



Hybrid ARIMA-Prophet-LSTM with SHAP Interpretation for Forecasting Outpatient Visits by Medical Specialty

Vlandimir Welem L.*

Universitas Bina Nusantara,
Indonesia

Sani Muhamad Isa

Universitas Bina Nusantara,
Indonesia

***Corresponding author:**

Vlandimir Welem L,
Universitas Bina Nusantara, Indonesia.

✉ vlandimir.lainsamputty@binus.ac.id

Article Info:

Article history:

Received: August 07, 2026

Revised: September 07, 2026

Accepted: September 18, 2026

Available Online: October 14, 2026

Keywords:

ARIMA;

Hybrid Forecasting;

LSTM;

Outpatient Visit Forecasting;

SHAP



This is an open access article under the
CC BY-SA license.

Copyright © 2026 by Author. Published by
Politeknik Siber Cerdika International

Abstract

Background: Accurate forecasting of outpatient visits is important for hospital staffing, scheduling, facility utilization, and resource allocation. Outpatient demand in the study data shows strong weekly seasonality and holiday-related variation, creating challenges for both forecasting accuracy and model interpretation.

Objective: This study evaluates ARIMA, Prophet, LSTM, and four hybrid configurations for daily outpatient-visit forecasting across five medical specialties and interprets the best-performing hybrid model using SHAP.

Method: Historical daily outpatient-visit data from Internal Medicine, Orthopedic and Traumatology Surgery, Cardiology, Pediatrics, and Neurology were modeled using ARIMA, Prophet, LSTM, ARIMA-Prophet, ARIMA-LSTM, Prophet-LSTM, and ARIMA-Prophet-LSTM. Time-series cross-validation with an expanding window was used, with MAE, RMSE, and MAPE as performance metrics. Friedman and post hoc Nemenyi tests were applied for model comparison, and SHAP's GradientExplainer was used for the LSTM component of Prophet-LSTM.

Results: Prophet achieved the best performance in four specialties, whereas Prophet-LSTM was best in Neurology. The triple hybrid was significantly worse than Prophet ($p = 0.017$) and did not differ significantly from ARIMA-Prophet ($p = 0.9996$). SHAP identified day-of-week cyclical features as the most influential predictors across specialties, while national holidays showed low global importance but were associated with large errors in specific cases.

Conclusion: Greater model complexity did not improve forecasting accuracy in these data. Prophet provided the strongest overall performance, while Prophet-LSTM remained a competitive, interpretable hybrid alternative.

To cite this article: Vlandimir, W. L., & Isa, S. M. (2026). Hybrid ARIMA-Prophet-LSTM with SHAP interpretation for forecasting outpatient visits by medical specialty. *Equivalent: Jurnal Ilmiah Sosial Teknik*, 8(4), 1750–1769. <https://doi.org/10.59261/jequi.v8i4.427>

INTRODUCTION

The increasing complexity of healthcare services has strengthened the need for efficient resource allocation, responsive operational planning, and data-driven decision-making. Outpatient services represent one of the main interfaces between hospitals and the communities they serve, making outpatient demand an important indicator for operational management. In the United States, approximately 1.0 billion physician office visits were reported in the National Ambulatory Medical Care Survey, corresponding to 320.7 visits per 100 persons, illustrating the substantial scale of ambulatory healthcare utilization (Chandekar et al., 2026; Statistics, 2017). Variations in outpatient demand can directly influence physician scheduling, nursing workloads, room utilization, diagnostic support services, and other hospital resources. Consequently, the ability to anticipate changes in outpatient volume is increasingly important for maintaining service efficiency and supporting evidence-based hospital management. Previous research has

likewise demonstrated that outpatient-demand forecasting can contribute to healthcare resource planning and operational decision-making ([Hariyanto, 2022](#); [Wei et al., 2022](#)).

Forecasting outpatient demand becomes more challenging when patient visits are examined at the level of individual medical specialties. Different specialties may display distinct patterns of weekly activity, seasonal variation, temporal dependence, and responses to calendar-related conditions. The outpatient data examined in this study demonstrate such heterogeneity. The five highest-volume specialties showed coefficients of variation ranging from 0.44 for Internal Medicine to 0.56 for Pediatrics. Weekly outpatient volumes also declined by approximately 12–18% during weeks containing one national holiday and by approximately 28–45% during periods containing multiple holidays. These differences indicate that outpatient demand cannot be adequately represented as a uniform hospital-wide pattern. In addition, Internal Medicine, which had the highest outpatient volume among the selected specialties, recorded daily visits ranging from 1 to 273 patients. Such variation creates practical challenges for hospital managers because inaccurate demand estimates may lead to mismatches between patient volume and available personnel, consultation rooms, or supporting services. Therefore, specialty-level forecasting is necessary to provide more specific information for operational planning.

Previous studies have applied various time-series approaches to healthcare-demand forecasting. Autoregressive Integrated Moving Average (ARIMA) remains a widely used statistical method because of its ability to represent autoregressive relationships, moving-average structures, and linear temporal dependencies. [Hariyanto \(2022\)](#), for example, applied ARIMA to forecast outpatient visits at a healthcare facility in Sukoharjo, Indonesia, demonstrating its applicability to healthcare service planning. Time-series approaches have also been extended by incorporating external variables. [Zhang et al. \(2023\)](#) showed that meteorological and environmental factors can be associated with variations in healthcare utilization and may contribute to improved forecasting. However, conventional statistical models may become less suitable when the observed series contains nonlinear relationships, complex temporal dependencies, or patterns that cannot be sufficiently represented by linear structures alone. These limitations have encouraged the increasing use of machine-learning and deep-learning approaches in healthcare forecasting.

Prophet offers another forecasting approach that explicitly represents trend, recurring seasonality, and holiday effects. These characteristics are particularly relevant for outpatient data because healthcare utilization may vary according to weekly schedules, annual periods, and public holidays. [Taylor and Letham \(2018\)](#) and [Sprangers et al. \(2024\)](#) introduced Prophet as a decomposable forecasting framework designed to model such recurring temporal components. More recently, [Fattouh et al. \(2026\)](#) demonstrated the applicability of Prophet to hospital bed-occupancy forecasting, showing that interpretable weekly, yearly, and holiday-related patterns could be incorporated into hospital forecasting. The explicit representation of seasonality provides an advantage in settings where recurring calendar patterns dominate the data. Nevertheless, Prophet may not capture all nonlinear dependencies remaining after trend and seasonal components have been modeled, particularly when the residual structure contains relationships that cannot be represented adequately through additive components.

Deep-learning methods provide another alternative for modeling complex temporal relationships. Long Short-Term Memory (LSTM), originally introduced by [Hochreiter & Schmidhuber \(1997\)](#) and [Krichen & Mihoub \(2025\)](#), is a recurrent neural network architecture capable of learning sequential dependencies and nonlinear temporal structures. In healthcare forecasting, [Wang et al. \(2023\)](#) applied LSTM and Gated Recurrent Unit models to predict outpatient visits for conjunctivitis, demonstrating the potential of deep learning to support medical resource planning. [Mahmud et al. \(2025\)](#) further showed that combining ARIMA and LSTM could improve forecasting performance for COVID-19 time-series data compared with selected individual models. These findings indicate that statistical and deep-learning models may provide complementary forecasting capabilities. However, LSTM models generally have more complex internal structures than conventional statistical models, making their predictions more difficult to interpret for hospital managers and other nontechnical users ([Di Martino & Delmastro, 2023](#)).

These limitations have encouraged the development of hybrid forecasting architectures that combine different modeling paradigms. A hybrid model may allow one component to capture

linear or systematic temporal structures while another component learns residual nonlinear relationships. Mahmud et al. (2025), for instance, demonstrated the potential of ARIMA-LSTM hybridization, while Saini & Sharma (2024) examined Prophet-LSTM forecasting by combining Prophet's trend and seasonal modeling with LSTM's nonlinear learning capability. The studies cited in the present research therefore provide evidence that hybridization may improve forecasting flexibility. However, they do not simultaneously provide a systematic comparison of standalone ARIMA, Prophet, and LSTM models with their pairwise and triple-hybrid configurations using specialty-level outpatient data. Mahmud et al. (2025) focused on COVID-19 forecasting rather than outpatient demand across medical specialties, while Saini & Sharma (2024) focused on Prophet-LSTM modeling and preprocessing rather than a comprehensive comparison among multiple standalone and hybrid configurations. Fattouh et al. (2026), meanwhile, applied Prophet to hospital bed occupancy without incorporating a deep-learning residual model combined with SHAP-based feature interpretation. Accordingly, an empirical gap remains regarding whether progressively more complex hybrid architectures actually provide better predictive performance than simpler forecasting models for specialty-level outpatient demand.

