavior of this indicator, even within the same hospital network.

In this context, predictive analytics has become an important tool for strategic planning in healthcare, enabling the anticipation of variations in average length of stay (ALS) and contributing to the optimization of bed allocation and the use of hospital resources. The review conducted by Urbi [Urbi and Tiva 2025] highlights the growing role of artificial intelligence in hospital management, emphasizing its potential to improve operational efficiency, support decision-making, and strengthen the performance of healthcare systems. In the field of time series forecasting, different approaches have been proposed over the years, ranging from traditional statistical methods to machine learning techniques and deep learning models. As highlighted by Cheng [Cheng et al. 2025], the evolution of this field has been marked by the development of different modeling approaches, driven by the need to represent temporal patterns and generate forecasts.

Building upon these advances, the present study conducts a comparative analysis of representative models from different modeling paradigms, including ARIMA, XGBoost, LightGBM, and LSTM, for forecasting the average length of stay in three units of the CHU/UPE: the Centro Integrado de Saúde Amaury de Medeiros (CISAM), the Hospital Universitário Oswaldo Cruz (HUOC), and the Pronto-Socorro Cardiológico Universitário de Pernambuco Professor Luiz Tavares (PROCAPE). Using historical hospital production data, the study aims to evaluate the predictive performance of these models through error metrics and statistical tests, in order to identify which techniques have a greater ability to capture the temporal patterns present in the analyzed time series.

This article is structured as follows: Section 2 presents the related works for this study; Section 3 describes the methodology employed; Section 4 presents the results obtained in the experiments conducted; and, finally, Section 5 presents the conclusions and suggestions for future work.

2. Related works

Time series modeling has been increasingly applied to healthcare data for forecasting and decision support. This section reviews studies relevant to the present work, considering both methodological and contextual aspects.

Regarding the application of ARIMA models in healthcare forecasting, Ospina et al. [Ospina et al. 2023] proposed an ARIMA-based predictive model during the COVID-19 pandemic, using data from Recife, Brazil. Although the pandemic period was characterized by structural breaks not directly comparable to the regular operational dynamics addressed in the present study, the methodological contribution remains relevant, as the study demonstrated the ability of ARIMA models to generate accurate short-term forecasts. The authors employed an automated parameter optimization process based on autocorrelation and partial autocorrelation plots, a procedure also adopted in the present study through the `auto.arima` function. However, the study observed that model performance deteriorated as the forecasting horizon extended, revealing a limitation shared with the approach proposed here.

In the context of hospital occupancy forecasting, Cunha [Cunha 2026] analyzed the occupancy rate of a private hospital in Sergipe using monthly patient-day and bed-day time series, segmented by accommodation type. The study combined exploratory analysis, stationarity tests, and classical time series modeling, such as ARIMA and ETS/Holt-Winters, as well as a machine learning model based on XGBoost, with the aim of supporting capacity planning and bed management. The results indicated that model performance varies according to the type of accommodation: XGBoost generally showed better predictive performance for the Apartment and ICU series, while the ETS(A,N,A) model proved more suitable for the Ward, especially in capturing the seasonal pattern.

In the same context as Cunha [Cunha 2026] work, Santos [Santos et al. 2025] evaluates different models applied to monthly time series in Brazil, constructed by integrating SUS hospital production data with epidemiological notifications. Classical statistical models and machine learning methods were compared, including lag-based approaches, Random Forest, and SARIMA models. The results indicate that Multiple-Lag Linear Regression achieved the best predictive performance, outperforming the other methods in metrics such as RMSE, MAE, and MAPE. In contrast, models such as

SARIMA showed sensitivity to extreme values, while more complex approaches like Random Forest and neural networks did not yield consistent gains, mainly due to the limitations of monthly granularity and the small size of the series. As a main limitation, the study highlights that the low temporal resolution and short time span of the dataset hinder the effective use of more complex models such as LSTM and 1D CNNs.

