

Article

Foundation Time-Series Models for Local Forecasting in Adults with Type 1 Diabetes Using Continuous Glucose Monitoring Signals

Roberto Carlos Diaz-Velazco ¹, Alberto Gudiño-Ochoa ¹, Julio Alberto García-Rodríguez ²,
Jorge Ivan Cuevas-Chávez ¹ and Eduardo Ruiz-Velázquez ^{1,*}

¹ División de Tecnologías para la Integración Ciber-Humana, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Guadalajara 44430, Jalisco, Mexico; roberto.diaz9193@alumnos.udg.mx (R.C.D.-V.); alberto.gudino9401@alumnos.udg.mx (A.G.-O.); jorge.cuevas9196@alumnos.udg.mx (J.I.C.-C.)

² Departamento de Eléctrica-Electrónica, Instituto Tecnológico de Ciudad Guzmán, Tecnológico Nacional de México, Ciudad Guzmán 49100, Jalisco, Mexico; julio.gr@cdguzman.tecnm.mx

* Correspondence: eduardo.ruiz@cucei.udg.mx

Abstract

Foundation time-series models have recently shown potential for forecasting complex temporal signals, but their behavior in patient-specific continuous glucose monitoring (CGM) forecasting remains insufficiently understood, particularly when only glucose history is available. This study provides a patient-level benchmark of foundation models for 30 min ahead glucose prediction in adults with type 1 diabetes mellitus (T1DM) under a strictly univariate CGM-only setting. Using the HUPA–UCM dataset from 25 individuals, we evaluated TimeGPT, Chronos, and Sundial against representative statistical, machine learning, and deep learning forecasters, including ARIMA, ETS, gradient-boosting models, recurrent networks, and neural forecasting architectures. Models were assessed using a local walk-forward validation strategy over the final 24 h of CGM data for each patient. Foundation models achieved the strongest global performance, with Sundial obtaining the lowest overall MAE (6.06 mg/dL), while TimeGPT and Chronos remained among the most competitive approaches. However, patient-level analyses showed that this advantage was not uniform: ARIMA remained highly competitive in selected individuals, and no single model consistently dominated across the cohort. These findings suggest that foundation time-series models are promising tools for short-horizon CGM forecasting, but their use should be framed within patient-specific model selection rather than as universal replacements for classical forecasting methods.

Keywords: blood glucose prediction; continuous glucose monitoring; time-series forecasting; foundation models; type 1 diabetes mellitus



Academic Editors: Everardo
Inzunza-González and Fabrizio
Stasolla

Received: 26 May 2026

Revised: 24 June 2026

Accepted: 26 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

For millions of people living with diabetes, daily life is shaped by the persistent risk of acute glycemic events and the long-term threat of organ damage affecting the heart, kidneys, eyes, and nervous system. Diabetes mellitus (DM) remains a major global public health challenge, with substantial consequences for quality of life, morbidity, and health-care systems worldwide. The International Diabetes Federation estimates that, by 2025, approximately 588.7 million people will be living with some form of DM, including nearly 9.5 million individuals with T1DM [1].

The management of DM requires continuous glycemic control to reduce the risk of acute and chronic complications, as emphasized in recent clinical guidelines and reviews [2]. DM is a chronic metabolic disorder defined by insufficient insulin production, ineffective insulin action, or both. Insulin, produced by pancreatic β cells, is essential for glucose regulation. T1DM, specifically, is an autoimmune disorder in which the immune system progressively destroys pancreatic β cells, resulting in insulin deficiency. Consequently, the body loses its capacity to regulate blood glucose levels, leading to recurrent hyperglycemia and hypoglycemia, which are associated with cardiovascular, neurological, renal, and microvascular complications [3].

CGM systems have become integral to modern T1DM management by providing frequent interstitial glucose measurements and supporting therapeutic decision-making [4]. From a signal-processing perspective, CGM data form a dense biomedical time series that reflects nonlinear, patient-specific, and rapidly changing glucose dynamics. However, physiological delays between blood and interstitial glucose, sensor noise, abrupt excursions, and high intra- and interpatient variability limit the effectiveness of relying solely on current measurements. Consequently, predicting future glucose levels has become a critical component of diabetes management, particularly for anticipating hypo- and hyperglycemic events and for supporting artificial pancreas systems and clinical decision-support tools [5–8].

Recent years have seen the development of various glucose prediction methods, ranging from classical statistical and regression models to advanced machine learning (ML) and deep learning (DL) approaches [9–11]. Due to challenges in capturing the nonlinear and nonstationary dynamics of glycemic behavior, alternative modeling strategies for glucose time series have been introduced [12–15].

Among DL-based methods, recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and gated recurrent units (GRUs) have been widely adopted because of their ability to represent temporal dependencies. Prior studies have demonstrated the effectiveness of LSTM and GRU models for both univariate and multivariate formulations, integrating signals from portable devices such as heart rate monitors, which enhance glucose prediction in T1DM patients over horizons from 5 to 60 min and in simulation environments based on physiological models [16–20].

