Time Series Forecasting with Irregular Intervals in Applied Behavior Analysis

Joanna Kołodziejczyk[1,2][0000−0002−8078−919X]
and Sebastian Limanowski[2]

[1] National Institute of Telecommunications, ul. Szachowa 1, Warsaw, 04-894, Poland
j.kolodziejczyk@il-pib.pl
[2] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, ul. Żołnierska 49, 71-210 Szczecin, Poland

Abstract. Applied Behavior Analysis (ABA) produces session records that are often irregularly spaced in time, which complicates the prediction because the interval between observations is itself variable and potentially informative. This study investigates a one-step-ahead forecast of the number of challenging-behavior episodes recorded during therapeutic sessions documented in the SYSABA information system. A deterministic data processing pipeline is proposed to convert raw session logs into an analytical data set while preserving temporal order, referential integrity, and data quality indicators. The feature set explicitly represents temporal irregularity through inter-session gaps, calendar attributes, and intra-session aggregates rather than imposing an artificial regular time grid. Forecasts are evaluated under two time-aware validation regimes: global splits across pooled patients and per-patient splits within individual histories. The empirical comparison includes SARIMAX, XGBoost, MLP, and LSTM models, together with the mean and persistence baselines. Point forecasts are assessed using MAE and patient-level MASE, model differences are examined with the Diebold–Mariano test, and predictive uncertainty is assessed through conformal prediction intervals. Results show that no single model dominates across all metrics; however, XGBoost provides the most stable overall performance across validation regimes, while LSTM obtains the best global micro-level MAE. The findings indicate that explicitly encoding temporal irregularity improves predictive usefulness and that individualized validation is essential for clinically interpretable uncertainty estimates.

Keywords: Applied Behavior Analysis · Irregularly Sampled Time Series · Forecasting · SYSABA

1 Introduction

Applied Behavior Analysis (ABA) is a therapeutic methodology in which intervention planning and evaluation are informed by systematic observation of behavior over time [10]. In routine practice, therapists document the outcomes

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

of therapeutic sessions at successive observation times, which allows quantitative assessment of change during intervention [1]. ABA is used in programs designed to strengthen adaptive behaviors and reduce challenging behaviors, particularly in services provided to individuals with developmental conditions, including autism spectrum disorder [8].

In many therapeutic programs, the development of adaptive skills is accompanied by efforts to reduce challenging behaviors, including aggression, disruption, and stereotypy. Because progress is monitored through repeated observation, ABA generates longitudinal behavioral records that can be examined quantitatively. These records have traditionally been interpreted through visual inspection and trend analysis, whereas recent developments in statistical learning have enabled formal prediction based on previously observed behavior [12].

The present study uses records of therapeutic sessions collected in ABA centers in Poland through the SYSABA information system [6]. Since 2018, this platform has accumulated structured observational data from multiple service settings. These records constitute a substantial analytical resource, but they also create a methodological challenge: observations are not collected at equal temporal intervals. The time between two consecutive therapeutic sessions can vary due to holidays, illness, scheduling constraints, or organizational factors. In addition, the completeness of behavioral documentation varies across sessions, and some event logs are only partially recorded.

For these reasons, reconstruction of the data into an artificial daily or weekly grid is not straightforward. The absence of a recorded session cannot be interpreted as the absence of challenging behavior, and interpolation may introduce information that was never observed. Furthermore, the interval between two observations may itself contain information relevant to the prediction. Therefore, the study treats temporal irregularity as an informative characteristic of the data rather than suppressing it through regularization.

The objective of this study is to forecast the number of recorded episodes of challenging behavior at the next observed therapeutic session when observations are irregularly spaced in time. To address this problem, the paper defines a reproducible procedure for transforming raw SYSABA records into an analytical data set suitable for forecasting. The proposed framework specifies the unit of observation, preserves referential integrity, and constructs predictors that represent temporal gaps, calendar information, and summaries of recorded events.

The empirical study compares a statistical forecasting model with exogenous regression models (SARIMAX), a tree-based ensemble method (XGBoost), and two neural architectures (MLP and LSTM). Evaluation is performed under two validation designs that preserve temporal order: one based on pooled observations across all individuals and one based on rolling partitions constructed separately for each individual history. Predictive precision is assessed with MAE and MASE [4], differences between competing forecasts are examined with the Diebold–Mariano test [3], and predictive uncertainty is analyzed by means of conformal prediction intervals.

The main contributions of this paper are as follows:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

1. We formalize next-session forecasting for irregular ABA session histories and define the observation unit, target, and temporal-causality constraint.
2. We propose a deterministic data-reconstruction and feature-engineering pipeline that preserves temporal irregularity and data-quality information.
3. We compare statistical, ensemble, and neural forecasting models under two time-aware validation designs corresponding to different deployment scenarios.
4. We evaluate not only point accuracy but also predictive uncertainty through conformal intervals and forecast-comparison testing.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data source and the rules used to define the analytical cohort. Section 4 formalizes the forecasting problem. Section 5 describes the forecasting methods and the validation design. Section 6 reports the results. Section 7 discusses the findings, limitations, and implications.