The need for model interpretation creates an additional research issue. Predictive accuracy alone may be insufficient for healthcare applications because hospital managers need to understand the factors associated with model predictions before using them in operational decisions. Complex machine-learning and deep-learning approaches are often treated as black-box models because their internal prediction mechanisms are difficult to interpret directly. Explainable artificial intelligence (XAI) has therefore become increasingly relevant for healthcare decision-support systems by providing mechanisms for explaining how model inputs contribute to prediction outputs (Di Martino & Delmastro, 2023; Jung et al., 2023). Among available XAI techniques, SHapley Additive exPlanations (SHAP) provides feature-level explanations based on Shapley values and enables both global and local interpretation of model predictions (Liu et al., 2023; Lundberg & Lee, 2017). Previous reviews have identified SHAP as a useful interpretability technique for machine-learning applications in healthcare (Band et al., 2023; Shaikh et al., 2023). For outpatient forecasting, this capability is particularly relevant because temporal variables such as day of the week, month, holiday status, and recent observations may influence predicted visit volumes.

Based on these considerations, the novelty of the present study lies not in assuming that a more complex hybrid model is necessarily superior, but in providing a systematic empirical evaluation of seven forecasting configurations on specialty-level outpatient data. The evaluated configurations consist of three standalone models ARIMA, Prophet, and LSTM and four hybrid models ARIMA-Prophet, ARIMA-LSTM, Prophet-LSTM, and ARIMA-Prophet-LSTM. These models are compared under a time-series validation framework using MAE, RMSE, and MAPE, complemented by statistical comparisons to determine whether observed performance differences are meaningful. The study further integrates SHAP-based interpretation into the best-performing hybrid model containing a nonlinear component. This approach enables the research to examine two issues simultaneously: whether increasing architectural complexity improves forecasting accuracy and which temporal features contribute most strongly to the predictions generated by the interpretable hybrid framework.

The urgency of this research is therefore both methodological and practical. Methodologically, outpatient forecasting requires empirical evidence regarding whether combining statistical, seasonal, and nonlinear modeling components provides meaningful performance gains compared with simpler alternatives. A more complex model is not necessarily preferable if the underlying outpatient series is predominantly explained by recurring seasonal structures. Practically, hospitals require forecasting approaches that provide sufficiently accurate predictions while remaining interpretable and usable for operational planning. The specialty-specific characteristics observed in the study data further emphasize that forecasting performance should not be assumed to be uniform across outpatient departments. In addition, the observed reductions in weekly visits during national-holiday periods demonstrate the importance of explicitly considering calendar-related patterns when developing forecasting systems.

Accordingly, this study aims to build and systematically evaluate ARIMA, Prophet, LSTM, and their hybrid configurations for forecasting daily outpatient visits across five medical specialties. Specifically, the study examines whether the ARIMA-Prophet-LSTM triple hybrid provides higher predictive accuracy than standalone and simpler hybrid alternatives based on MAE, RMSE, and MAPE, and evaluates differences among the forecasting models using statistical comparison procedures. The study also applies SHAP to the best-performing hybrid model with a nonlinear component to identify the temporal features contributing to predictions at both global and local levels. Through this framework, the study is expected to contribute empirical evidence regarding the relationship between forecasting-model complexity and predictive performance while providing an interpretable basis for hospital personnel scheduling, facility utilization, specialty-level capacity planning, and healthcare-resource allocation.

METHOD

This study employed a retrospective quantitative design with a comparative time-series forecasting approach to predict daily outpatient visits by medical specialty. The analysis used historical outpatient-visit records from a private hospital in Jakarta, Indonesia, covering five specialties with the highest outpatient volumes: Internal Medicine, Orthopedic and Traumatology Surgery, Cardiology, Pediatrics, and Neurology. The dataset included temporal information such as day of the week, month, quarter, year, weekend status, and national holidays. Data preparation involved data integration and consistency checking, treatment of missing values and outliers, and construction of cyclical temporal features using sine-cosine transformations. Stationarity was assessed using the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, with differencing applied when necessary to obtain a stationary series.

Seven forecasting configurations were evaluated, consisting of three standalone models ARIMA, Prophet, and Long Short-Term Memory (LSTM) and four hybrid models ARIMA-Prophet, ARIMA-LSTM, Prophet-LSTM, and ARIMA-Prophet-LSTM. The hybrid models were developed using a residual-cascade approach in which subsequent models learned patterns remaining in the residuals generated by the preceding model. Model evaluation was performed using time-series cross-validation with an expanding-window scheme to preserve chronological order and minimize data leakage. Forecasting performance was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). For the overall comparison of the seven models, identical reduced three-fold cross-validation splits were used across all specialties to provide matched observations for statistical testing. Differences in model performance were evaluated using the Friedman test followed by post hoc Nemenyi comparisons.

To improve model interpretability, SHapley Additive exPlanations (SHAP) were applied to the LSTM component of the Prophet-LSTM model, which represented the best-performing hybrid configuration suitable for nonlinear interpretation. SHAP's GradientExplainer was used to evaluate the contributions of temporal features, including cyclical day-of-week and month variables, national-holiday indicators, and lagged Prophet residuals. Interpretation was conducted at both the global and local levels. Global analysis was used to identify the relative importance and directional contributions of features across predictions, while local analysis examined selected cases, such as peak and low outpatient visits, national holidays, large prediction errors, and substantial differences between hybrid and standalone model predictions. This approach enabled the forecasting results to be evaluated not only in terms of predictive accuracy but also in terms of the temporal factors contributing to individual model predictions.

RESULTS AND DISCUSSION

Stationarity Test Results (ADF Test)

The stationarity test was conducted using the Augmented Dickey-Fuller (ADF) test, complemented by the KPSS test as a cross-check. Both tests were selected because they have opposing null hypotheses: the ADF test has H_0 : the series has a unit root (nonstationary), whereas the KPSS test has H_0 : the series is stationary. Therefore, a consistent conclusion from both tests provides greater confidence than reliance on a single test alone, a practice recommended in econometric time-series analysis (Cousineau, 2025; Hyndman & Athanasopoulos, 2021; Kwiatkowski et al., 1992). The tests were performed at the level (without differencing), shown in Table 1.

Table 1*Results of the Augmented Dickey–Fuller (ADF) and KPSS Stationarity Tests*

Specialty	ADF statistic	ADF p value	ADF lags	ADF conclusion	KPSS statistic	KPSS p-value	KPSS conclusion	Recommended
Internal Medicine	-6.077	0	22	Stationary	0.099	0.100	Stationary	0
Orthopedic and Traumatology Surgery	-6.147	0	22	Stationary	0.442	0.059	Stationary	0
Cardiology	-7.314	0	22	Stationary	0.542	0.032	Non-stationary	1
Pediatrics	-5.516	0.000002	21	Stationary	0.254	0.100	Stationary	0
Neurology	-6.069	0	22	Stationary	0.351	0.098	Stationary	0

The ADF test rejected the null hypothesis for all five specialties ($p < 0.001$), indicating no evidence of a stochastic unit root based on the ADF results. The ADF statistics were more negative than the 5% critical value (-2.86) for all specialties. However, because the ADF and KPSS tests evaluate stationarity under different null hypotheses, the results of both tests were considered jointly before determining the appropriate differencing order for each specialty.

However, the KPSS test for Cardiology produced a p-value of 0.032 (< 0.05), rejecting the null hypothesis of stationarity and therefore differing from the ADF result for this specialty. After first differencing ($d = 1$), both the ADF test ($p < 0.001$) and the KPSS test ($p = 0.100$) supported stationarity. Accordingly, $d = 1$ was used for Cardiology in ARIMA model identification, whereas $d = 0$ was retained for Internal Medicine, Orthopedic and Traumatology Surgery, Pediatrics, and Neurology.

These results provide the empirical basis for determining the differencing order (d) in the per-specialty ARIMA model identification and underscore the importance of testing stationarity individually for each specialty rather than assuming a uniform order across all models.

Identification of Seasonal Patterns and Holiday Effects

This subsection extends the weekly-pattern findings from the STL Decomposition section by focusing specifically on quantifying the effects of national holidays, including possible bridging effects, whereby visit patterns increase or decrease on the days surrounding a holiday. A more meaningful comparison is total weekly visits across weeks with no holiday, weeks with one holiday, and weeks with two or more holidays (e.g., the Eid al-Fitr period close to Nyepi/Labor Day):

Table 2*Weekly Seasonal Patterns and National Holiday Effects Across Outpatient Specialties*

Specialty	Avg. regular week	Avg. week w/ 1 holiday	Avg. week w/ 2 holidays	Decrease w/ 1 holiday (%)	Decrease w/ 2 holidays (%)
Internal Medicine	886.34	782.2	550	-11.7	-37.9
Orthopedic and Traumatology Surgery	384.55	329.46	212.25	-14.3	-44.8
Cardiology	372.33	322.91	226.25	-13.3	-39.2
Pediatrics	340.03	281.14	244.75	-17.3	-28
Neurology	316.2	257.97	198.5	-18.4	-37.2

Based on 118 regular weeks, 35 weeks with one holiday, and 4 weeks with two or more holidays, a proportional and realistic pattern emerges: losing one working day to a national holiday reduces total weekly visits by 12–18%, while two adjacent holidays (e.g., Eid al-Fitr) reduce visits by 28–45%. Interestingly, the decrease is not exactly $1/6$ ($\approx 16.7\%$) or $2/6$ ($\approx 33.3\%$), which would be expected if the other days of the week were unaffected, indicating a partial redistribution effect: some patients who would have visited on the holiday appear to shift to the nearest working day within the same week, although this effect is not strongly detected as a directional daily $H-1/H+1$ effect in the daily analysis presented in the Holiday Effects section.