Other works have explored multitask learning and personalized ML strategies to improve adaptation to individual patients [21,22]. Hybrid CNN–LSTM architectures, which combine local pattern extraction with temporal modeling, have also attracted attention because of their reported benefits at longer forecasting horizons [23,24]. In parallel, feature transformation and feature selection techniques have been studied to improve the identification of clinically relevant patterns at the individual level [25,26].

In addition, physiology-based modeling and parameter estimation strategies have been investigated to achieve more realistic representations of glycemic dynamics. Parameter identification approaches based on glucose–insulin models and food intake dynamics aim to personalize physiological representations and support prediction and control frameworks grounded in metabolic mechanisms [27,28]. However, their practical use often depends on the availability and reliability of additional information, such as insulin doses, carbohydrate intake, and patient-specific physiological parameters, which may be incomplete, noisy, or inconsistently recorded in real-world settings.

More recently, foundation time-series models and large language model (LLM)-inspired forecasting approaches have opened a new direction for glucose prediction. Unlike conventional DL models trained for a specific dataset or task, foundation models are pre-trained on large and heterogeneous collections of time series, allowing them to learn general temporal representations that can be transferred to downstream forecasting problems. This

paradigm is particularly relevant for CGM forecasting because glucose signals contain recurring local patterns, abrupt transitions, and patient-specific variability, while the amount of available data for each individual may be limited.

Recent studies have begun to explore this direction. For example, ref. [29] evaluated the use of Time-LLM for glucose forecasting in individuals with T1DM treated with multiple daily insulin injections and insulin pumps, comparing its performance against conventional neural networks, multilayer perceptron (MLP)-based models, and transformer-based architectures. Their results highlighted the value of personalized training and adaptive model selection, especially at short prediction horizons. Similarly, ref. [30] investigated the use of the foundation model TimeGPT for glucose forecasting in pediatric T1DM patients, benchmarking it against models such as ARIMAX, LSTM, and MLP-based architectures, including TiDE and TSMixer, using CGM and insulin data. Their findings showed that TimeGPT achieved performance comparable to or better than several baselines at short horizons without requiring patient-specific training, suggesting that pretrained foundation models may be useful in clinical settings with limited data. However, these studies did not specifically address whether foundation models remain competitive when forecasting is based only on CGM history.

Despite these advances, there is still limited systematic evidence on how fundamentally different forecasting families behave on real-world CGM data when evaluated at the patient level rather than only through aggregate performance summaries. This gap is clinically important because population-level averages can obscure substantial differences in glycemic dynamics across individuals, particularly in the presence of severe hypoglycemia, extreme hyperglycemia, and rapid glucose excursions. Moreover, many existing studies rely on multivariate inputs or family-specific experimental designs, which can complicate direct comparisons across methods. For this reason, a strictly CGM-only setting was adopted to evaluate the intrinsic forecasting ability of each model family under a common and reproducible information setting, while avoiding the confounding influence of additional variables that may differ in availability, quality, or temporal resolution. The present study examines this question by including newer pretrained models such as Sundial and Chronos, applying patient-level Diebold–Mariano tests, and analyzing clinically relevant episodes of severe hypoglycemia and extreme hyperglycemia.

This study examines foundation time-series models for 30 min ahead local forecasting of CGM signals in adults with T1DM using the real-world HUPA–UCM dataset. TimeGPT, Chronos, and Sundial are evaluated alongside representative statistical models, tree-based ML methods, and DL architectures under the same strictly univariate setting. The analysis combines global forecasting performance, patient-level variability, and clinically focused case studies involving severe hypoglycemia and extreme hyperglycemia. This design clarifies what foundation models add beyond strong conventional baselines when only glucose history is available, and whether their advantage is consistent across individual patients rather than limited to aggregate performance. In this sense, this study positions foundation time-series models as promising tools for short-horizon CGM forecasting, but not as universal replacements for patient-specific model selection.

The remainder of this paper is organized as follows. Section 2 describes the HUPA–UCM dataset, forecasting models, and experimental design. Section 3 reports and discusses the evaluation results. Finally, Section 4 concludes the paper and outlines future research directions.

2. Materials and Methods

2.1. Dataset

This study utilized the HUPA-UCM real-world clinical dataset comprising continuous glucose monitoring data from individuals with T1DM [31]. This dataset comprises at least

14 days of real-life data from 25 individuals with T1DM. It includes CGM data acquired every 15 min using the Abbott FreeStyle Libre system (Abbott Diabetes Care, Alameda, CA, USA) and subsequently interpolated to a 5 min resolution, together with records of insulin administration, carbohydrate intake, and physiological variables such as heart rate, physical activity, caloric expenditure, and sleep, collected via insulin devices, mobile applications, and Fitbit wearables (Fitbit LLC, San Francisco, CA, USA). As indicated in Table 1, the clinical characteristics show that 52% of participants were female, with a mean age of 39.23 ± 11.84 years; 56% used continuous subcutaneous insulin infusion (CSII), and 44% followed a multiple daily injection (MDI) therapy regimen.