2 Related Work

In Applied Behavior Analysis, decisions about intervention effects and expectations concerning future behavior are traditionally based on repeated observation displayed in single-case graphs. Visual analysis remains the dominant interpretive procedure, with attention focused on changes in level, trend, variability, overlap, and proximity across phases [7, 9].

Statistical procedures have long been introduced as complements to graph-based interpretation, particularly when trend evaluation requires more explicit quantification. Methods used in single-case analysis include time-series approaches, piecewise regression, nonoverlap measures such as Tau-U, and multilevel models. Recent reviews emphasize that these methods serve different analytical purposes and should be selected based on the data structure and the analysis objective [11, 12, 9].

More recently, machine learning has been introduced into ABA-related decision support, including automatic interpretation of single-case graphs and recommendation or personalization of treatment goals [5, 2, 7]. However, these studies do not directly address one-step-ahead forecasting of challenging-behavior counts from irregularly spaced therapeutic-session histories. The present study, therefore, lies at the intersection of ABA outcome monitoring, irregular clinical time-series modeling, and predictive uncertainty quantification.

3 Data Preparation and Analytical Representation

The study uses anonymized records extracted from the SYSABA information system, which documents therapeutic sessions in ABA centers in Poland. The retained observation window spans 22 May 2019 to 20 December 2024 and includes 21,200 unique sessions from 195 patients.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

The analytical workflow consists of deterministic extraction, standardization, and transformation steps (Fig. 1). The SYSABA registry provides raw records and reference dictionaries. Because the source table is not strictly session-level, cohort characteristics are reported for analytical rows, unique sessions, and patient-specific unwanted-behavior series. The extract contains 32,233 rows, representing 946 distinct teaching programs series, 461 behavior programs, and 322 unwanted-behavior types.

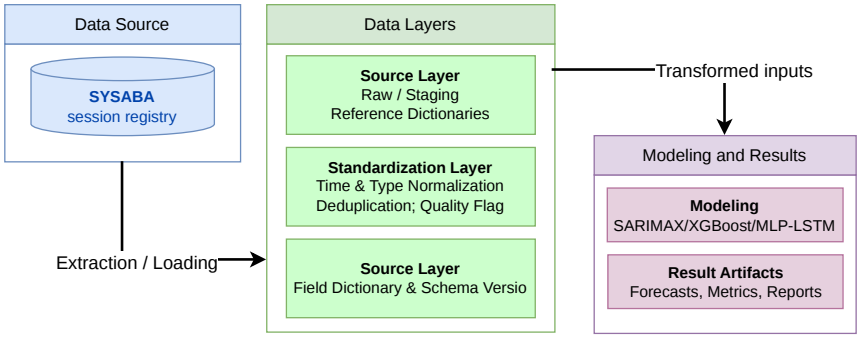

Fig. 1. Overview of the deterministic data-processing and forecasting workflow.

Standardization applies deterministic rules to ensure temporal validity and semantic consistency, including timestamp harmonization, type and range validation, event-log deduplication, removal of personal identifiers, and distinction between true zero counts and missing event documentation. The final analytical stage produces session-level aggregates and engineered predictors.

For forecasting, the cleaned records are converted into ordered one-step-ahead instances. For each patient-specific series, predictors available up to session k are used to predict the number of challenging-behavior episodes at session $k+1$. This representation preserves temporal order and models irregular inter-session spacing explicitly. Of the 946 Behavior programs series, 851 had at least two observations and were therefore eligible for one-step-ahead forecasting.

4 Problem Formulation

This section defines the forecasting problem addressed in the study. It introduces the observational structure of the therapeutic records, the predictor representation constructed from those records, and the one-step ahead forecasting task under the constraint of temporal causality.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

4.1 Observational Structure

Let $i \in \{1, \ldots, N\}$ index patients and let $k \in \{1, \ldots, n_i\}$ index sessions observed for patient i. For each patient, sessions occur at strictly increasing timestamps

$$\mathcal{T}_i = \{t_{i,1}, t_{i,2}, \ldots, t_{i,n_i}\}, \qquad t_{i,k} < t_{i,k+1}. \tag{1}$$

The sequence $\mathcal{T}_i$ is irregular in the sense that the intervals (gaps) between consecutive observations,

$$\Delta t_{i,k} = t_{i,k} - t_{i,k-1}, \qquad k \geq 2, \tag{2}$$

are not assumed to be constant. Let $y_{i,k} \in \{0, 1, 2, \ldots\}$ denote the number of challenging-behavior episodes recorded during session k for patient i. Each patient therefore contributes an irregularly sampled count-valued time series

$$\{(t_{i,k}, y_{i,k})\}_{k=1}^{n_i}.$$

4.2 Predictor Representation

For each patient i and session $k < n_i$, a deterministic predictor vector $X_{i,k}$ is constructed from information available no later than time $t_{i,k}$.