In the use of Long Short-Term Memory (LSTM) neural networks for prediction, Fontoura [Fontoura 2024] proposes a multilayer LSTM-based approach for forecasting new daily COVID-19 cases. As in the case of Ospina et al. [Ospina et al. 2023], the epidemic context introduces volatility patterns that differ from the regular operational dynamics of hospital indicators, still the reported model behavior offers useful methodological insight. The results indicate that the model performs well for prediction horizons of up to 15 days, especially during periods of greater stability in the time series. However, performance deteriorates in high-volatility scenarios, such as epidemic peaks, and it is also sensitive in situations with values close to zero. The limitations include the high computational cost of training, dependence on the quality of epidemiological data, and reduced performance under abrupt regime changes.

Taken together, the studies analyzed highlight the relevance of time series methods applied to healthcare, especially for hospital indicators, emphasizing the importance of predictive approaches in supporting bed management and healthcare planning. At the same time, it is observed that, although there are investigations involving classical statistical models and machine learning techniques, there is still room for more comprehensive comparative evaluations that consider different model families under the same experimental protocol. In this context, the present study contributes by expanding this perspective through the comparison of statistical and machine learning approaches, evaluated using walk-forward validation. This strategy enables a more consistent analysis of predictive performance in sequential scenarios, contributing to decision-making support in hospital management.

3. Methodology

This study is characterized as quantitative research based on the analysis of hospital data, focusing on the average length of stay. The methodological process adopted is presented in Figure 1.

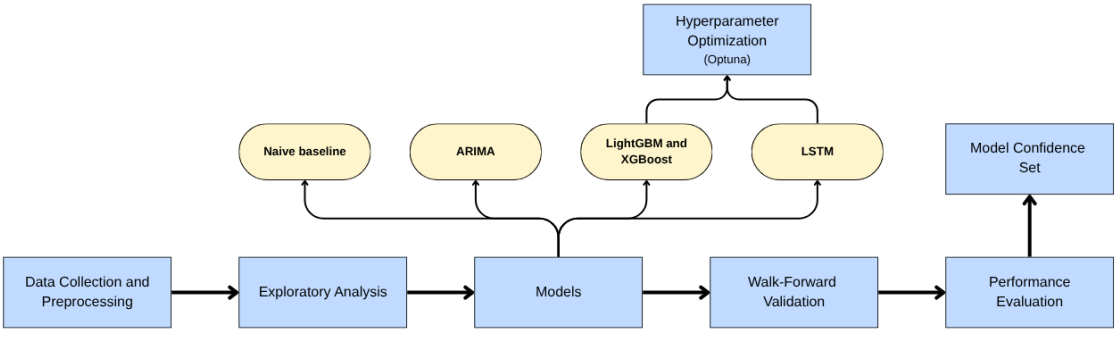


Figure 1. Methodological process

3.1. Database, preprocessing and exploratory analysis

The database used in this study was obtained from TABNET, the tabulation system of DataSUS (the Department of Informatics of the Brazilian Unified Health System). The extracted information consists of monthly records covering the period from January 2018 to December 2025, referring to hospital production related to bed utilization by specialty in three units of the CHU/UPE.

Since the aim of the study is to analyze the average length of stay (ALS), time series were constructed for this indicator, calculated as the ratio between length of stay, corresponding to the total number of hospitalization days, and frequency, referring to the number of admissions, for each month of the analyzed interval, within each specialty. As the units present several specialties, reflecting different types of beds, the specialty with the highest demand in each unit was selected for analysis, resulting in three time series.The series were organized chronologically and subjected to consistency checks, including the identification of missing values and duplicate monthly records, considering that each series should contain 96 monthly observations.