Pediatrics shows the clearest indication of redistribution: the one-holiday decrease of 17.3% slightly exceeds the ideal proportion of 16.7%; however, the two-holiday decrease is only 28.0%, well below the ideal proportion of 33.3%. This suggests that some pediatric visits delayed by the longer holiday period were shifted to other working days within the same week.

This weekly comparison indicates a substantial holiday-related reduction in outpatient visits, although the magnitude of the reduction is not proportional to the number of holiday days. Prophet's built-in holiday-effects component, which models holiday effects at the individual-date level, therefore remains relevant for representing holiday-related variation. Nevertheless, forecasting performance during weeks containing multiple holidays, such as the Eid al-Fitr period, should be interpreted cautiously because these periods were associated with larger changes in outpatient volume.

Baseline Model Results and Evaluation

ARIMA Modeling Results

ARIMA order identification followed the Box–Jenkins methodology: the differencing order (d) was determined from the ADF/KPSS test results in the Stationarity Test section ($d = 1$ specifically for Cardiology and $d = 0$ for the other four specialties), while the autoregressive (p) and moving-average (q) orders were identified through a combination of visual ACF/PACF analysis (Figures 4.6–4.10) and an AIC-based grid search ($p, q = 0-5$) on the initial training set (the first 766 days, 70% of the data).

The initial grid search showed that the lowest-AIC order for several specialties had AR/MA polynomial roots nearly on the unit circle (modulus close to 1.000), causing the model to fail to converge numerically and producing multi-step predictions that were not reliably mean-reverting. Instead of decaying toward the mean, these predictions exhibited persistent oscillations resembling the training pattern. Therefore, the order-selection procedure was revised by adding convergence and root-stability filters (AR and MA root modulus > 1.05), selecting the lowest-AIC model among candidates that were both convergent and stable:

Table 3

ARIMA Model Identification and Diagnostic Results

Specialty	d	ARIMA order (p, d, q)	AIC	Converged	Min AR root modulus	Min MA root modulus	Residual white noise (5%)
Internal Medicine	0	(5, 0, 0)	8490.74	Yes	1.1167	-	No
Orthopedic and Traumatology Surgery	0	(4, 0, 2)	7485.85	Yes	1.2331	1.1751	No
Cardiology	1	(5, 1, 1)	7179.24	Yes	1.0862	1.4227	No
Pediatrics	0	(2, 0, 4)	7310.97	Yes	1.3628	1.0768	No
Neurology	0	(5, 0, 1)	7073.24	Yes	1.1108	3.7516	No

All models now converge fully, with roots well away from the unit circle, producing numerically stable predictions. Nevertheless, the Ljung–Box test still rejects the white-noise null hypothesis for all five specialties ($p < 0.001$ at both lag 10 and lag 20); the residuals still contain systematic patterns that have not been captured, consistent with the research design that deliberately delegates the seasonal component to Prophet. Evaluation used a TSCV expanding-window scheme (initial training of 766 days, a 30-day horizon, and 10 folds per specialty):

Table 4

Forecasting Performance of the ARIMA Model

Specialty	Order	MAE	RMSE	MAPE %
Internal Medicine	(5, 0, 0)	51.5	62.84	749.08
Orthopedic and Traumatology Surgery	(4, 0, 2)	27.78	33.29	40.81
Cardiology	(5, 1, 1)	24.29	28.27	50.69
Pediatrics	(2, 0, 4)	20.76	26.12	155.06
Neurology	(5, 0, 1)	20.97	25.46	40.73

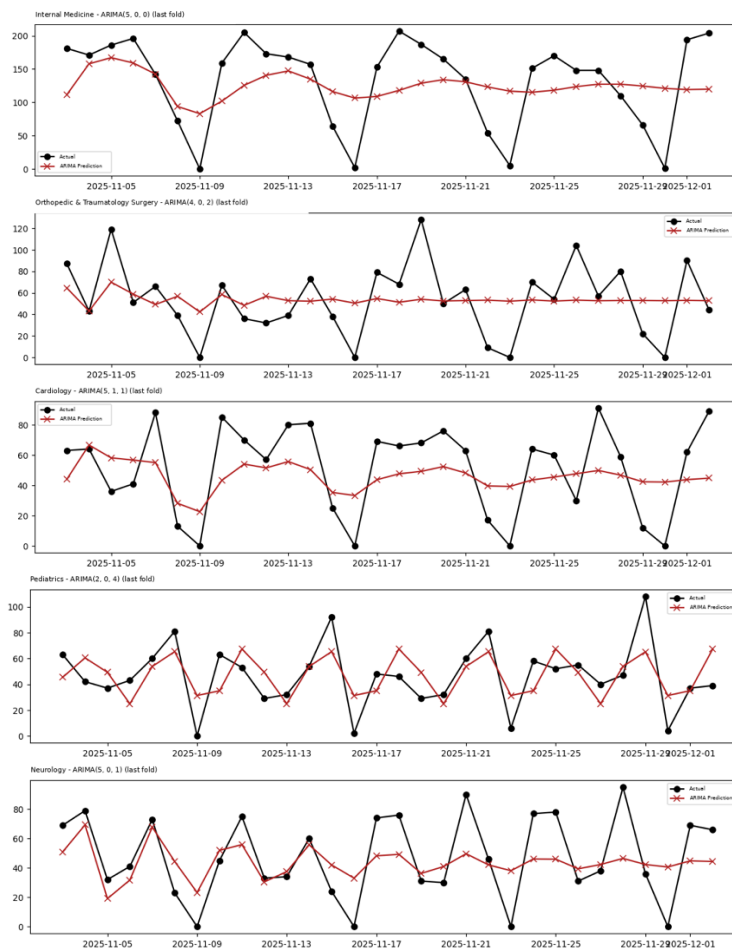


Figure 1. ARIMA Forecasting Results for Five Outpatient Specialties

Compared with the lowest-AIC order without the stability filter (reported in the initial iteration of this subsection), applying the convergence/stability filter consistently increased all error metrics across the five specialties (e.g., the Orthopedic and Traumatology Surgery RMSE increased from 21.94 to 33.29, and the Cardiology MAPE increased from 27.0% to 50.7%). This is not an implementation error but rather direct empirical evidence that the discarded near-unit-root structure provided the only way for nonseasonal ARIMA to approximate the highly persistent weekly cycle in the data (ACF remains high, approximately 0.65–0.75, up to lag 35 without meaningful decay; Figure 1). In other words, pure ARIMA faces an unavoidable trade-off when the data exhibit such strong seasonality: sacrificing numerical stability for accuracy or sacrificing accuracy for stability and more reliable long-term forecasts. This study prioritizes stability (the second option), because non-mean-reverting predictions cannot be relied upon for operational planning, even though the consequence is that the ARIMA baseline reported here represents a conservative but robust lower bound on performance.

This finding reinforces the core justification for the study: the structural limitation of nonseasonal ARIMA in handling a strong weekly cycle is the direct motivation for the hybrid architecture, in which Prophet is expected to capture the seasonal component that ARIMA proved difficult to represent without sacrificing model stability.

Prophet Modeling Results

The Prophet model was configured with a national-holiday calendar containing 52 dates from the actual `is_libur_nasional` data, as well as additional regressors in the form of cyclical encoding (sine/cosine terms for day of week and month), alongside Prophet's built-in weekly and yearly seasonal components. Hyperparameters were tuned through a 64-combination grid search on the validation set (the last 60 days of the initial training set), with the configuration having the lowest MAPE selected:

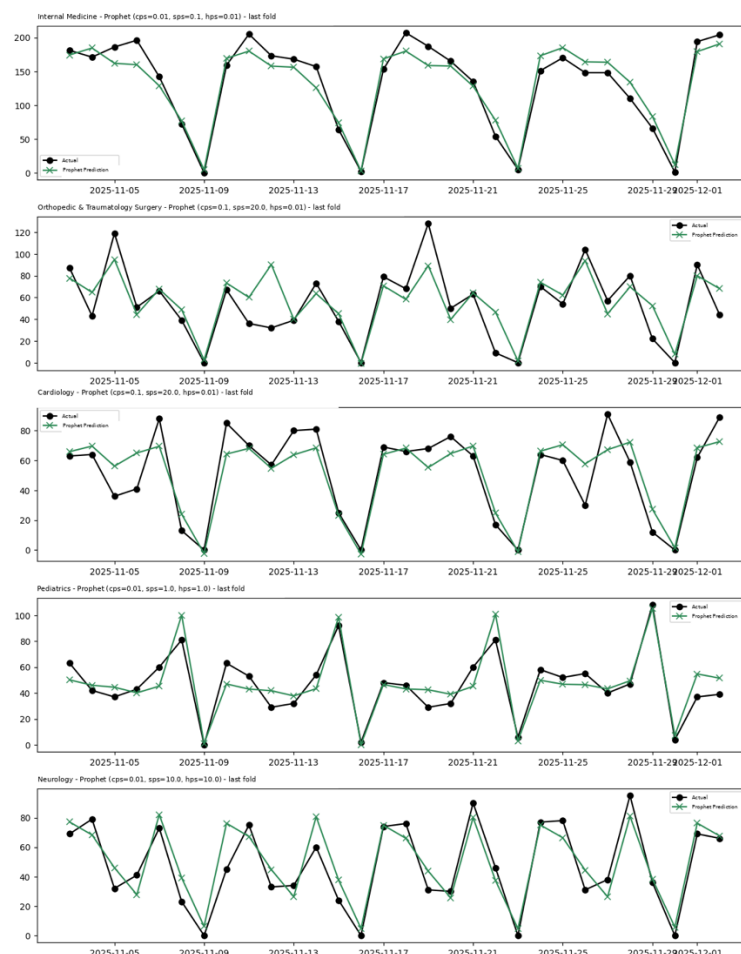
Table 5*Optimal Prophet Hyperparameters Selected by Grid Search*