Table 1. Patients from the HUPA-UCM dataset.

Characteristic	Value
Total number of patients	25
Female	13 (52.0%)
Male	12 (48.0%)
Age (years)	39.23 ± 11.84
Weight (kg)	69.06 ± 14.12
Height (cm)	169.04 ± 10.41
HbA1c (%)	7.37 ± 0.82
Years since diagnosis	17.8 ± 10.5
CSII therapy (%)	56.0
MDI therapy (%)	44.0
Minimum recording duration	≥ 14 days
Sampling frequency	5 min

2.2. Architectures Used

The evaluated architectures are grouped by modeling principle in Table 2. Statistical, ML, and DL models are summarized briefly, whereas the foundation time-series models are described in more detail, their forecasting mechanisms differing substantially in pretraining strategy, input representation, and autoregressive generation.

Table 2. Summary of statistical, machine learning, and deep learning architectures used in this study.

Model Family	Model	Architecture
Statistical	ARIMA [32,33]	Autoregressive integrated moving-average model that combines autoregressive terms, differencing, and moving-average components to represent temporal dependence.
	ETS [34,35]	Exponential smoothing framework that represents time series through combinations of error, trend, and seasonal components updated recursively.
Machine learning	CatBoost [36]	Gradient-boosting decision-tree algorithm that uses ordered boosting to improve stability and reduce overfitting.
	LightGBM [37,38]	Efficient gradient-boosting framework based on decision-tree ensembles for learning nonlinear relationships from structured input features.
	XGBoost [39]	Regularized gradient-boosting decision-tree model that builds trees sequentially to minimize residual errors and improve predictive accuracy.
Deep learning	MLP [40,41]	Feedforward neural network composed of fully connected layers that learn nonlinear mappings between historical inputs and future values.
	LSTM [42–44]	Recurrent neural network with memory cells and gating mechanisms designed to capture temporal dependencies in sequential data.
	GRU [19,45,46]	Recurrent neural network that uses update and reset gates to model temporal dependencies with a simpler structure than LSTM.
	NHITS [47]	Neural forecasting architecture based on hierarchical interpolation blocks for multi-step time-series prediction.
	NBEATSx [48]	Neural basis-expansion architecture that uses stacked blocks to decompose the input signal into backcast and forecast components.
	KAN [49]	Neural architecture that replaces fixed activation functions with learnable univariate functions defined on network connections.

Foundation Models

TimeGPT is a foundation model for time-series forecasting based on a generative pretrained Transformer architecture specialized for temporal data [50,51]. Given an input sequence, TimeGPT embeds the temporal observations, incorporates positional information, and processes the resulting representation through stacked self-attention layers. The core self-attention operation can be written as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q , K , and V denote the query, key, and value matrices, and d_k is the dimensionality of the key vectors. Through this mechanism, TimeGPT learns dependencies among temporal positions and transfers representations learned from large heterogeneous time-series collections to downstream forecasting tasks. In this study, its relevance lies in assessing whether pretrained temporal representations can improve local CGM forecasting using only recent glucose history.

Sundial is a generative foundation model for time-series forecasting that operates directly on continuous-valued signals, avoiding the discrete tokenization step used by language-model-based approaches [52]. The model first divides the input sequence into contiguous temporal patches. If m denotes the patch length, the i -th patch can be expressed as

$$\mathbf{p}_i = [y_{(i-1)m+1}, y_{(i-1)m+2}, \dots, y_{im}], \quad (2)$$

and each patch is projected into an embedding space as

$$\mathbf{e}_i = \mathbf{W}_p \mathbf{p}_i + \mathbf{b}_p. \quad (3)$$

The embedded patches are processed by a decoder-only transformer with causal self-attention, so that each representation depends only on current and previous patches. Sundial then models future trajectories in the continuous domain through a generative TimeFlow objective based on flow matching. In abstract form, the generation process can be described as

$$\frac{d\mathbf{z}_\tau}{d\tau} = v_\theta(\mathbf{z}_\tau, \tau \mid \mathbf{H}), \quad \mathbf{z}_0 \sim p_0, \quad \hat{\mathbf{p}}_{n+1} = \mathbf{z}_1, \quad (4)$$

where v_θ is the learned flow field, $\tau \in [0, 1]$ is the flow time, $\mathbf{H}$ denotes the transformer representation, and $\hat{\mathbf{p}}_{n+1}$ is the generated future patch. This continuous generative formulation is especially relevant for CGM signals because short-term glucose dynamics may include abrupt excursions and local uncertainty.