In the exploratory analysis stage, the series were represented by line graphs and histograms, with the aim of inspecting temporal patterns, data distribution, and possible deviations from normality. Monthly seasonality was investigated using the Kruskal-Wallis test, applied independently to each of the three hospital series. Since no statistically significant differences were observed among the months ($p > 0.05$), the simple Naïve model was adopted as the baseline, in lieu of a Seasonal Naïve approach. Next, stationarity was assessed using the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests, in order to verify the suitability of the series for forecasting models.

3.2. Forecasting models

Four forecasting approaches were evaluated: a Naïve baseline, ARIMA, LightGBM/XGBoost and LSTM. This design allows the study to test, under a common experimental protocol, whether the added complexity of machine learning and deep learning approaches translates into measurable forecasting gains over simpler alternatives. The Naïve model was used as a reference benchmark. In this approach, the forecast for each time step is equal to the most recent observed value. For ARIMA, the parameters (p,d,q) were automatically selected at each iteration using the *auto.arima* function from the forecast package in R [Hyndman and Khandakar 2008], using the Akaike Information Criterion (AIC).

For the LightGBM and XGBoost models, each observation was represented by features derived from past values, including lags (1, 2, 3, 6, and 12 months), moving-window statistics (mean and standard deviation over 3 and 6 months), cyclic encoding of the calendar month (sine and cosine), and a temporal variable to represent trend. For series classified as non-stationary in the ADF test ($p > 0.05$), first differencing was incorporated as an additional variable, without changing the target variable.

The LSTM network received as input univariate sequences of 12 consecutive months, normalized by *MinMaxScaler* fitted in each training window. The architecture consisted of two stacked LSTM layers with dropout, followed by an intermediate dense layer. Training was performed using the Adam optimizer, MSE loss function, and early stopping with a patience of 10 epochs.

The hyperparameters of LightGBM, XGBoost, and LSTM were optimized with the Optuna framework [Akiba et al. 2019], based on the Tree-structured Parzen Estimator (TPE) algorithm and internal validation through TimeSeriesSplit with 5 folds. A total of 60 trials were executed for LightGBM and XGBoost, and 20 for LSTM, due to the higher computational cost of the latter.

3.3. Validation and performance evaluation

Model evaluation was performed through progressive validation (walk-forward) [Wahyuddin et al. 2025]. The first 72 observations were used for initial training, and the remaining 24 for testing, preserving chronological order. At each iteration, the training window was expanded and a one-step-ahead forecast was generated. All models were re-estimated at each step, except for LSTM, which was updated every six iterations due to its higher computational cost. In this case, only the normalization was readjusted in the intermediate steps. Performance was evaluated using the MAE, RMSE, and MAPE metrics. Based on the errors obtained at the evaluation points, the Model Confidence Set (MCS) [Hansen et al. 2011] test was applied, with 5000 bootstrap resamples and significance level $\alpha = 0.10$, to identify the set of models with statistically equivalent predictive performance.

4. Results

Figure 2 presents the temporal behavior of the analyzed series through line charts and histograms. Distinct patterns can be observed among the specialties of the units over the period. CISAM, for instance, shows lower variability and a slight downward trend over time, indicating greater stability of the indicator. HUOC, in turn, exhibits a declining trend up to a certain period, followed by a recent sharp increase, suggesting a possible structural change in the series behavior. PROCAPE, on the other hand, displays an oscillatory pattern, with frequent fluctuations and no clear trend.

Based on these characteristics, stationarity tests were applied to the series, adopting a significance level of 5%. The results, presented in Table 1, indicate that the series for CISAM and HUOC do not fully satisfy the stationarity condition, whereas the series for PROCAPE was classified as stationary. For HUOC, the ADF and KPSS tests yielded conflicting results, which may reflect the structural change observed in the series. Given this ambiguity, the series was treated as non-stationary. In machine learning-based models, such as LightGBM and XGBoost, the non-stationary series received the first difference of the series as an additional feature, aiming to incorporate information about local variations without altering the original scale of the target variable. For the ARIMA model, differencing was handled automatically during the fitting process.