For each observed session, a deterministic predictor vector $X_{i,k}$ is constructed from information available no later than time $t_{i,k}$. Its general form is

$$X_{i,k} = \left(\Delta t_{i,k}, \mathrm{cal}_{i,k}, \mathrm{proc}_{i,k}, \mathrm{context}_{i,k}\right), \tag{3}$$

where:

- $\Delta t_{i,k}$ represents the interval between two consecutive observed sessions;
- $\mathrm{cal}_{i,k}$ contains calendar attributes associated with the session time, such as the day of the week and month;
- $\mathrm{proc}_{i,k}$ contains summaries derived from the event log, including the total number of recorded events, the number of distinct behavioral categories, counts of type-specific entries, total recorded duration of behavior when available, session duration measured from the first to the last logged event, and the counts of quality-related indicators;
- $\mathrm{context}_{i,k}$ contains descriptors of the therapeutic context, including intervention attributes recorded in SYSABA.

4.3 Forecasting Task

Let

$$\mathcal{H}_{i,k} = \{(X_{i,j}, y_{i,j}) : 1 \leq j \leq k\} \tag{4}$$

denote the information available for the individual i after the k-th observed session. The forecasting objective is to estimate the number of challenging-behavior episodes at the next observed session:

$$\hat{y}_{i,k+1} = \mathcal{M}(\mathcal{H}_{i,k}), \tag{5}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

where $\mathcal{M}$ denotes a predictive model of a specified class.

The forecasting rule is subject to a temporal causality constraint: the prediction of $\hat{y}_{i,k+1}$ may depend only on information available at or before time $t_{i,k}$.

Two forecasting settings are considered:

1. a model is estimated from the pooled histories of all individuals $\mathcal{M}_{\text{pool}} : \bigcup_{i=1}^{N} \mathcal{H}_{i,k} \to \hat{y}_{\cdot,k+1}$,
2. a separate model is estimated for each individual $\mathcal{M}_i : \mathcal{H}_{i,k} \to \hat{y}_{i,k+1}$.

4.4 Temporal Irregularity

In contrast to approaches that assume equidistant sampling, the present formulation does not reconstruct the data onto a fixed temporal grid. Instead, temporal irregularity is kept and represented explicitly through the interval variable $\Delta t_{i,k}$ and related predictors. In this way, the original temporal structure of the therapeutic record is preserved.

5 Experimental Protocol

As described in Section 4, the predictive task consists in estimating the number of episodes of challenging behavior in a session based on historical observations.

Explanatory variables include calendar features (day of week, day of month, month), the inter-session gap $\Delta t_{i,k}$, intra-session aggregates (number of events, number of event types, number of entities involved, log span, age).

Numeric preprocessing was performed within each training fold only. First, a predefined subset of nonnegative, right-skewed predictors (e.g., event counts, durations, and gap length) was transformed using $\log(1+x)$. Second, transformed numeric variables were winsorized to the empirical training quantiles $q_{0.01}$ and $q_{0.99}$. Third, numeric variables were centered on the training median and scaled by the training median absolute deviation (MAD).

5.1 Models and Baselines

Three classes of predictive models are considered: (i) a statistical time-series model (SARIMAX, where the regression component is based on exogenous variables (calendar features, inter-session gap, session-level aggregates, and age), while the residual component is examined using ACF and PACF diagnostics together with the Ljung–Box test), (ii) a gradient-boosted tree ensemble (XGBoost), and (iii) neural network models (MLP and LSTM).

All models are evaluated under identical temporal splits and a consistent set of performance metrics, ensuring fair comparison.

Two baseline predictors are included for reference:

- Historical mean: the prediction equals the mean value of the target variable computed on the training set.
- Naive forecast: the prediction equals the most recent observed value.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

5.2 Validation Protocol

Evaluation is performed using time-aware rolling-origin validation. Two evaluation protocols are used:

- Global protocol: Time blocks are defined using global timestamps, with a strict temporal boundary between training and test sets. This protocol reflects real-world forward forecasting and prevents information leakage across folds.
- Per-patient protocol: For each individual, fixed-length test windows are extracted from the end of the available history, while the training set contains only earlier observations.

To ensure robustness, five-fold cross-validation is performed under both protocols, as summarized in Table 1. The use of consistent folds and identical preprocessing of exogenous variables enables a statistically valid comparison of model variants under identical conditions.