Specialty	changepoint prior scale	seasonality prior scale	holidays prior scale	Validation MAPE (%)
Internal Medicine	0.01	0.1	0.01	62.62
Orthopedic and Traumatology Surgery	0.1	20	0.01	38.47
Cardiology	0.1	20	0.01	24.66
Pediatrics	0.01	1	1	27.85
Neurology	0.01	10	10	23.42

The four specialties other than Internal Medicine consistently selected a low changepoint_prior_scale (0.01–0.10), indicating a relatively stable trend with limited changes in trend rate, consistent with the STL findings in the STL Decomposition section, which showed no sharp growth or decline. In contrast, the selected seasonality_prior_scale varied widely (0.1–20.0) across specialties, indicating that the strength of the seasonal component differs substantially by specialty, consistent with the varied monthly patterns found in Section 4.1.2. Using a TSCV expanding-window scheme identical to that in the preceding section (for a fair comparison):

Table 6*Forecasting Performance of the Prophet Model*

Specialty	Mean MAE	Mean RMSE	Mean MAPE (%)
Internal Medicine	22.84	31.45	63.28
Orthopedic and Traumatology Surgery	14.32	18.94	31.26
Cardiology	11.45	16.01	29.4
Pediatrics	9.42	13.34	29.5
Neurology	11.5	15.25	33.49

**Figure 2.** Prophet Forecasting Results for Five Outpatient Specialties

Prophet consistently outperformed ARIMA across all specialties and all metrics, with RMSE improvements of 40–50% and MAPE improvements of 18–92%. The most pronounced improvement occurred in Internal Medicine (a 91.6% reduction in MAPE), a result consistent with the observation that ARIMA's inflated MAPE for this specialty was associated with its failure to capture the weekly cycle (ARIMA Results section). Prophet's explicit holiday and seasonal components directly addressed this structural weakness. This finding supports the hypothesis established in Section 4.2.1: nonseasonal ARIMA is inadequate for data with such a strong weekly cycle, whereas Prophet, which is specifically designed to model trend, seasonality, and holiday effects, substantially addresses this limitation.

Nevertheless, Prophet's MAPE for Internal Medicine (63.3%) remains considerably higher than those of the other four specialties (approximately 29–33%), indicating that a nonlinear residual component remains insufficiently captured by either ARIMA or Prophet for this highest-volume specialty, making it a potential candidate for improvement by the LSTM component in the subsequent hybrid modeling stage.

ARIMA-Prophet-LSTM Hybrid (Triple) Model Results

The Triple Hybrid is a full three-stage cascade, exactly as shown in the earlier general diagram: ARIMA (Stage 1) is fitted to the raw data, Prophet (Stage 2) is fitted to residual_arima, and LSTM (Stage 3) is then fitted to residual_after_prophet = residual_arima – Prophet_prediction. In other words, the Triple Hybrid extends the ARIMA–Prophet cascade (Triple Hybrid section) by adding one LSTM stage at the end rather than constituting an architecture independent of the other three variants.

A direct consequence of this structure is that the LSTM in the Triple Hybrid inherits two layers of residual modeling: first from ARIMA and then from Prophet operating on the ARIMA residual. Therefore, the remaining signal for the LSTM to learn should ideally approach noise if the two preceding stages have adequately captured the systematic structure. However, because the ARIMA residual itself is already poorly structured (ARIMA Results section), and the Prophet stage operating on that residual does not completely eliminate the remaining structure (Triple Hybrid section), the LSTM in the Triple Hybrid may inherit the imperfections of both preceding stages simultaneously. This is precisely what the results below test.

Table 7

Forecasting Performance of the ARIMA–Prophet–LSTM Triple Hybrid Model

Specialty	MAE	RMSE	MAPE
Pediatrics	14.2	18.08	141.97
Orthopedic and Traumatology Surgery	18.98	23.68	32.57
Internal Medicine	41.3	52.97	248.94
Cardiology	20.82	25.87	34
Neurology	15.54	19.45	39.74

The Triple Hybrid, as the core architecture proposed by this study, performs close to the ARIMA–Prophet model but does not outperform Prophet–LSTM in any of the five specialties. Adding an ARIMA stage at the beginning of the cascade provides no apparent advantage over starting directly from Prophet; instead, the ARIMA stage produces a residual that may be less informative for the subsequent stages, creating an additional modeling burden without a commensurate benefit, consistent with the structural explanation above.

Performance Comparison of All Models

The results do not support the study's main hypothesis that the hybrid models, particularly the Triple ARIMA–Prophet–LSTM model, would outperform the baseline models.

Table 8

Performance Comparison of Forecasting Models Across Outpatient Specialties

Specialty	ARIMA	Prophet	LSTM	ARIMA-Prophet	ARIMA-LSTM	Prophet-LSTM	ARIMA-Prophet-LSTM	Best Model
Internal Medicine	63.4	38.09	53.55	51.17	57.21	39.04	52.97	Prophet
Orthopedic and	32.03	19.88	31.46	23.63	29.6	19.97	23.68	Prophet

Specialty	ARIMA	Prophet	LSTM	ARIMA-Prophet	ARIMA-LSTM	Prophet-LSTM	ARIMA-Prophet-LSTM	Best Model
Traumatology								
Surgery								
Cardiology	32.52	17.23	23.77	25.81	31.93	17.36	25.87	Prophet
Pediatrics	30.13	13.39	21.73	17.12	27.1	15.25	18.08	Prophet
Neurology	25.41	15	26.72	19.77	22.2	14.49	19.45	Prophet-LSTM

Note: The RMSE values in the overall comparison table were computed using a reduced 3-fold TSCV scheme (5 specialties $\times$ 3 folds = 15 matched blocks) to ensure a fair paired comparison across all seven models under identical fold splits. The individual model tables above used 10 folds per specialty. This methodological difference accounts for the numerical discrepancies between tables (e.g., Internal Medicine Prophet RMSE: 31.45 with 10 folds vs. 38.09 with 3 folds).

Key Findings: (1) Standalone Prophet had the lowest RMSE in four of the five specialties, outperforming all hybrid variants, including the Triple Hybrid that constitutes the core contribution of this study. (2) The Triple Hybrid had the lowest RMSE in none of the specialties and consistently underperformed relative to Prophet-LSTM. And (3) Prophet-LSTM had a slightly lower RMSE than standalone Prophet in one specialty (Neurology: 14.49 vs. 15.00, a 3.4% difference), whereas in the other four specialties it had a higher RMSE than standalone Prophet.

The additive residual-cascade design assumes that each stage isolates and passes the remaining signal to the next stage. However, when the first stage (ARIMA) produces residuals that are not white noise (ARIMA Results: residuals are not white noise in any specialty), the subsequent stages inherit systematic structure rather than an ideally isolated nonlinear signal. Computational simplifications (reduced epochs and folds, without per-fold retuning) may also have limited the optimal performance of LSTM, both in standalone mode and within the hybrid models. Iterative rollout prediction in LSTM can accumulate error across the 30-day horizon, a weakness that may have a greater impact on hybrid configurations in which LSTM serves as the final stage.

Statistical Significance Test Results

Normality Test of Metric Data

The Shapiro-Wilk test was applied to the metric data (MAE, RMSE, and MAPE) for each model, with the unit of observation being 15 blocks (5 specialties $\times$ 3 TSCV folds), constituting a matched-pairs design because each model was evaluated on identical specialty-fold blocks.

Only 1 of 21 combinations (metric $\times$ model) was normally distributed (LSTM on MAE, $p = 0.518$); all other 20 combinations rejected the null hypothesis of normality ($p < 0.05$). This is consistent with the skewed nature typically observed in forecasting metric data, largely due to several folds with extreme values (e.g., MAPE spiking on days when the actual value is near zero, as repeatedly observed in the ARIMA Results section—Section 4.2.3). With most groups being nonnormal and a small sample size ($N = 15$ per group), a nonparametric test (Friedman test) was chosen over parametric repeated-measures ANOVA.

Friedman Test for Multi-Model Comparison

The Friedman test was applied to all three metrics, with H_0 : all seven models (ARIMA, Prophet, LSTM, ARIMA-Prophet, ARIMA-LSTM, Prophet-LSTM, and ARIMA-Prophet-LSTM) have equivalent performance.