Table 1. Stationarity tests for the original series

Unit	Specialty	ADF Stat.	ADF p-value	KPSS Stat.	KPSS p-value
CISAM	Obstetric	-1.732	0.4148	1.4321	0.0100
HUOC	Clinical	-1.461	0.5527	0.2865	0.1000
PROCAPE	Clinical	-3.368	0.0121	0.3913	0.0809

Table 2 presents the best hyperparameters found by Optuna for the XGBoost and LightGBM models. For the Long Short-Term Memory (LSTM) model, the search space

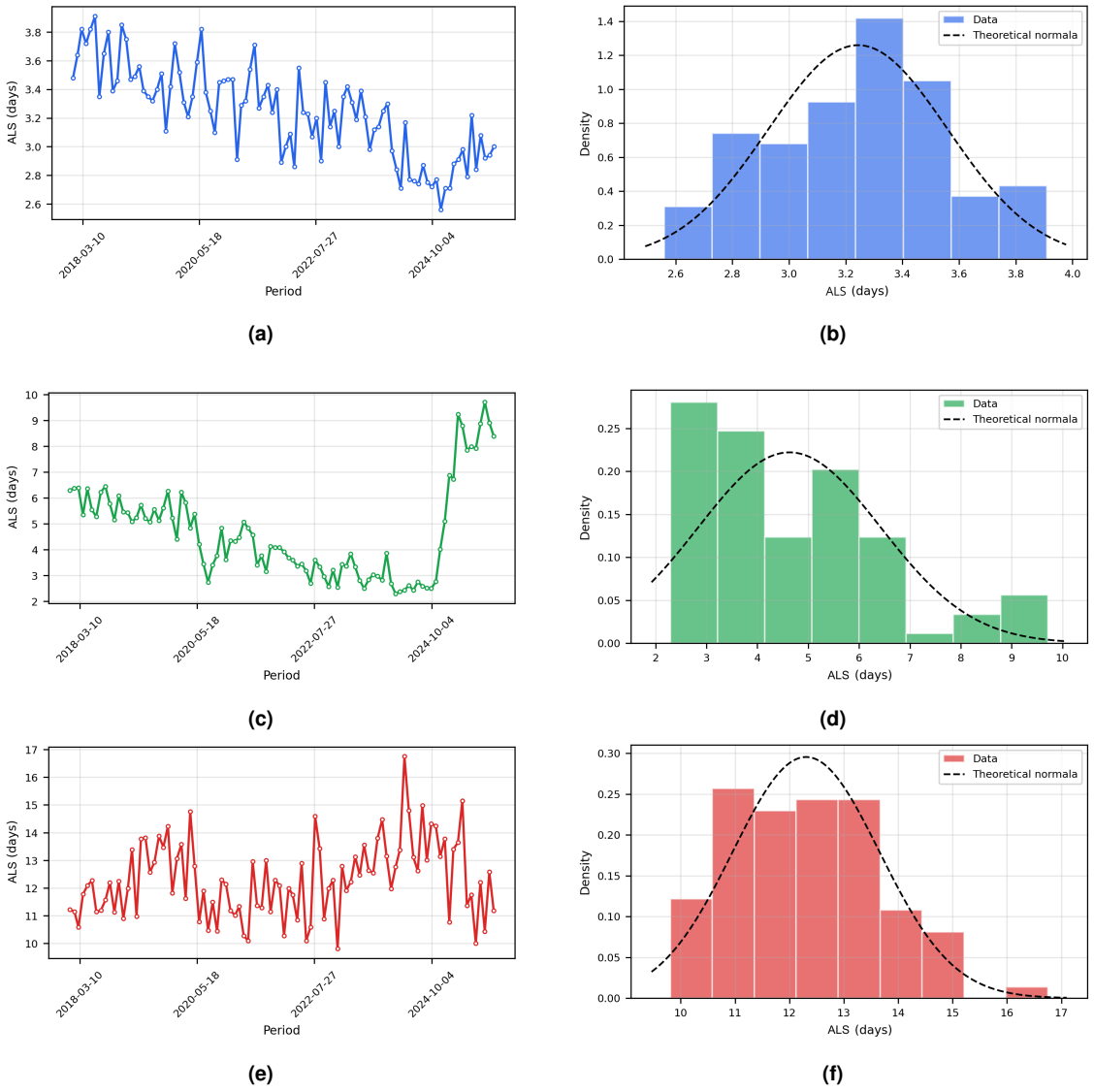


Figure 2. Temporal behavior and distribution of the selected specialty in each hospital unit: (a)-(b) CISAM (Obstetric), (c)-(d) HUOC (Clinical), and (e)-(f) PROCAPE (Clinical). Subfigures (a), (c), and (e) show the time series, while (b), (d), and (f) present the corresponding histograms.

was restricted to three hyperparameters, number of units, dropout rate, and batch size, which resulted in identical configurations across the series after the optimization process. The selected values were: 32 units in the LSTM layer, a dropout rate of 0.279598, and a batch size equal to 4.

The performance of the models evaluated in the walk-forward validation process is presented in Table 3, while Figure 3 illustrates the predicted versus observed values during the test period for each unit. It can be observed that there was no universally superior model across the three analyzed series, indicating that the suitability of each method depends on the specific characteristics of each hospital unit. In CISAM, the ARIMA model showed the best performance across all metrics (MAE, RMSE, and MAPE), suggesting

Table 2. Optimized hyperparameters for XGBoost and LightGbm

Unit	Model	n_estimators	lr	depth/leaves	child	subsample	colsample	α	λ
CISAM	LightGBM	122	0.0217	14	18	0.694	0.569	0.345	0.157
CISAM	XGBoost	188	0.0166	2	4	0.516	0.779	0.848	1.251