Training observations always preceded test observations strictly in time. For each fold, the training subset ended before the first timestamp included in the corresponding test subset.

Table 1. Cross-validation folds for the two evaluation protocols: global forward-chaining and per-patient validation.

Protocol	Fold	Train range	Test range	#Train	#Test
Global	0	baseline	2019-05-27 – 2020-01-30	0	504
Global	1	$\leq$ 2020-01-30	2020-01-31 – 2021-02-19	504	407
Global	2	$\leq$ 2021-02-19	2021-02-22 – 2021-12-15	911	398
Global	3	$\leq$ 2021-12-15	2021-12-15 – 2022-11-25	1309	393
Global	4	$\leq$ 2022-11-25	2022-11-28 – 2024-12-17	1702	357
Per-patient	0	$<$ 2024-10-23	2024-10-23 – 2024-12-17	1803	222
Per-patient	1	$<$ 2024-04-09	2024-04-09 – 2024-10-23	1599	185
Per-patient	2	$<$ 2024-02-23	2024-02-23 – 2024-04-09	1426	175
Per-patient	3	$<$ 2024-01-12	2024-01-12 – 2024-02-23	1256	165
Per-patient	4	$<$ 2023-05-11	2023-05-11 – 2023-05-25	1109	142

5.3 Ablation Variants

To isolate the contribution of temporal irregularity, two feature variants are considered:

1. FULL — complete feature set: calendar features ($\text{cal}_{i,k}$), inter-session gap ($\Delta t_{i,k}$), intra-session aggregates ($\text{proc}_{i,k}$), contextual variables ($\text{context}_{i,k}$), and (for neural models) sequential windows of length W.
2. N-Gap — which excludes the inter-session gap feature $\Delta t_{i,k}$; the model does not explicitly account for irregular time intervals.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

Table 2. Model variants, feature scope, and interpretability mechanisms.

Method	Variant	Feature Scope	Interpretability / Diagnostics
SARIMAX	FULL	$\{\Delta t_{i,k}, \text{cal}_{i,k}, \text{proc}_{i,k}, \text{context}_{i,k}\}$; optional autoregressive window	regression coefficients; residual diagnostics
SARIMAX	N-Gap	$\{\text{cal}_{i,k}, \text{proc}_{i,k}, \text{context}_{i,k}\}$	as above
XGBoost	FULL	$\{\Delta t_{i,k}, \text{cal}_{i,k}, \text{proc}_{i,k}, \text{context}_{i,k}\}$; interaction terms; lag/window features	SHAP (global summary and dependence plots)
XGBoost	N-Gap	$\{\text{cal}_{i,k}, \text{proc}_{i,k}, \text{context}_{i,k}\}$	as above
MLP/LSTM	FULL	$\{\Delta t_{i,k}, \text{cal}_{i,k}, \text{proc}_{i,k}, \text{context}_{i,k}$, sequential windows (W)	learning curves; temporal validation analysis
MLP/LSTM	N-Gap	$\{\text{cal}_{i,k}, \text{proc}_{i,k}, \text{context}_{i,k}\}$; sequential windows (W)	as above

5.4 Evaluation Metrics

The performance of the model is assessed using the cross-validation protocols defined in Table 1. For a given fold, let $\mathcal{S}_{\text{test}}$ denote the set of test pairs (i, k). Its cardinality is denoted by $|\mathcal{S}_{\text{test}}|$. Let $y_{i,k}$ denote the target value observed in session k of patient i, and let $\hat{y}_{i,k}$ (Eq. 5) denote the prediction of the corresponding point.

Point Forecast Evaluation. For each test $(i, k) \in \mathcal{S}_{\text{test}}$, the point prediction error is defined as

$$e_{i,k} = y_{i,k} - \hat{y}_{i,k}. \tag{6}$$

All point forecast metrics, including MSE and MASE, use the same test set and error terms $e_{i,k}$, with their formal definitions provided in Table 3. MAE is treated as the primary point-error metric because it is easily interpretable on the scale of the response and is less sensitive to isolated large deviations than squared-error measures. MASE is also reported because it scales the forecast error relative to a naive benchmark and therefore supports the comparison between patients histories with different levels and variability.

Table 3. Forecast evaluation metrics with consistent (i, k) notation.