Chronos is a foundation model for probabilistic time-series forecasting that reformulates forecasting as a language modeling problem [53]. Unlike Sundial, Chronos transforms continuous-valued observations into discrete tokens. Given an observed value y_i , the model first applies scaling,

$$\tilde{y}_i = \frac{y_i}{s}, \quad (5)$$

where s is a series-dependent scale factor. The scaled value is then mapped into a finite vocabulary through a quantization function,

$$q_i = Q(\tilde{y}_i), \quad q_i \in \{1, \dots, V\}, \quad (6)$$

where V is the vocabulary size. Forecasting is then performed as autoregressive token prediction:

$$p_{\theta}(q_{t+1:t+h} | q_{1:t}) = \prod_{j=1}^h p_{\theta}(q_{t+j} | q_{1:t+j-1}). \quad (7)$$

The model is trained using a categorical objective over future tokens, and predicted token sequences are mapped back to the original numerical scale through inverse quantization and rescaling. This formulation allows Chronos to use standard transformer language-modeling machinery while producing probabilistic forecasts for continuous time-series data.

2.3. Walk-Forward Cross-Validation and Model Configurations

The models were evaluated using local walk-forward cross-validation with a patient-specific fixed-size rolling-window scheme. For each patient, the last 24 h of available CGM data were used as the evaluation block and were covered by 48 iterations. At each iteration, the model was evaluated at a 6-step-ahead horizon (30 min), and the next iteration advanced by the same horizon. The first training window ended immediately before the first cutoff of this evaluation block and started at the beginning of the patient's available CGM record. In subsequent iterations, both the start and end of the training window were advanced by 30 min, keeping the number of training observations constant within that patient. This number differed across patients according to the amount of historical CGM data available for each individual. The exact number of training points per patient is reported in Table 3. Overall, most patients provided more than one week of historical CGM data for local training, while the longest historical records corresponded to HUPA0026P, HUPA0027P, and HUPA0028P. This setup preserves temporal causality, avoids information leakage, and reflects a realistic local forecasting scenario. Figure 1 illustrates this procedure. Figure 2 summarizes the overall workflow.

Table 3. Patient-specific training-window size used in the walk-forward cross-validation scheme. Training points are exact; training days are approximate.

Patient	Training Points	Approx. Training Days
HUPA0001P	3810	13
HUPA0002P	2898	10
HUPA0003P	3486	12
HUPA0004P	2898	10
HUPA0005P	3570	12
HUPA0006P	2004	7
HUPA0007P	3570	12
HUPA0009P	3528	12
HUPA0010P	2688	9
HUPA0011P	3552	12
HUPA0014P	3546	12
HUPA0015P	3504	12
HUPA0016P	3552	12
HUPA0017P	3312	12
HUPA0018P	3612	13
HUPA0019P	3426	12
HUPA0020P	2574	9
HUPA0021P	2058	7
HUPA0022P	3738	13
HUPA0023P	3636	13
HUPA0024P	2616	9
HUPA0025P	3720	13
HUPA0026P	40,320	140
HUPA0027P	165,018	573
HUPA0028P	25,614	89

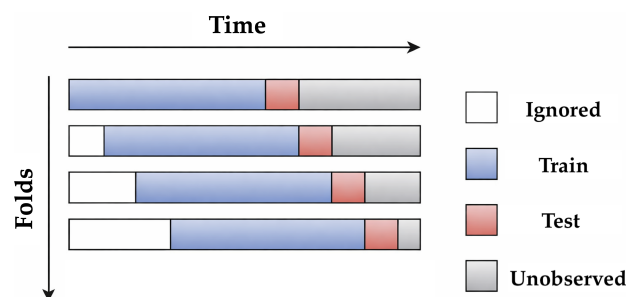


Figure 1. Walk-forward cross-validation with fixed-length rolling windows. In each fold, the training segment keeps the same length and is shifted forward in time, while the test block always corresponds to the immediately following 30 min forecasting horizon.