HUOC	LightGBM	338	0.1723	57	6	0.879	0.543	0.002	0.0004
HUOC	XGBoost	407	0.1449	2	6	0.585	0.609	0.029	0.0002
PROCAPE	LightGBM	142	0.1769	30	17	0.519	0.665	4.393	0.0031
PROCAPE	XGBoost	269	0.0224	6	1	0.987	0.502	0.0007	0.0159

that the temporal dynamics of this series are predominantly linear and can be adequately represented by classical statistical approaches. For the HUOC series, the Naïve model achieved the lowest error, indicating greater temporal persistence, in which the most recent value already represents a good approximation for the next period. In this case, the use of more complex models did not provide gains that would justify their application. Finally, in the PROCAPE series, the machine learning-based models showed better performance, with XGBoost standing out. This result suggests the presence of more complex or non-linear relationships, which were better captured by decision tree-based methods.

Table 3. Performance comparison of predictive models in the three hospital units (walk-forward validation, 24 test points)

Unit	Model	MAE	RMSE	MAPE (%)
CISAM - Obstetric	Naïve	0.1631	0.2146	5.61
CISAM - Obstetric	ARIMA	**0.1608**	**0.1976**	**5.59**
CISAM - Obstetric	LightGBM	0.2162	0.2630	7.72
CISAM - Obstetric	XGBoost	0.2212	0.2690	7.92
CISAM - Obstetric	LSTM	0.1668	0.2050	5.92
HUOC - Clinical	Naïve	**0.5998**	**0.8587**	**11.42**
HUOC - Clinical	ARIMA	0.6615	0.9745	12.49
HUOC - Clinical	LightGBM	0.8720	1.1959	17.58
HUOC - Clinical	XGBoost	0.8658	1.2242	18.23
HUOC - Clinical	LSTM	2.0801	2.7256	31.76
PROCAPE - Clinical	Naïve	1.5988	1.8680	12.59
PROCAPE - Clinical	ARIMA	1.3259	1.7003	10.45
PROCAPE - Clinical	LightGBM	1.2208	1.5996	9.57
PROCAPE - Clinical	XGBoost	**1.1876**	**1.5376**	**9.21**
PROCAPE - Clinical	LSTM	1.2671	1.6006	9.86