Table 9

Results of the Friedman Test for Forecasting Model Comparison

Metric	N blocks	K models	Chi2 statistic	p value	Significant (5%)
MAE	15	7	65.4571	0	Yes
RMSE	15	7	64.6571	0	Yes
MAPE	15	7	45.0571	0	Yes

The null hypothesis was rejected for all three metrics ($p < 0.001$), indicating statistically significant differences in performance among the seven models. This provides a basis for post hoc

testing to identify which pairs of models differ significantly.

Post-hoc Nemenyi Test and Pairwise Comparison

The Nemenyi test was applied to RMSE (the primary metric) for all 21 model pairs. The mean ranks (1 = best) provide a summary of the ordering observed in the Model Comparison section:



Figure 3. Pairwise Post-hoc Nemenyi Test Heatmap (RMSE)

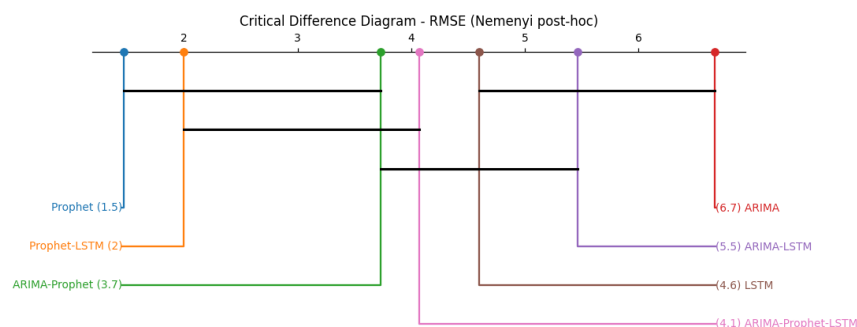


Figure 4. Critical Difference Diagram of Average RMSE Rankings

The following findings statistically reinforce the observations in the Model Comparison section: (1) Prophet vs. Prophet-LSTM: not significantly different ($p = 0.9939$). The two models did not differ significantly in performance, indicating that adding LSTM to the Prophet residual did not produce a statistically significant improvement, even though Prophet-LSTM had a slightly lower RMSE in one specialty (Model Comparison section). (2) Prophet vs. Triple Hybrid: significantly different ($p = 0.0170$). This is the key finding indicating that the Triple Hybrid (ARIMA-Prophet-LSTM), the core architectural contribution of this study, performed significantly differently from standalone Prophet. Because lower RMSE indicates better performance and the mean-rank results favor Prophet, the finding does not support the study's hypothesis that the Triple Hybrid would outperform standalone Prophet. (3) Triple Hybrid vs. ARIMA-Prophet: not significantly different ($p = 0.9996$). Adding LSTM to the ARIMA-Prophet cascade did not produce a statistically significant difference in performance. (4) All models vs. standalone ARIMA: all other models differed significantly from standalone ARIMA ($p < 0.05$ in all pairwise comparisons), consistent with the findings in the ARIMA Results section.

Because Prophet and Prophet-LSTM did not differ significantly in the post hoc comparison and both occupied the upper end of the performance ranking, Prophet-LSTM was selected for interpretation not because it was statistically proven to be the best model, but because it includes

a nonlinear LSTM component that is particularly relevant for interpretation with SHAP. Standalone Prophet is more directly interpretable through its explicit trend, seasonal, and holiday components and therefore has less need for an additional post hoc explanation method such as SHAP.

Model Interpretation with SHAP

Model Selection for Interpretation

SHAP interpretation in this study was applied to the Prophet-LSTM model, with the following justification: (1) The Friedman and post hoc Nemenyi test results (Statistical Test section) show that Prophet-LSTM did not differ significantly from standalone Prophet ($p = 0.9939$; mean ranks 2.00 vs. 1.47) and that both models differed significantly from the other models, including the Triple Hybrid (ARIMA-Prophet-LSTM), which was originally proposed as the core architecture. (2) Standalone Prophet is inherently interpretable because its trend, seasonal, and holiday components are explicitly separated and can be examined directly using `plot_components`, as shown in the Prophet Results section. The added value of SHAP is therefore relatively limited for Prophet. Prophet-LSTM, in contrast, contains an LSTM component that functions as a nonlinear model of the residuals after Prophet, making an interpretation tool such as SHAP useful for examining which features contribute to its residual predictions. (3) Choosing Prophet-LSTM rather than the Triple Hybrid for interpretation preserves narrative coherence: this study has shown that the Triple Hybrid was not supported as a superior model, so shifting the interpretation focus back to the Triple Hybrid would require additional justification.

Global Interpretation Results

SHAP interpretation was applied to the LSTM component of the Prophet-LSTM model (Model Selection section), using SHAP's GradientExplainer, a method suitable for gradient-based neural network models such as LSTM. Because the LSTM input is sequential (14 time steps $\times$ 6 features), SHAP values for each (sample, time step, feature) combination were aggregated by summing their absolute values across all time steps, producing a single representative mean absolute SHAP value per feature per sample. This represents a practical simplification for time-series model interpretation and is noted as a limitation in the Limitations subsection. The analysis was conducted on 150 sequence samples, selected through stratified random sampling (random seed = 42) to ensure proportional representation across months, from a total of 752 sequences available in the initial training set for the Internal Medicine specialty, the main illustrative case and the highest-volume specialty.

Table 10

Global SHAP Feature Importance for the Prophet-LSTM Model

Feature	Mean abs SHAP
dow_cos	0.109528336
dow_sin	0.089894514
month_sin	0.074168112
month_cos	0.068082534
is_libur_nasional	0.019618713
residual_prophet_lag	0.017246322

Day-of-week encoding (dow_sin, dow_cos) dominates the LSTM's contribution, jointly accounting for approximately 48% of the total mean absolute SHAP magnitude, far exceeding the monthly encoding and holiday features. This is a methodologically important finding and warrants attention: the LSTM in Prophet-LSTM still appears to be capturing part of the weekly seasonality, a pattern that should already have been substantially modeled by Prophet in the preceding stage. The residual targeted by LSTM should ideally be largely free of the explicit weekly seasonal structure. This suggests that Prophet has not perfectly captured every nuance of weekly seasonality, leaving residual structure that remains associated with day of week, consistent with standalone Prophet's relatively high MAPE for Internal Medicine (63.3%, Section 4.2.2).

The `residual_prophet_lag` feature, representing the historical value of the residual itself, has the lowest contribution, indicating that the Prophet residual contains relatively limited lagged

autocorrelation that is useful for prediction. Instead, the residual appears to be more strongly influenced by calendar factors such as day of week and month that Prophet has not perfectly captured.

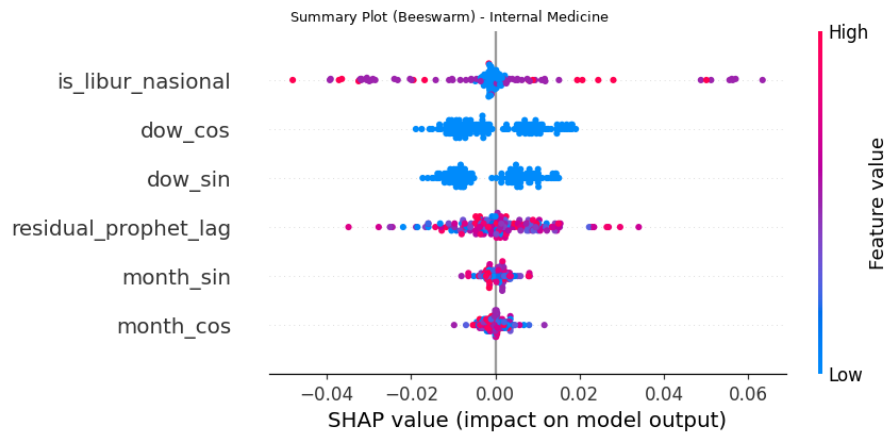


Figure 5. Global SHAP Beeswarm Plot for the Prophet-LSTM Model (Internal Medicine)

The beeswarm plot confirms the direction of influence: high and low values of `dow_cos` and `dow_sin`, representing positions within the weekly cycle, are associated with a wide spread of SHAP values in both directions. The LSTM therefore uses the day's position within the weekly cycle to correct the Prophet residual substantially upward or downward depending on the day, rather than applying a single uniform directional effect.