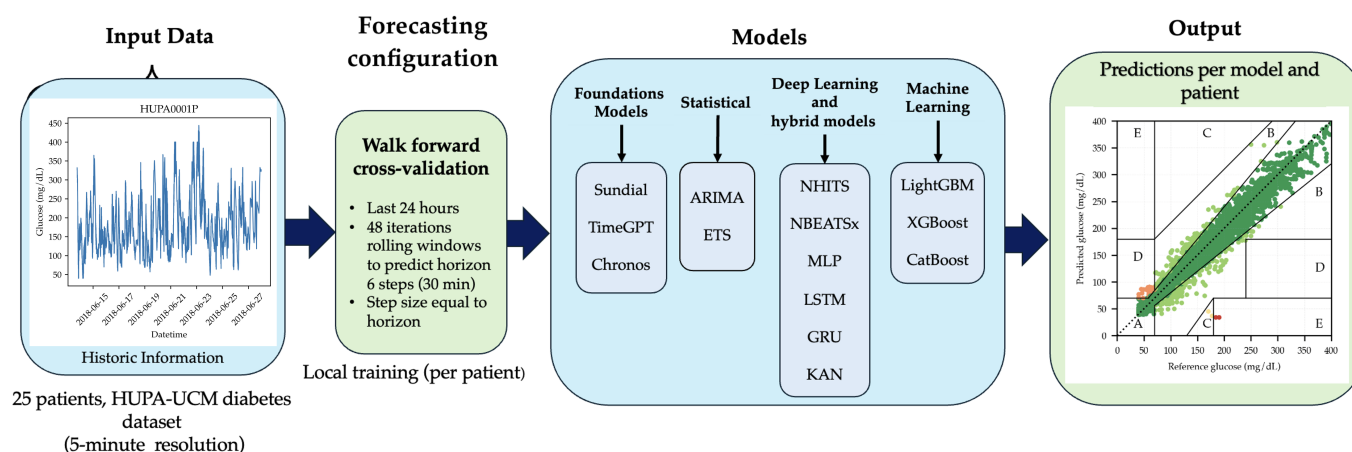


Figure 2. Overall workflow; historic CGM data from the HUPA-UCM dataset were evaluated under a local training strategy using walk-forward cross-validation over the last 24 h of each patient.

Model-specific hyperparameter optimization was not introduced as an additional experimental factor. The forecasting approaches considered in this study differ substantially in their parameterization, training procedures, and degree of configurability, making a common tuning strategy inappropriate across statistical models, tree-based methods, neural architectures, and pretrained foundation models. Extensive model-specific searches would also add a source of heterogeneity, making the results more dependent on unequal optimization effort and model-specific search spaces than on the forecasting paradigm being evaluated. Therefore, representative configurations from standard implementations were adopted to preserve a controlled comparison under the same strictly univariate CGM-only information setting. This choice was particularly relevant given the patient-specific walk-forward design, which required repeated training and evaluation across multiple models, patients, and temporal splits. Table 4 summarizes the main implementations and shared configurations used in the experiments.

For machine learning models, predictors were built exclusively from historical CGM information using lagged glucose values from the previous 24 h, along with rolling means over short and medium temporal windows and an expanding historical mean. For deep learning models, a common training configuration was adopted across architectures to maintain a fair comparison.

Table 4. Main model families and shared configurations used in the study.

Model Family	Models	Implementation	Main Configuration
Foundation	Sundial, TimeGPT, Chronos	timecopilot (0.3.0) (sundial-base-128m); Nixtla (0.7.2) (timegpt-1); Amazon (2.3.0) (chronos-bolt-base)	Only historical CGM values were used as input. Default pretrained versions were evaluated under the same local forecasting setting.
Machine learning	CatBoost, LightGBM, XGBoost	MLForecast (1.0.2)	Only historical CGM values were used. Predictors included 288 lag features (24 h at 5 min resolution), rolling means over 15, 30, 60, and 120 min, and an expanding mean. Shared settings included 200 estimators and a learning rate of 0.05. The Tweedie objective was used when supported.
Deep learning	GRU, LSTM, NHITS, NBEATSx, KAN, MLP	NeuralForecast (3.1.2)	Input size of 240 observations, MAE loss, random seed 42, a maximum of 200 training steps, learning rate of 10^{-3} , early stopping patience of 10 steps, and batch size of 8. Only historical CGM values were used.
Statistical	ARIMA, ETS	StatForecast (2.0.3)	Only historical CGM values were used. ARIMA was implemented as <code>ARIMA(1,1,1)</code> , while ETS was implemented as <code>AutoETS(model='ZZZ')</code> .

3. Results and Discussion

In this study, all experiments were conducted on a Lenovo Legion 5i laptop equipped with an Intel Core i9-14900HX processor, 32 GB of DDR5 RAM, a 1 TB SSD, and an NVIDIA GeForce RTX 4060 GPU. Model training was accelerated using GPU computation. To ensure experimental reproducibility, a fixed random seed of 42 was used in all experiments.

3.1. Error Analysis

Model performance was assessed using complementary error and goodness-of-fit metrics. The main comparative analysis focused on the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and bias factor, allowing for an evaluation of overall fit, average deviation, error dispersion, and systematic directional bias.

The MAE is defined as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (8)$$

where n is the number of observations, y_i is the observed CGM glucose value at time step i , and $\hat{y}_i$ is the corresponding prediction.

The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (9)$$

The coefficient of determination was computed as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (10)$$

where $\bar{y}$ denotes the mean observed CGM glucose value.

To quantify systematic over- or underestimation, the mean bias was first calculated as

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i), \quad (11)$$

The bias factor was then defined as

$$\text{Bias Factor (\%)} = \frac{\text{Bias}}{\bar{y}} \times 100. \quad (12)$$

Negative values indicate systematic underprediction, while positive values indicate systematic overprediction.

Table 5 summarizes the global performance of the evaluated models and shows a consistent separation between the forecasting families. Foundation models achieved the strongest overall results, with Sundial ranking first (MAE = 6.06 mg/dL), followed by TimeGPT, ARIMA, and Chronos, all of which remained below 7.2 mg/dL in MAE. This pattern indicates that pretrained foundation models were particularly effective at capturing short-term CGM dynamics in a strictly local, patient-specific setting.