Figure 4 presents the distributions of the absolute errors of the models for the three analyzed series. In CISAM, the ARIMA, Naïve, and LSTM models showed similar error distributions, with lower medians and smaller dispersion compared to the machine learning-based models. In contrast, LightGBM and XGBoost exhibited greater variability and higher median errors, indicating lower predictive stability for this series. In HUOC, it can be observed that the Naïve model presented the most concentrated error distribution and the lowest median, while the ARIMA model showed similar behavior, although with slightly higher dispersion. On the other hand, LightGBM and XGBoost showed greater variability and higher median errors. The LSTM model presented the worst performance among the evaluated models, with wide dispersion and high extreme values, reinforcing its lower predictive capability for this series. Finally, in PROCAPE, a greater balance

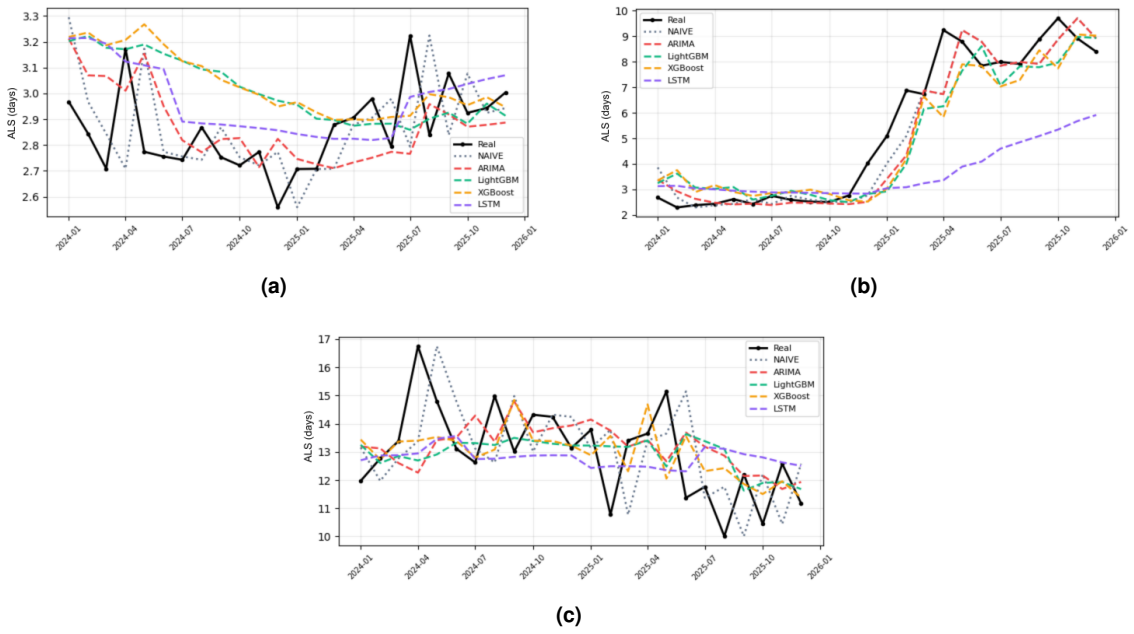


Figure 3. Observed and predicted values during the test period: (a) CISAM (b) HUOC, and (c) PROCAPE.

among the evaluated models can be observed. Although XGBoost achieved the best average performance, its error distribution was similar to that of the LightGBM and ARIMA models, suggesting less pronounced differences in terms of stability. The Naïve model showed a higher median and greater dispersion, while the LSTM presented intermediate performance, although with a higher occurrence of extreme values.