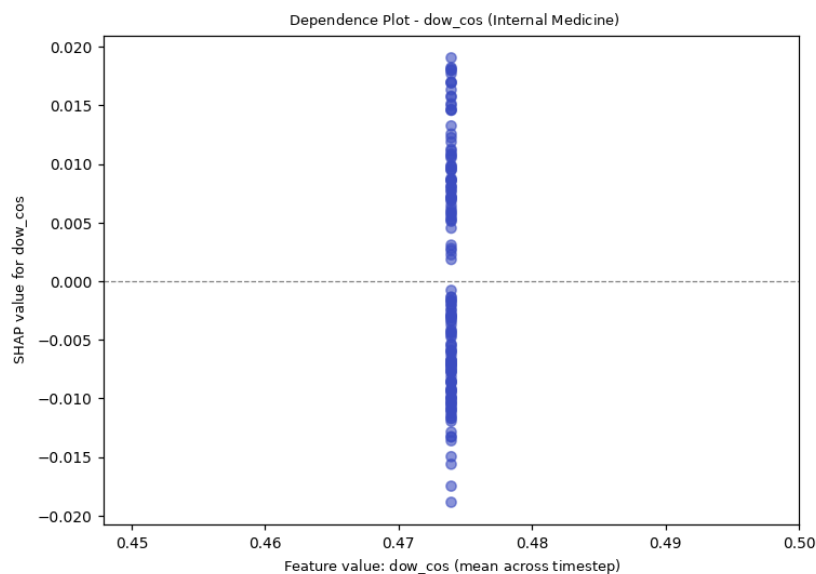


Figure 6. SHAP Dependence Plot for `dow_cos` (Internal Medicine)

The dependence plot shows a nonlinear relationship between the `dow_cos` value and its SHAP contribution, forming a U-shaped or wave-like pattern consistent with the cyclical nature of cosine encoding rather than a simple linear relationship. This indicates that LSTM exploits the nonlinear representation provided by cyclical encoding to distinguish between days of the week, a capability that the linear ARIMA component does not provide (consistent with the findings in the ARIMA Results section).

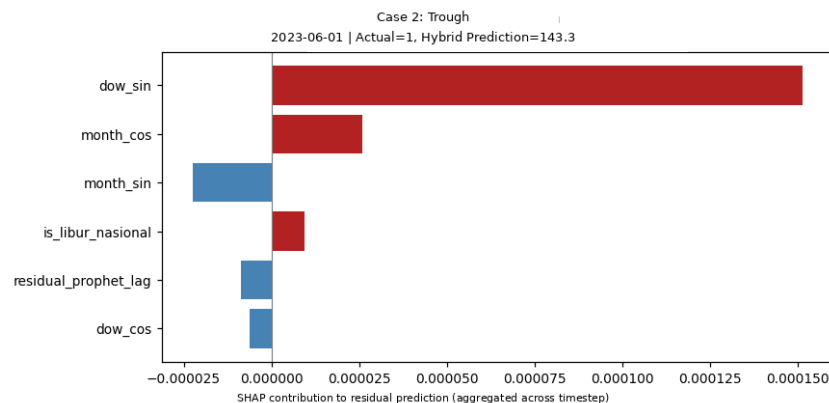
Local Interpretation Results

Five specific cases were interpreted using SHAP on the LSTM component of the Prophet-LSTM model (Internal Medicine specialty, consistent with the Global Interpretation section) to understand model behavior in operationally critical situations.

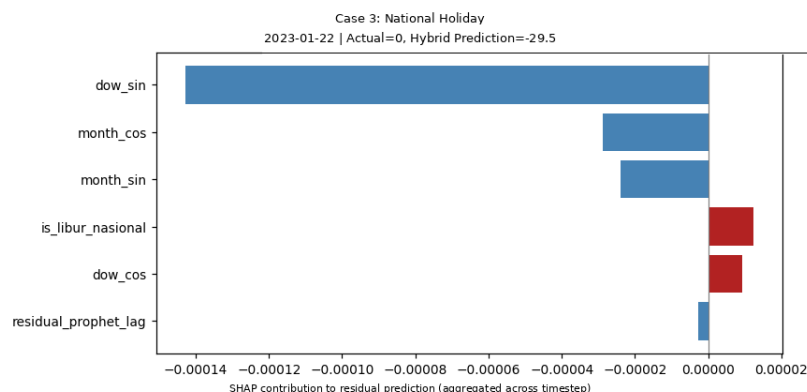
Table 11*Summary of Local SHAP Interpretation Cases*

Case	Date	Actual	Hybrid Prediction	Hybrid Error	Standalone LSTM Prediction	Hybrid vs Standalone Difference
Case 1: Peak	5/20/2024	273	182.1	90.9	155.54	26.56
Case 2: Trough	6/1/2023	1	143.3	-142.3	143.11	0.19
Case 3: National Holiday	1/22/2023	0	-29.46	29.46	22.49	51.94
Case 4: Hybrid vs Standalone LSTM (Max Divergence)	4/23/2023	0	-28.24	28.24	29.99	58.23
Case 5: Highest Prediction Error	6/17/2024	0	160.7	-160.7	154.14	6.56

On May 20, 2024, actual visits reached the highest peak in the data (273), but both the hybrid and standalone LSTM failed to anticipate this surge (hybrid prediction = 182.1; error = +90.9). SHAP shows a dominant contribution from `dow_cos/dow_sin`, representing the day's position in the weekly cycle, which pushed the prediction toward a more typical level rather than an extreme spike. This indicates that the model has limited ability to recognize unexpected anomalies or spikes that are not associated with the usual calendar pattern, potentially because such events may be driven by specific disease outbreaks or external factors not represented among the model's features.

**Figure 7.** Local SHAP Explanation for the Lowest Non-zero Observation

The lowest nonzero trough case (actual = 1) falls exactly on June 1, 2023, the single anomalous row already identified in the Data Description section (Pancasila Day holiday with one recorded data row rather than an empty placeholder, as occurred for the other 51 holidays). The model predicted 143.3, far from the actual value, because structurally the model treats this day as an ordinary working day; the calendar features are insufficient to flag this recording anomaly. This case therefore reflects a data-quality limitation (Data Description section) more than a weakness of the model itself.

**Figure 8.** Local SHAP Explanation for a National Holiday Case

On Chinese New Year 2023 (actual = 0), the model predicted a negative value (−29.5). Although this is mathematically possible within the unconstrained additive forecasting framework, it demonstrates that the hybrid may overcorrect on holidays: Prophet already captures part of the holiday effect, while LSTM adds an additional residual correction in the negative direction. Compared with the standalone LSTM prediction (22.5, which is closer to zero but still far from the actual value), the hybrid produces a larger absolute error (29.5 vs. 22.5) and a negative prediction that is operationally meaningless because outpatient visits cannot be negative. This highlights an important nuance: the accumulation of two correction stages (Prophet + LSTM) can lead to overcorrection in extreme situations such as a full holiday.

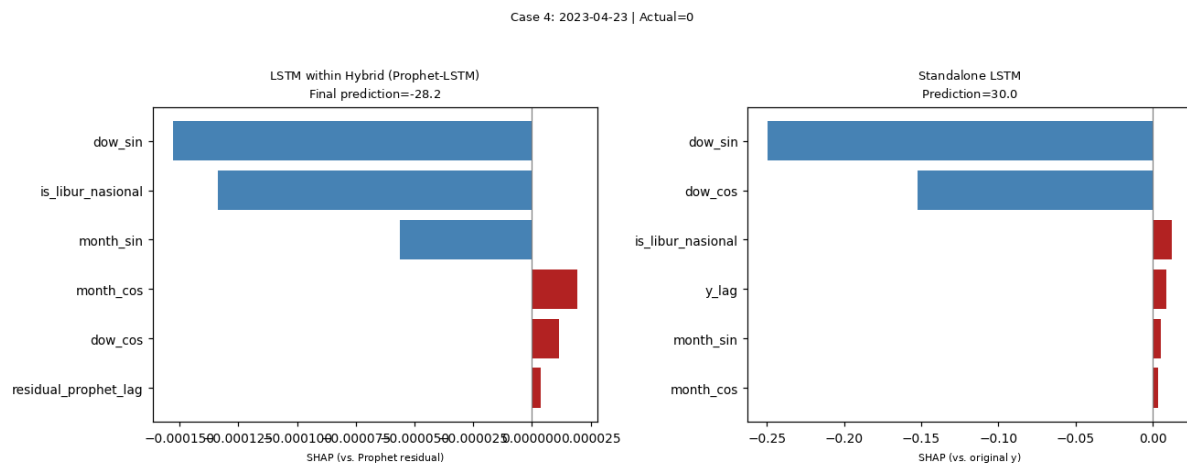


Figure 9. Comparison of SHAP Contributions Between Prophet-LSTM and Standalone LSTM

During Eid al-Fitr 2023, the largest divergence between the hybrid prediction (−28.2) and the standalone prediction (30.0) occurred, with a difference of 58.2, even though both predictions were far from the actual value (0). A side-by-side SHAP comparison shows a similar directional pattern of feature contributions (*dow_cos*/*dow_sin* remain dominant in both models), but the magnitude of the contributions differs because the two models operate on different target scales (Prophet residual vs. raw visits). This suggests that the difference in final predictions is driven more by the Prophet component incorporated into the hybrid, which can take strongly negative values on holidays, than by a fundamental difference in how LSTM processes the calendar features.