Metric	Definition	Interpretation
Mean Absolute Error (MAE)	$\frac{1}{\lvert\mathcal{S}_{\text{test}}\rvert} \sum_{(i,k)\in\mathcal{S}_{\text{test}}} \lvert e_{i,k}\rvert$	Mean absolute deviation
Mean Absolute Scaled Error (MASE)	$\frac{\frac{1}{\lvert\mathcal{S}^{(i)}_{\text{test}}\rvert} \sum_{(i,k)\in\mathcal{S}_{\text{test}}} \lvert e_{i,k}\rvert}{\frac{1}{\lvert\mathcal{S}^{(i)}_{\text{train}}\rvert - 1} \sum_{(i,k)\in\mathcal{S}_{\text{train}}} \lvert y_{i,k} - y_{i,k-1}\rvert}$	Error scaled by naive forecast
Coverage $(1-\alpha)$	$\frac{1}{\lvert\mathcal{S}_{\text{test}}\rvert} \sum_{(i,k)\in\mathcal{S}_{\text{test}}} \mathbf{1}\{L_{i,k} \le y_{i,k} \le U_{i,k}\}$	Empirical interval calibration
IntervalWidth	$\frac{1}{\lvert\mathcal{S}_{\text{test}}\rvert} \sum_{(i,k)\in\mathcal{S}_{\text{test}}} (U_{i,k} - L_{i,k})$	Average prediction interval width

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

Interval Forecast Evaluation. For models that provide prediction intervals at nominal level $1-\alpha$, denote the bounds by $L_{i,k}$ and $U_{i,k}$:

$$[L_{i,k}, U_{i,k}], \qquad \mathbb{P}(L_{i,k} \leq y_{i,k} \leq U_{i,k}) \approx 1-\alpha. \tag{7}$$

The quality of the interval is assessed using two complementary criteria defined formally in Table 3. Coverage measures empirical calibration, i.e., the proportion of observed values falling within the predicted interval, and should be close to the nominal confidence level $1-\alpha$ ($\alpha \in (0,1)$). The interval width quantifies sharpness, reflecting the average width of the interval.

5.5 Aggregation Strategy

The results are reported in two complementary aggregation schemes.

Micro Aggregation. Micro-aggregation treats each pair (i,k) as equally weighted and computes the metric over the union of all test observations across folds within a given validation protocol. It corresponds to a fold-weighted average with weights proportional to test set sizes.

$$m_{\text{micro}} = \frac{\sum_{f=1}^{F} |\mathcal{S}_{\text{test}}^{(f)}|\, m\left(\mathcal{S}_{\text{test}}^{(f)}\right)}{\sum_{f=1}^{F} |\mathcal{S}_{\text{test}}^{(f)}|}, \tag{8}$$

where $m(\cdot)$ is the metric, $\mathcal{S}_{\text{test}}^{(f)}$ are test pairs in fold f and $|\mathcal{S}_{\text{test}}^{(f)}|$ is test fold cardinality. This perspective reflects the expected error for a randomly selected test observation. The same procedure applies to empirical coverage and average interval width.

Macro Aggregation. Macro aggregation assigns equal weight to each patient:

$$m_{\text{macro}} = \text{median}_{i=1,\ldots,N}\ m\left(\mathcal{S}_{\text{test}}^{(i)}\right).$$

This perspective reflects typical model behavior at the patient level and reduces sensitivity to extreme cases.

6 Results

This section presents the selected experimental results, organized according to the adopted evaluation strategy. Different model classes offer distinct analytical perspectives on the predictions and their interpretation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

Table 4. Summary across validation protocols and aggregation schemes.

Model	MAE (micro)		MAE_{med} (macro)		MASE_{med} (macro)	
	Global	Per-patient	Global	Per-patient	Global	Per-patient
XGBoost	4.736	**2.525**	3.025	**1.348**	1.211	**0.649**
LSTM	**4.578**	2.803	**1.767**	1.589	1.286	0.848
MLP	5.363	3.565	1.988	2.252	1.224	0.997
SARIMAX	11.487	5.026	2.871	1.765	1.185	1.002
Baseline: mean	6.600	2.657	2.490	1.177	0.919	0.801
Baseline: persistence	8.387	3.043	3.308	1.314	1.185	0.746

6.1 Models' effectiveness in micro and macro aggregation strategy

Table 4 summarizes the predictive effectiveness in validation protocols and aggregation schemes. Micro-level results report the Mean Absolute Error (MAE) calculated in all test pairs $(i, k) \in \mathcal{S}_{\text{test}}$. In contrast, macro-level results report the median per-patient MAE_{med} and the median Mean Absolute Scaled Error (MASE_{med}).

At the micro level, the LSTM achieves the lowest MAE (4.578), outperforming XGBoost (4.736) and improving over the baseline mean (6.600) and SARIMAX (11.487) in the global protocol. The per-patient protocol determines this order. In per-patient setting, XGBoost achieves the lowest MAE (2.525), followed by the baseline mean (2.657) and the LSTM (2.803). This indicates that while sequence models capture global temporal structure effectively, the tree-based model generalizes better when evaluation is aligned with individual patient histories.

The macro-aggregation provides a complementary perspective. In the global protocol, LSTM produces the lowest MAE_{med} (1.767). In contrast, under the per-patient protocol, XGBoost achieves the lowest MAE_{med} among the learned models (1.348), outperforming LSTM (1.589) and SARIMAX (1.765). This shift suggests that relative advantages depend on the evaluation protocol and that patient-level generalization differs from pooled forecasting.