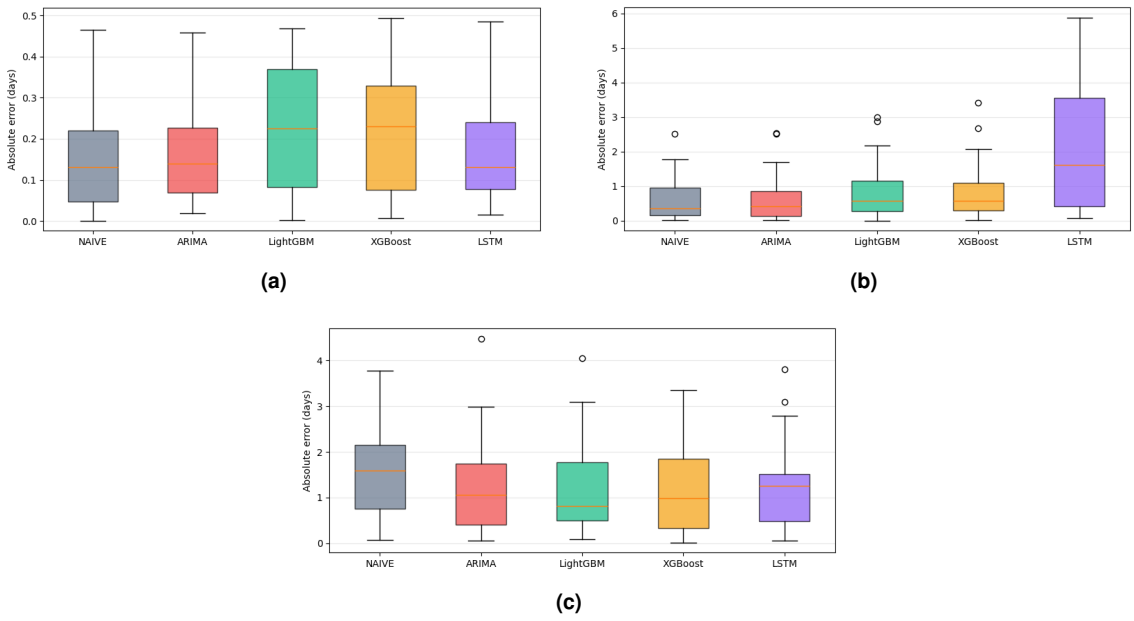


Figure 4. Distribution of absolute errors: (a) CISAM (b) HUOC, and (c) PROCAPE.

To complement the comparison among the models, the Model Confidence Set (MCS) test was applied at a 10% significance level in order to identify the set of models

that are statistically indistinguishable from the best-performing model in terms of mean absolute error, as shown in Table 4. The results show that, in CISAM, all models remained in the confidence set, indicating the absence of statistically significant differences among them. In HUOC, only the Naïve and ARIMA models were retained in the confidence set, while LightGBM, XGBoost, and LSTM were excluded. This result reinforces that, for this series, simpler approaches achieved superior or equivalent performance, with no evidence of statistical gains from more complex models. In turn, in PROCAPE, the Naïve model was excluded from the confidence set, while ARIMA, LightGBM, XGBoost, and LSTM remained. Although XGBoost achieved the lowest mean error, the test indicates that its performance was not statistically superior to the other models included in the set.

Table 4. MCS test results

Model	p-MCS (CISAM)	p-MCS (HUOC)	p-MCS (PROCAPE)
Naïve	0.9512	1.0000	0.0062
ARIMA	1.0000	0.2962	0.5082
LightGBM	0.3478	0.0092	0.8180
XGBoost	0.2972	0.0092	1.0000
LSTM	0.9512	0.0092	0.5596

5. Final Considerations

This study aimed to compare predictive models for the average length of stay (ALS) in three units of the CHU/UPE, based on historical hospital production data related to bed utilization. For this purpose, five predictive models were evaluated (Naïve, ARIMA, LightGBM, XGBoost, and LSTM), combining classical statistical methods and machine learning techniques. The comparison was carried out using walk-forward validation with an expanding window and statistically assessed using the Model Confidence Set (MCS) test.