Table 5. Global performance of the evaluated models for local CGM-based glucose forecasting.

Model	R^2	MAE (mg/dL)	RMSE (mg/dL)	Bias Factor (%)
Sundial	0.9799	6.06	10.38	−0.22
ARIMA	0.9726	6.54	12.14	−0.21
TimeGPT	0.9742	6.91	11.78	0.16
Chronos	0.9746	7.14	11.68	0.07
LightGBM	0.9698	7.61	12.74	−0.19
XGBoost	0.9692	7.64	12.85	−0.93
ETS	0.9671	7.72	13.29	−0.14
CatBoost	0.9653	8.63	13.65	−0.45
NHITS	0.9518	10.46	16.09	−0.57
NBEATSx	0.9436	11.69	17.39	−0.39
MLP	0.9014	16.09	23.01	−1.15
LSTM	0.8873	17.22	24.60	−0.71
KAN	0.8405	17.32	29.26	−0.06
GRU	0.8795	18.28	25.44	0.47

Best values are highlighted in blue and worst values in red. For bias factor, the best value corresponds to the estimate closest to zero in absolute terms.

ARIMA was a notable exception to the complexity–performance trend. Despite its simpler structure, it remained highly competitive, reaching an MAE of 6.54 mg/dL, close to the leading foundation models and comparable to or better than several machine learning approaches. This result suggests that the recent history of CGM alone contains a substantial short-term predictive structure.

Among conventional machine learning models, LightGBM and XGBoost showed nearly identical performance, with MAE values of 7.61 and 7.64 mg/dL, respectively, while ETS also remained competitive. By contrast, deep learning models were less consistent. NHITS and NBEATSx formed an intermediate tier, with MAE values greater than 10 mg/dL, while the MLP, LSTM, GRU and KAN showed substantially larger errors. The GRU produced the highest MAE (18.28 mg/dL), while the KAN yielded the lowest R^2 and the highest RMSE.

The bias factor values remained small across all models, ranging from −1.15% to 0.47%, indicating limited global directional bias. This suggests that the observed differences were driven mainly by forecast accuracy rather than by pronounced systematic over- or underestimation. The strongest models, particularly Sundial, ARIMA, Chronos, and TimeGPT, combined low error with only minimal directional bias.

Figure 3 complements Table 5 by showing the distributions of MAE and RMSE in all patients for each model. The GRU, LSTM, and KAN exhibit wide interquartile ranges and long whiskers in both panels, indicating broader error distributions in the local short-horizon setting. This pattern is consistent with less stable predictive behavior in the cohort. By contrast, Sundial, Chronos, TimeGPT, ARIMA, and LightGBM exhibit lower and more concentrated error distributions. This tighter spread is especially clear for Sundial, whose boxplots remain compact in both MAE and RMSE. NHITS and NBEATSx occupy an intermediate position, with distributions that are more contained than those of

the pure deep learning architectures but still broader and shifted upward relative to the strongest models.

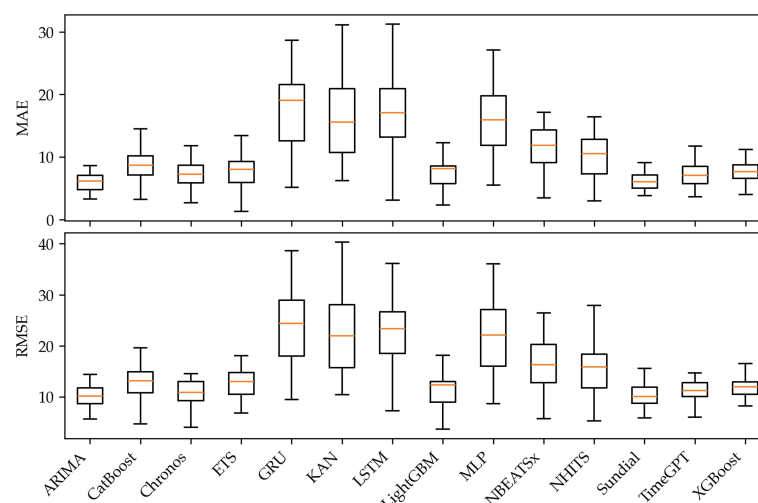


Figure 3. Distributions of MAE and RMSE across all patients for the evaluated models in the local 30 min ahead CGM forecasting task.

To rank the forecasting approaches on a relative scale, we conducted a multiple comparisons with the best (MCB) analysis based on average MAE ranks, where lower values indicate better performance. As shown in Figure 4, the Friedman test was significant ($p < 0.05$), confirming that the models differed systematically in rank performance.