The scale-normalized results (MASE_{med}) further clarify these differences. In the global protocol, all learned models exhibit MASE_{med} values greater than 1 (e.g., XGBoost: 1.211; LSTM: 1.286), indicating that for a typical patient they do not outperform the naive baseline. According to the per-patient protocol, XGBoost reaches $\text{MASE}_{\text{med}} = 0.649$, substantially below 1, demonstrating a clear improvement over the naive predictor for the median patient. LSTM also improves over the baseline (0.848), though to a lesser extent.

In general, the results reveal two consistent patterns. First, performance rankings depend on the validation protocol, which highlights the importance of aligning the evaluation with the intended deployment scenario. Second, the tree-based model (XGBoost) demonstrates greater stability across protocols and achieves the most significant relative improvement at the patient level, as reflected in both MAE and MASE_{med}.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

6.2 Stratification by Irregularity

To assess how predictive accuracy varies with temporal irregularity, test observations were stratified according to quartiles (Q1–Q4) of the inter-session gap. For each quartile, MAE was computed separately under the global and per-patient validation protocols.

Table 5 reports selected results for representative models. In the global protocol, the distribution between quartiles is highly imbalanced, with the vast majority of observations concentrated in Q1. Results for Q2–Q4 should be interpreted with caution. The per-patient protocol shows substantially more balanced quartile sizes, enabling a more reliable comparison between irregularity levels.

Table 5. MAE by quartiles of inter-session gap ($\Delta t_{i,k}$). Best (lowest) MAE in each quartile and protocol is shown in bold.

	Global				**Per-patient**			
Model	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
XGBoost	4.681	1.864	15.028	1.703	**3.348**	**1.301**	**1.416**	**1.242**
LSTM	**4.102**	**1.188**	17.861	**0.982**	3.800	1.698	1.446	1.839
MLP	4.965	2.481	16.493	1.399	4.415	2.357	2.237	2.378
SARIMAX	11.692	6.489	**12.110**	2.227	6.021	2.920	5.031	5.536
$\lvert\mathcal{S}_{\mathbf{test}}\rvert$	4035	15	6	6	768	342	408	426

Under the global protocol, reliable comparison is effectively limited to Q1, which contains the vast majority of observations. In this quartile, both XGBoost (4.681) and LSTM (4.102) outperform SARIMAX (11.692). The remaining quartiles (Q2–Q4) contain only a small number of observations, resulting in unstable MAE estimates and limiting the strength of conclusions that can be drawn for longer gaps under global pooling.

The per-patient protocol reveals more consistent pattern. For XGBoost, MAE decreases monotonically from Q1 (3.348) to Q4 (1.242), corresponding to an approximate 63% reduction in error. LSTM exhibits a similar, though less regular, decline (3.800 $\rightarrow$ 1.839).

The stratified analysis suggests that model performance may vary across levels of temporal irregularity, but the strength of this conclusion depends on the validation regime. In the global protocol, quartile-level inference is weak because observations are highly concentrated in Q1, leaving Q2–Q4 too sparse for stable comparison. In the per-patient protocol, the quartile counts are more balanced and the pattern is more interpretable, with XGBoost showing decreasing MAE as inter-session gaps increase.

6.3 Diebold–Mariano Test

The Diebold–Mariano analysis focuses on the SARIMAX and XGBoost comparison because both models use the same tabular exogenous-feature representation and therefore provide a direct contrast between a classical forecasting specification and a nonlinear ensemble model. Neural models are excluded from this particular test because their input structure differs due to sequential windowing.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

Table 6. Diebold–Mariano test (Absolute Error). Positive statistics indicate the advantage of XGBoost over SARIMAX.

Protocol	Variant	DM stat.	p-value
Global	FULL	4.962	8.7e-07
Global	N_Gap	4.815	1.8e-06
Per-patient	FULL	2.246	0.0251
Per-patient	N_Gap	3.089	0.0021

As shown in Table 6, XGBoost significantly outperforms SARIMAX in all configurations under both validation protocols. In the global protocol, the most significant difference is observed in the FULL specification (DM $= 4.962$, $p < 10^{-6}$), followed by N_Gap (DM $= 4.815$). In the per-patient protocol, the advantage remains statistically significant, although the effect sizes are smaller (e.g., FULL: DM $= 2.246$, $p = 0.025$).

Importantly, the reduction in DM statistics when moving from FULL to N_Gap suggests that explicitly modeling the inter-session gap contributes to the observed performance.

Overall, DM analysis confirms that XGBoost is superior to SARIMAX, with statistical robustness across validation protocols and feature specifications.