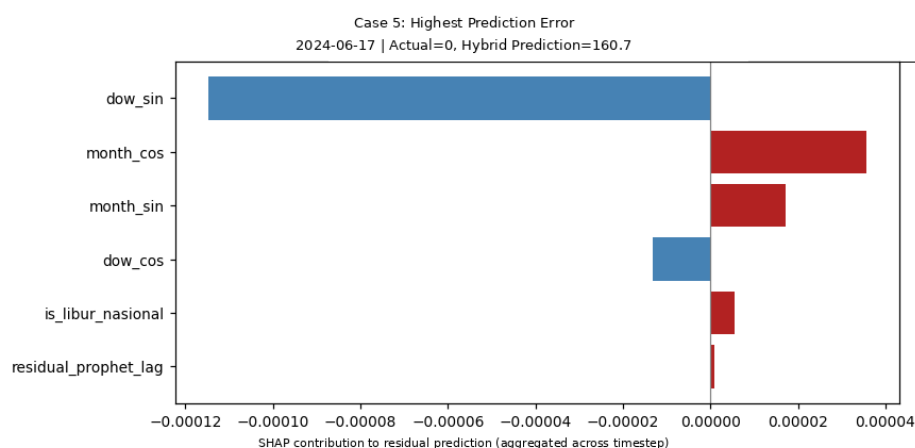


Figure 10. Local SHAP Explanation for the Largest Prediction Error

The highest absolute error (160.7) was found on June 17, 2024 (Eid al-Adha); this date is also a national holiday. An important cross-case finding is that three of the five interpreted cases (Cases 3, 4, and 5) fall on national holidays, and all three produced large errors. This reveals an apparent paradox relative to the Global Interpretation finding that *is_libur_nasional* has relatively low SHAP feature importance (ranked fifth of six features). Although its average contribution is

small, its impact on prediction error can be large on the specific days when the feature is active. A plausible explanation is that holidays are rare events (52 of 1,096 days, approximately 4.7%); consequently, their contribution to the global average SHAP magnitude is diluted by the large majority of ordinary days, whereas their impact on individual predictions can be substantial.

Synthesis of Interpretation Findings by Specialty

SHAP feature importance was calculated for all five specialties: Internal Medicine used the 150-sample analysis described in the Global Interpretation section, whereas the other four specialties were analyzed using stratified random samples of 100 sequences each (random seed = 42) for computational efficiency. To enable cross-specialty comparison, SHAP values were normalized by dividing each feature's mean absolute SHAP value by the sum of all features' mean absolute SHAP values within each specialty, thereby removing the effect of differing residual scales.

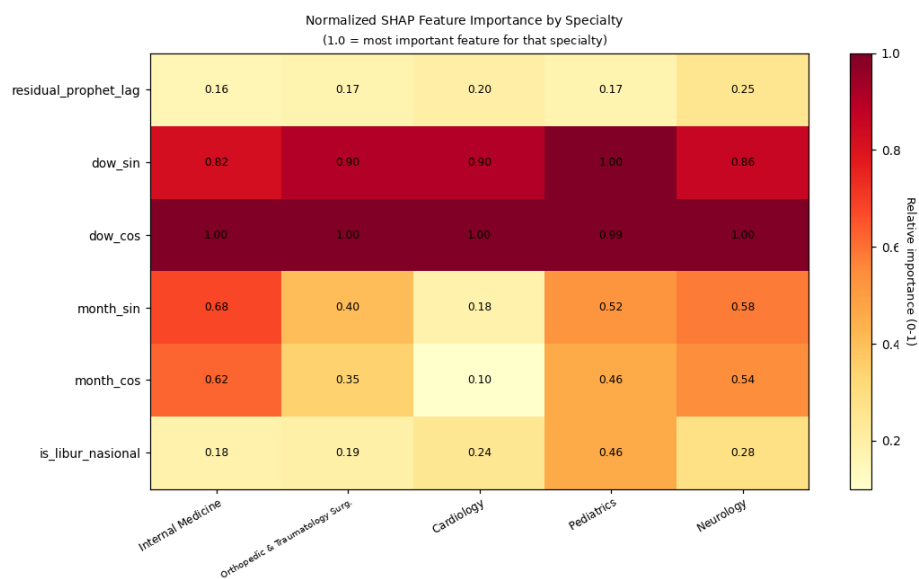


Figure 11. Normalized Global SHAP Feature Importance Across Outpatient Specialties

Day-of-week encoding (dow_cos/dow_sin) dominates as the two most important features across all five specialties without exception, confirming that the finding in the Global Interpretation section is not specific to Internal Medicine but represents a systematic pattern: the LSTM component in the Prophet–LSTM architecture consistently captures part of the weekly seasonal structure across specialties. This indicates that, even with Prophet's built-in weekly_seasonality, Prophet has not perfectly captured every nuance of the weekly pattern in this outpatient data.

Pediatrics is the only specialty in which dow_sin slightly exceeds dow_cos (0.0190 vs. 0.0189), although the values are nearly tied. This is consistent with the finding in the Data Description section that Pediatrics has the highest coefficient of variation (CV) among the five specialties, indicating that its weekly pattern is the least regular or least aligned with the standard symmetric sinusoidal cycle represented by the cyclical encoding.

Internal Medicine (highest volume) shows the highest mean absolute SHAP value (0.1095) for its top feature, decreasing sequentially to Pediatrics (0.0190). This is mathematically plausible because a larger residual scale in high-volume specialties produces a larger absolute SHAP scale as well. However, the absolute SHAP magnitudes should not be interpreted directly as evidence that high-volume specialties have stronger or more predictable weekly patterns. The relative, normalized SHAP values are more appropriate for cross-specialty comparisons. The finding is nevertheless consistent with the CV results in the ARIMA Results section (Internal Medicine CV = 0.44, the lowest; Pediatrics CV = 0.56, the highest among the five specialties).

is_libur_nasional and residual_prophet_lag consistently occupy the bottom two ranks across all five specialties, reinforcing the pattern identified in the Local Interpretation section: low global importance can coexist with large prediction errors on individual holidays. This suggests

that special holiday handling may be relevant across specialties rather than only for Internal Medicine.

Comprehensive Discussion

Performance Evaluation of the Triple Hybrid Model

This subsection title is deliberately retained in line with the study's original structure, but its content reflects the actual findings of this study: the hypothesis that the Triple Hybrid (ARIMA–Prophet–LSTM) would consistently outperform simpler models was not supported. The Triple Hybrid model performed significantly differently from the standalone Prophet model ($p = 0.017$), with the mean-rank results favoring Prophet, and showed no statistically significant difference compared with the simpler ARIMA–Prophet model ($p = 0.9996$). These findings indicate that increasing model complexity does not necessarily translate into higher forecasting accuracy, particularly when the underlying time series is dominated by stable seasonal and trend components rather than highly nonlinear dynamics. Similar observations have been reported in recent healthcare forecasting studies, where model performance depends more on data characteristics than on algorithmic complexity (Fattouh et al., 2026; Wei et al., 2022).

Previous studies have demonstrated the potential advantages of hybrid forecasting approaches. For example, Saini and Sharma (2024) reported that a Prophet–LSTM hybrid improved forecasting accuracy by combining Prophet's ability to model trend and seasonality with LSTM's capability to capture nonlinear residual patterns. Likewise, Mahmud et al. (2025) found that a hybrid ARIMA–LSTM model achieved lower forecasting errors than individual statistical models for COVID-19 prediction. These findings are partially consistent with the present study because all hybrid variants produced better performance than standalone ARIMA. However, unlike previous studies, the current research additionally compared the hybrid approaches with standalone Prophet, which had the lowest RMSE in four of the five specialties. This suggests that when temporal patterns are predominantly seasonal and relatively stable, Prophet alone may sufficiently capture the underlying data structure without requiring additional deep-learning components.

The present findings are also consistent with recent evidence questioning the assumption that deep-learning models inherently outperform simpler forecasting techniques. Comparative forecasting studies have shown that simpler models frequently match or exceed the predictive performance of complex deep-learning approaches Makridakis et al. (2018) and Makridakis et al. (2023), and Prophet can match or exceed the predictive performance of LSTM on datasets characterized by strong seasonality and long-term trends while requiring substantially fewer computational resources (Saini & Sharma, 2024). Similarly, Fattouh et al. (2026) demonstrated that Prophet achieved highly competitive forecasting accuracy for hospital bed occupancy while offering strong interpretability, lower computational complexity, and easier integration into routine hospital management. These characteristics make Prophet potentially suitable for operational healthcare forecasting, where transparency and ease of implementation can be important alongside predictive accuracy.

Another explanation for the observed findings relates to the characteristics of outpatient-visit data. Weekly cycles, monthly trends, and recurring calendar effects dominate outpatient demand, while abrupt nonlinear fluctuations occur relatively infrequently. Under these conditions, Prophet can model much of the meaningful temporal variation through its additive representation of trend, seasonality, and holiday effects. Introducing additional nonlinear learners such as LSTM may therefore contribute limited additional predictive information while increasing model complexity and the risk of overfitting. Similar conclusions have been reported by Wei et al. (2022), who emphasized that healthcare demand forecasting should prioritize models that appropriately match the statistical characteristics of the data rather than assuming that more sophisticated algorithms will always provide better performance.

Implications of Findings for Hospital Management

The findings have several practical implications for hospital operations and healthcare resource planning. First, based on the present results, Prophet or Prophet–LSTM could be considered for routine implementation rather than the Triple Hybrid model. Prophet achieved the lowest RMSE in four of the five specialties and provides a more directly interpretable model

structure, while Prophet-LSTM showed a similar statistical performance profile. The Triple Hybrid, by contrast, did not demonstrate a corresponding performance advantage despite its greater architectural complexity. This finding is consistent with recent hospital forecasting research emphasizing that simpler forecasting models can be preferable because they are easier to deploy, maintain, validate, and communicate to healthcare practitioners (Fattouh et al., 2026).