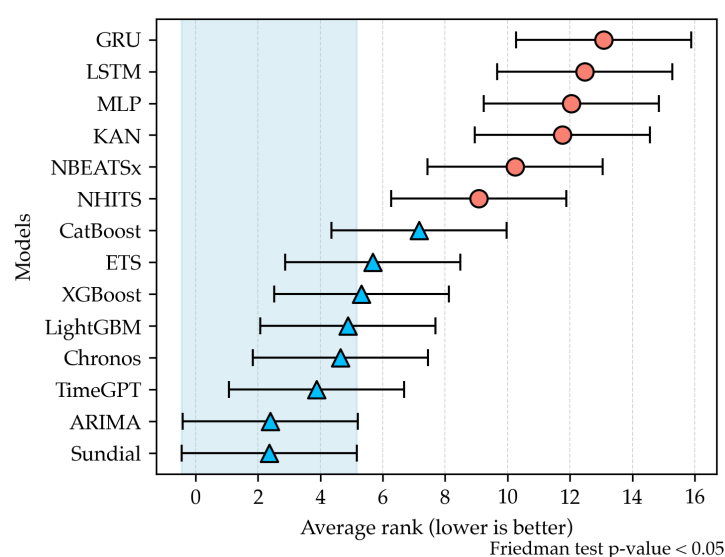


Figure 4. MCB–Nemenyi test using average MAE ranks for the evaluated forecasting models. Lower ranks indicate better predictive performance. The Friedman test p -value is reported in the lower-right corner ($p < 0.05$).

The lowest average ranks were obtained by Sundial and ARIMA, followed by TimeGPT, Chronos, LightGBM, and XGBoost. ETS and CatBoost occupied an intermediate position, whereas NHITS and NBEATSx remained below the strongest statistical, foundation and tree-based approaches. The highest average ranks were concentrated in the KAN, MLP, LSTM, and GRU, confirming their weaker relative performance in the local short-horizon CGM forecasting setting. This rank-based analysis is consistent with the metric-based results and reinforces the same comparative structure: foundation models

occupied the leading positions, ARIMA remained strongly competitive, and several deep learning architectures were displaced toward poorer ranks.

To further assess whether Sundial's superiority was preserved at the individual level, we performed pairwise Diebold–Mariano comparisons against the best-performing model from each remaining forecasting family: ARIMA (statistical), LightGBM (machine learning), NHITS (deep learning), and TimeGPT (foundation model). For each patient-wise comparison, the result was classified according to both the direction of the DM statistic and its statistical significance at $\alpha = 0.05$. Sundial was considered significantly better only when the DM statistic favored Sundial and the corresponding two-sided p -value was below 0.05. Non-significant differences were reported separately. Figure 5 summarizes these patient-wise comparisons.

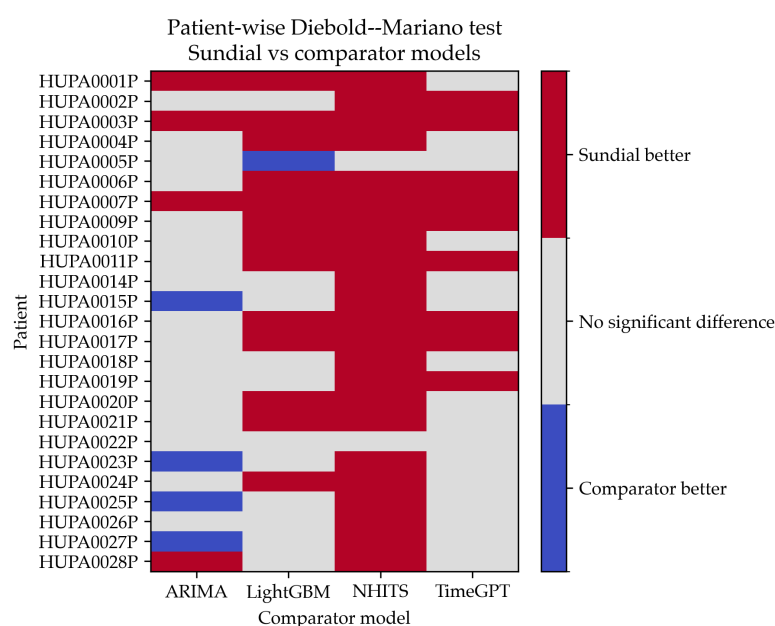


Figure 5. Patient-wise Diebold–Mariano comparisons between Sundial and the best-performing model from each remaining forecasting family across all patients. Red cells indicate that Sundial was significantly better than the comparator, gray cells indicate no statistically significant difference, and blue cells indicate that the comparator was significantly better than Sundial.

The clearest separation was observed against NHITS, for which Sundial was significantly better in 23 of 25 patients (92.0%), with no cases significantly favoring NHITS. Against LightGBM, Sundial was significantly better in 13 patients (52.0%), LightGBM was significantly better in one patient (4.0%), and 11 comparisons were not statistically significant (44.0%). Comparisons with ARIMA and TimeGPT were more balanced. Against ARIMA, Sundial was significantly better in four patients (16.0%), ARIMA was significantly better in four patients (16.0%), and 17 comparisons showed no statistically significant difference (68.0%). Against TimeGPT, Sundial was significantly better in nine patients (36.0%), while the remaining 16 comparisons were not statistically significant (64.0%).