6.4 Prediction Interval Properties

Prediction intervals were constructed using a split-conformal procedure applied within each training fold. A dedicated calibration subset was separated from the model-fitting subset in temporal order to preserve causality. For models producing lower and upper conditional quantiles, conformalized quantile regression was used to adjust interval bounds to the desired nominal coverage level.

Table 7 reports empirical coverage and average interval width for nominal 90% prediction intervals.

Table 7. Empirical coverage and average prediction interval width (nominal level $1 - \alpha = 0.9$).

Model	Global		Per-patient	
	Coverage	Width	Coverage	Width
SARIMAX	0.769	23.99	0.918	18.10
XGBoost	0.661	13.33	0.926	8.95

Under the global validation protocol, both models exhibit substantial undercoverage (SARIMAX: 0.769; XGBoost: 0.661), indicating insufficient calibration. Although SARIMAX achieves higher coverage than XGBoost, it does so at the cost of markedly wider intervals (23.99 vs 13.33). This suggests that pooling heterogeneous time series leads to unstable uncertainty quantification.

In contrast, under the per-patient protocol, both models achieve coverage close to the nominal level (SARIMAX: 0.918; XGBoost: 0.926). Importantly,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

XGBoost attains this calibration with substantially narrower intervals (8.95 vs 18.10), indicating superior sharpness without sacrificing reliability.

These findings demonstrate that individualized modeling not only improves point accuracy but also yields better-calibrated and more informative uncertainty estimates.

6.5 Practical Example

To illustrate the practical application, we present predictions generated by XGBoost in the FULL configuration, previously identified as the strongest-performing specification. Figure 2 shows results for an illustrative patient example under both validation protocols (global and per-patient).

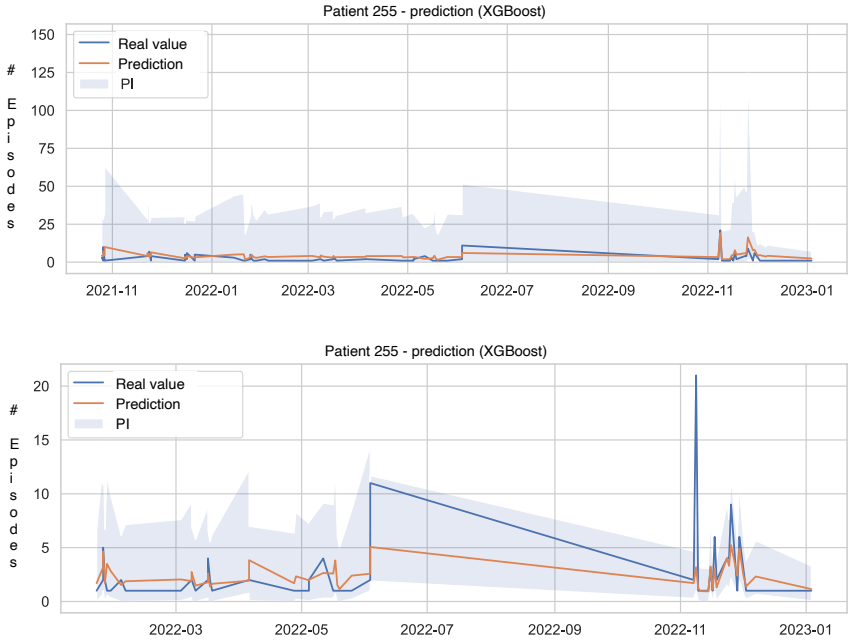

Fig. 2. Forecasting results for patient ID 255 using XGBoost (FULL specification). Top: global validation protocol. Bottom: per-patient validation protocol. The blue line denotes observed values $y_{i,k}$, the orange line represents point forecasts $\hat{y}_{i,k}$, and the shaded region corresponds to 90% prediction intervals.

Under the global protocol, prediction intervals are substantially wider and frequently exceed the empirical range of observations. This reflects the increased uncertainty when the model is trained on pooled data and applied across heterogeneous time series. In contrast, under the per-patient protocol, prediction

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

intervals are noticeably narrower and more closely aligned with the observed variability. This indicates improved calibration and sharper uncertainty quantification when training is restricted to individual patient histories.

7 Discussion and Conclusion

This study investigated time-series forecasting of one-step-ahead challenging behavior under irregular observation intervals in Applied Behavior Analysis (ABA), treating gaps as part of the signal structure. The empirical results indicate that explicit encoding of inter-session gaps can improve predictive usefulness, especially when evaluation is aligned with individual patient histories. At the same time, the results do not support a universal winner across all metrics: LSTM performs best on global micro-level MAE, whereas XGBoost offers the most stable overall trade-off between accuracy, interpretability, and interval sharpness. The XGBoost advantage remained statistically significant in both feature configurations (FULL and N_Gap).