The results indicate that no model was universally superior across the analyzed units, highlighting that the choice of approach should be guided by the temporal characteristics of each series. ARIMA stood out in CISAM and HUOC, suggesting that the dynamics of these series are predominantly linear. In HUOC, in particular, the occurrence of an abrupt structural change limited the performance of all models. In PROCAPE, tree-based models showed better numerical performance, although without statistically significant differences compared to the others.

For future work, it is suggested to extend the analysis period to enlarge the test set and increase the statistical power of the tests, a relevant limitation of this study. Additionally, it is recommended to investigate, in collaboration with clinical literature, how specialty-specific patient behavior shapes ALS dynamics, since this understanding could guide the selection of forecasting approaches more effectively than a purely numerical comparison. It is recommended to explore new variables that may influence the target indicator, such as patient profiles and clinical factors.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2623–2631, Anchorage, AK, USA. Association for Computing Machinery.

Cheng, M., Liu, Z., Tao, X., Liu, Q., Zhang, J., Pan, T., Zhang, S., He, P., Zhang, X., Wang, D., Wang, J., and Chen, E. (2025). A comprehensive survey of time series forecasting: Concepts, challenges, and future directions. *TechRxiv*, 2025(0410).

Cunha, I. M., Mendes, M. C. B., Beserra, J. V. F., Bolognani, C. V., Göttems, L. B. D., and de Carvalho, E. M. P. (2023). Factors that contribute to the increase in the length of hospital stay. *New Trends in Qualitative Research*, 18:e835.

Cunha, J. M. M. (2026). Análise e previsão da taxa de ocupação hospitalar em um hospital privado de sergipe. Trabalho de Conclusão de Curso (Graduação em Ciências Atuariais), Universidade Federal de Sergipe, São Cristóvão, SE, Brasil.

de Souza, A. A., da Silva, O. F., Avelar, E. A., and Lamego, L. F. (2016). Análise de eficiência dos hospitais: Um estudo com foco em indicadores operacionais. Technical report, Universidade Federal de Minas Gerais.

Fontoura, M. A. R. (2024). Modelagem preditiva de surtos epidêmicos usando redes neurais LSTM: Uma aplicação para a COVID-19. Dissertação de mestrado em informática e gestão do conhecimento, Universidade Nove de Julho (UNINOVE).

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3):1–22.

Ospina, R., Gondim, J. A. M., Leiva, V., and Castro, C. (2023). An overview of forecast analysis with arima models during the covid-19 pandemic: Methodology and case study in brazil. *Mathematics*, 11(14).

Santos, B. N. G., de Almeida Ferreira Azevedo, G., Carneiro, L. B., dos Santos, R. B. G., Nogueira, S. E. F., de Brito Simões, M. M., and de Brito Boechat, G. C. M. (2025). Previsão de ocupação em hospitais públicos utilizando modelos de séries temporais e aprendizado de máquina. Trabalho de Conclusão de Curso (Graduação), UCSal – Universidade Católica do Salvador.

Urbi, S. R. C. and Tiva, M. G. (2025). Technology and innovation in healthcare: Adoption of ai and predictive analytics in hospital management. *Pathfinder of Research*, 3(2):22–45.

Ventura, T., Gomes, R., Pedroso, G., Vecchiato, D., and Rezende, R. (2025). Estimativa da ocupação de leitos para tratamento da covid-19 usando dados temporais. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 24–31, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação, SBC.

Wahyuddin, E. P., Caraka, R. E., Kurniawan, R., Caesarendra, W., Gio, P. U., and Pardamean, B. (2025). Improved lstm hyperparameters alongside sentiment walk-forward validation for time series prediction. *Journal of Open Innovation: Technology, Market, and Complexity*, 11(1):100458.