Second, national holidays should be managed using explicit operational rules rather than relying solely on statistical or machine-learning predictions. The SHAP analysis revealed that holiday variables contributed relatively little to global feature importance despite being associated with some of the largest individual forecasting errors. This pattern is likely related to the low frequency of holiday observations, which limits their influence on aggregate feature-importance measures. Similar challenges have been identified in healthcare forecasting literature, where rare events can contribute disproportionately to forecasting uncertainty (Wei et al., 2022). Therefore, hospitals may improve forecasting reliability by implementing a separate operational adjustment or rule-based layer for officially designated public holidays rather than relying exclusively on the learned holiday effect.

Third, hospital resource planning should be performed separately for each medical specialty rather than applying a uniform forecasting strategy across all outpatient departments. The observed differences in seasonal patterns, coefficients of variation, and forecasting accuracy among specialties indicate that each specialty possesses distinct temporal dynamics requiring individualized planning. Recent studies on outpatient forecasting similarly emphasize that specialty-specific prediction models can provide more effective support for workforce scheduling, clinic capacity planning, and resource allocation than aggregated hospital-level forecasts (Hariyanto, 2022; Wang et al., 2023).

Finally, hospitals should complement calendar-based forecasting models with an early-warning monitoring system capable of detecting unexpected surges in patient demand. Case analyses demonstrated that forecasting errors were largest during sudden increases in outpatient visits that could not be explained by routine calendar variables. Such events may result from infectious disease outbreaks, environmental conditions, policy changes, or other external factors beyond historical temporal patterns. Because models such as Prophet, ARIMA, and LSTM primarily rely on historical observations, they cannot reliably anticipate unprecedented events without additional external information. Recent healthcare forecasting studies therefore recommend integrating forecasting models with real-time surveillance systems and operational monitoring dashboards to improve responsiveness during unexpected demand fluctuations (Di Martino & Delmastro, 2023; Zhang et al., 2023).

Overall, the present findings contribute to the growing evidence that forecasting performance in healthcare depends not only on algorithmic sophistication but also on selecting models that appropriately reflect the temporal characteristics of the underlying data. For outpatient services characterized by regular weekly seasonality and relatively stable long-term trends, Prophet provides an effective balance between predictive accuracy, computational efficiency, and interpretability in this study. Furthermore, combining forecasting with Explainable AI techniques such as SHAP can help hospital managers understand the factors contributing to model predictions, thereby supporting evidence-based decision-making for staffing, facility utilization, and healthcare resource management (Band et al., 2023; Jung et al., 2023).

CONCLUSION

This study evaluated seven forecasting configurations ARIMA, Prophet, LSTM, ARIMA-Prophet, ARIMA-LSTM, Prophet-LSTM, and ARIMA-Prophet-LSTM across five outpatient specialties using time-series cross-validation and Friedman-Nemenyi statistical testing. The results showed that the superiority of the ARIMA-Prophet-LSTM triple hybrid was not confirmed. Prophet achieved the best forecasting performance in four of the five specialties, while Prophet-LSTM performed best only in Neurology, with a 3.4% improvement in RMSE over Prophet. The triple hybrid was significantly inferior to standalone Prophet ($p = 0.017$) and showed no significant difference from ARIMA-Prophet ($p = 0.9996$), whereas Prophet-LSTM was statistically equivalent to Prophet ($p = 0.9939$). These findings demonstrate that increasing model complexity does not necessarily improve forecasting accuracy, particularly for outpatient data characterized by strong and recurring temporal patterns.

SHAP interpretation of the Prophet-LSTM model consistently identified cyclical day-of-week features as the most influential predictors across the five specialties, accounting for approximately 48% of the total SHAP contribution in Internal Medicine. Although the national-holiday indicator showed relatively low global feature importance, several of the largest prediction errors occurred on national holidays, indicating that rare calendar events may have substantial effects in specific forecasting cases. Overall, the findings suggest that Prophet provides the strongest overall balance between forecasting performance and interpretability among the evaluated models, while Prophet-LSTM remains a competitive hybrid alternative when nonlinear residual patterns require further interpretation. The study therefore provides empirical evidence for selecting forecasting models according to the temporal characteristics of specialty-level outpatient demand rather than relying solely on greater model complexity.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Bina Nusantara for providing academic support and research facilities during the completion of this study. Appreciation is also extended to the hospital management and all staff who supported the data collection process for this research.

AUTHOR CONTRIBUTION STATEMENT

Vlandimir Welem contributed to conceptualization, data collection, data analysis, and writing the original manuscript. Sani Muhamad Isa contributed to supervision, validation, methodology refinement, and manuscript review. Both authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

REFERENCES

- Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40, 101286. <https://doi.org/10.1016/j.imu.2023.101286>
- Chandekar, A., Akande, O., Makar, C., Grace, K., Abduselam, S., Balogun, O., Nwoke, M., Okeke, C., Gebreyes, M., & Ricks-Oddie, J. (2026). Impact of primary care visit frequency on non-urgent emergency department visits in a large urban medical center. *Journal of Primary Care & Community Health*, 17, 21501319261420570.
- Cousineau, D. (2025). There are no alternative hypotheses in tests of null hypotheses. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1708313>
- Di Martino, F., & Delmastro, F. (2023). Explainable AI for clinical and remote health applications: A survey on tabular and time series data. *Artificial Intelligence Review*, 56(6), 5261–5315. <https://doi.org/10.1007/s10462-022-10304-3>
- Fattouh, M., Lyssenko, L., Heilmeyer, F., Ball, T., & Haverkamp, C. (2026). Forecasting hospital bed occupancy: A time series approach with Prophet. *BMC Medical Informatics and Decision Making*, 26(1), 162. <https://doi.org/10.1186/s12911-026-03542-w>
- Hariyanto, S. (2022). Forecasting the number of outpatient patient visits using ARIMA at Mitra Sehat Clinic: Case study in Sukoharjo, Central Java Indonesia. *Universal Journal of Public Health*, 10(5), 482–491. <https://doi.org/10.13189/ujph.2022.100506>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.
- Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 9(5), e16110. <https://doi.org/10.1016/j.heliyon.2023.e16110>
- Krichen, M., & Mihoub, A. (2025). Long Short-Term Memory Networks: A Comprehensive Survey. In *AI (Switzerland)* (Vol. 6, Number 9). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/ai6090215>
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). 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? *Journal of Econometrics*, 54(1–3), 159–178. <https://doi.org/10.1016/0304->

4076(92)90104-Y

- Liu, J., Li, C., Ouyang, P., Liu, J., & Wu, C. (2023). Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. *Journal of Forecasting*, 42(5). <https://doi.org/10.1002/for.2931>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Mahmud, A., Noor, S. H. N. S. H., Musa, K. I., Hamzah, F. M., Yudin, Z. M., Kamaruddin, N., Madawana, A. M., & Nawawi, M. A. A. (2025). Hybrid ARIMA-LSTM for COVID-19 forecasting: A comparative AI modeling study. *PeerJ Computer Science*, 11, e3195. <https://doi.org/10.7717/peerj-cs.3195>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), e0194889. <https://doi.org/10.1371/journal.pone.0194889>
- Makridakis, S., Spiliotis, E., Assimakopoulos, V., Semenoglou, A. A., Mulder, G., & Nikolopoulos, K. (2023). Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *Journal of the Operational Research Society*, 74(3). <https://doi.org/10.1080/01605682.2022.2118629>
- Saini, K. P., & Sharma, A. (2024). Impact of data preprocessing on Prophet-LSTM hybrid models for time series forecasting. *Educational Administration: Theory and Practice*, 30(5), 14647–14650.
- Shaikh, A. S., Samant, R. M., Patil, K. S., Patil, N. R., & Mirkale, A. R. (2023). Review on Explainable AI by using LIME and SHAP models for healthcare domain. *International Journal of Computer Applications*, 185(45), 18–23. <https://doi.org/10.5120/ijca2023923263>
- Sprangers, O., Wadman, W., Schelter, S., & de Rijke, M. (2024). Hierarchical forecasting at scale. *International Journal of Forecasting*, 40(4). <https://doi.org/10.1016/j.ijforecast.2024.02.006>
- Statistics, C. for D. C. N. C. for H. (2017). Fast stats. *Ambulatory Care Use and Physician Office Visits*.
- Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>
- Wang, Y., Yi, X., Luo, M., Wang, Z., Qin, L., Hu, X., & Wang, K. (2023). Prediction of outpatients with conjunctivitis in Xinjiang based on LSTM and GRU models. *PLOS ONE*, 18(9), e0290541. <https://doi.org/10.1371/journal.pone.0290541>
- Wei, Y., Li, W., Tan, J., Yuan, J., Wu, Z., Li, Y., Mao, Y., & Huang, D. (2022). A method for improving the prediction of outpatient visits for hospital management: Bayesian autoregressive analysis. *Computational and Mathematical Methods in Medicine*, 2022, 4718157. <https://doi.org/10.1155/2022/4718157>
- Zhang, Y., Wang, L., Li, H., & Chen, X. (2023). Time-series forecasting of healthcare utilization incorporating meteorological factors: A machine learning approach. *BMC Health Services Research*, 23, 1189.