The study has several limitations. First, the data come from a single operational platform and one national service context, which limits external validity. Second, the target variable is a session-level count and does not capture behavioral severity, context specificity, or functional class beyond what is encoded in the available logs. Third, the analysis is observational and predictive rather than causal; improved forecast accuracy does not imply that the identified predictors are treatment drivers. Fourth, the irregularity analysis is partly constrained by an imbalance in quartile support under the global protocol. Fifth, patient-specific models can be unstable when individual histories are short. Sixth, uncertainty calibration was evaluated retrospectively rather than prospectively in a live clinical workflow.

Theoretical implications arise at two levels. First, the paper supports the view that, in event-based behavioral records, observation timing carries information and should be explicitly modeled. Second, it shows that model assessment in such settings must distinguish pooled forecasting performance from patient-level deployment performance, because the preferred model depends on the evaluation perspective.

The practical implications are equally important. A validated session-level forecasting system could support therapist review by identifying elevated next-session risk of episodes, quantifying uncertainty around the predicted escalation, and helping prioritize supervisory attention. However, such use should remain decision-supportive rather than automated, particularly because calibration and error profiles vary between validation regimes.

From a decision-support perspective, modeling irregular time series enables more realistic forecasting in ABA contexts, where session timing is inherently variable.

Future work should extend the analysis to external validation across centers, richer behavioral outcomes, and time-series-specific uncertainty methods tailored to dependent observational data.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34

References

1. Armstrong, K.H., Ogg, J.A., Sundman-Wheat, A.N., St. John Walsh, A.: Evidence-based interventions for children with challenging behavior. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-7807-2
2. Cox, D.J., Sosine, J.: A Data-Driven, Algorithmic Approach to Recommending Hours of ABA for Individuals With ASD. Behavioral Interventions **40**(2), e70014 (Apr 2025). https://doi.org/10.1002/bin.70014, https://onlinelibrary.wiley.com/doi/10.1002/bin.70014
3. Diebold, F.X., Mariano, R.S.: Comparing Predictive Accuracy. Journal of Business & Economic Statistics **13**(3), 253–263 (Jul 1995). https://doi.org/10.1080/07350015.1995.10524599, http://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524599
4. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. International Journal of Forecasting **22**(4), 679–688 (Oct 2006). https://doi.org/10.1016/j.ijforecast.2006.03.001, https://linkinghub.elsevier.com/retrieve/pii/S0169207006000239
5. Kohli, M., Kar, A.K., Bangalore, A., Ap, P.: Machine learning-based ABA treatment recommendation and personalization for autism spectrum disorder: an exploratory study. Brain Informatics **9**(1), 16 (Dec 2022). https://doi.org/10.1186/s40708-022-00164-6, https://braininformatics.springeropen.com/articles/10.1186/s40708-022-00164-6
6. Kołodziejczyk, J.: Uncovering patterns in training skills with aba: Rule extraction from the sysaba database. In: Hernes, M., Wątróbski, J. (eds.) Emerging Challenges in Intelligent Management Information Systems. pp. 3–14. Springer Nature Switzerland, Cham (2024)
7. Lanovaz, M.J., Hranchuk, K.: Machine learning to analyze single-case graphs: A comparison to visual inspection. Journal of Applied Behavior Analysis **54**(4), 1541–1552 (2021). https://doi.org/10.1002/jaba.863
8. Leaf, J.B., Cihon, J.H., Ferguson, J.L., Weinkauf, S.M.: An Introduction to Applied Behavior Analysis, pp. 25–42. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-71210-9_3
9. Manolov, R., Rochat, L.: Analyzing data in single-case experimental designs: Objectives and available software options. Journal of Behavioral and Cognitive Therapy **34**(4), 100511 (2024). https://doi.org/10.1016/j.jbct.2024.100511
10. Poling, A., Fuqua, R.W. (eds.): Research Methods in Applied Behavior Analysis. Springer US, Boston, MA (1986). https://doi.org/10.1007/978-1-4684-8786-2, http://link.springer.com/10.1007/978-1-4684-8786-2
11. Tryon, W.W.: A SIMPLIFIED TIME-SERIES ANALYSIS FOR EVALUATING TREATMENT INTERVENTIONS. Journal of Applied Behavior Analysis **15**(3), 423–429 (Sep 1982). https://doi.org/10.1901/jaba.1982.15-423, https://onlinelibrary.wiley.com/doi/10.1901/jaba.1982.15-423
12. Xu, T.L., De Barbaro, K., Abney, D.H., Cox, R.F.A.: Finding Structure in Time: Visualizing and Analyzing Behavioral Time Series. Frontiers in Psychology **11**, 1457 (Jul 2020). https://doi.org/10.3389/fpsyg.2020.01457, https://www.frontiersin.org/article/10.3389/fpsyg.2020.01457/full

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_34