These results indicate that Sundial's patient-wise advantage was strongest relative to the deep learning baseline and moderate relative to the machine learning baseline. However, the comparisons with ARIMA and TimeGPT showed that Sundial did not uniformly dominate the strongest statistical and foundation-model alternatives at the individual level. This reinforces the interpretation that the relative benefit of foundation time-series models is patient-dependent and should not be inferred from aggregate performance alone.

3.2. Clarke Error Grid Analysis

A more clinically oriented view is provided by Table 6, which summarizes the distribution of predictions across Clarke error grid zones for each model. These numerical patterns are consistent with the visual structure of Figure 6. The separation already observed in the numerical metrics is preserved in terms of clinical safety. Sundial, Chronos, LightGBM, ARIMA, XGBoost and TimeGPT achieved the highest proportions of Zone A, all close to 96%, and all exceeded 99% when Zones A and B were considered jointly. Their contributions to the clinically adverse Zones C and E were minimal, indicating a very low frequency of potentially unsafe prediction errors.

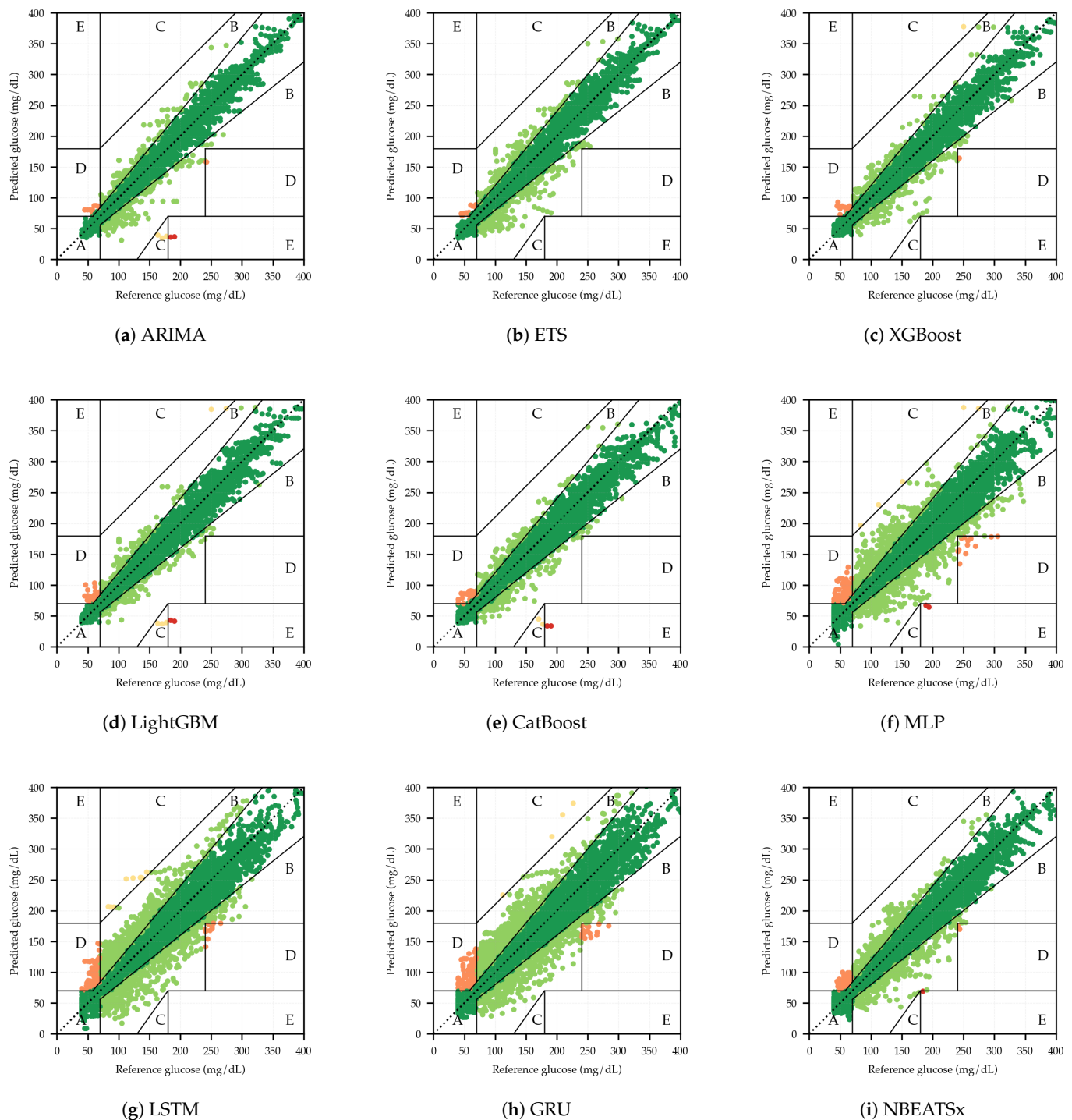


Figure 6. Cont.

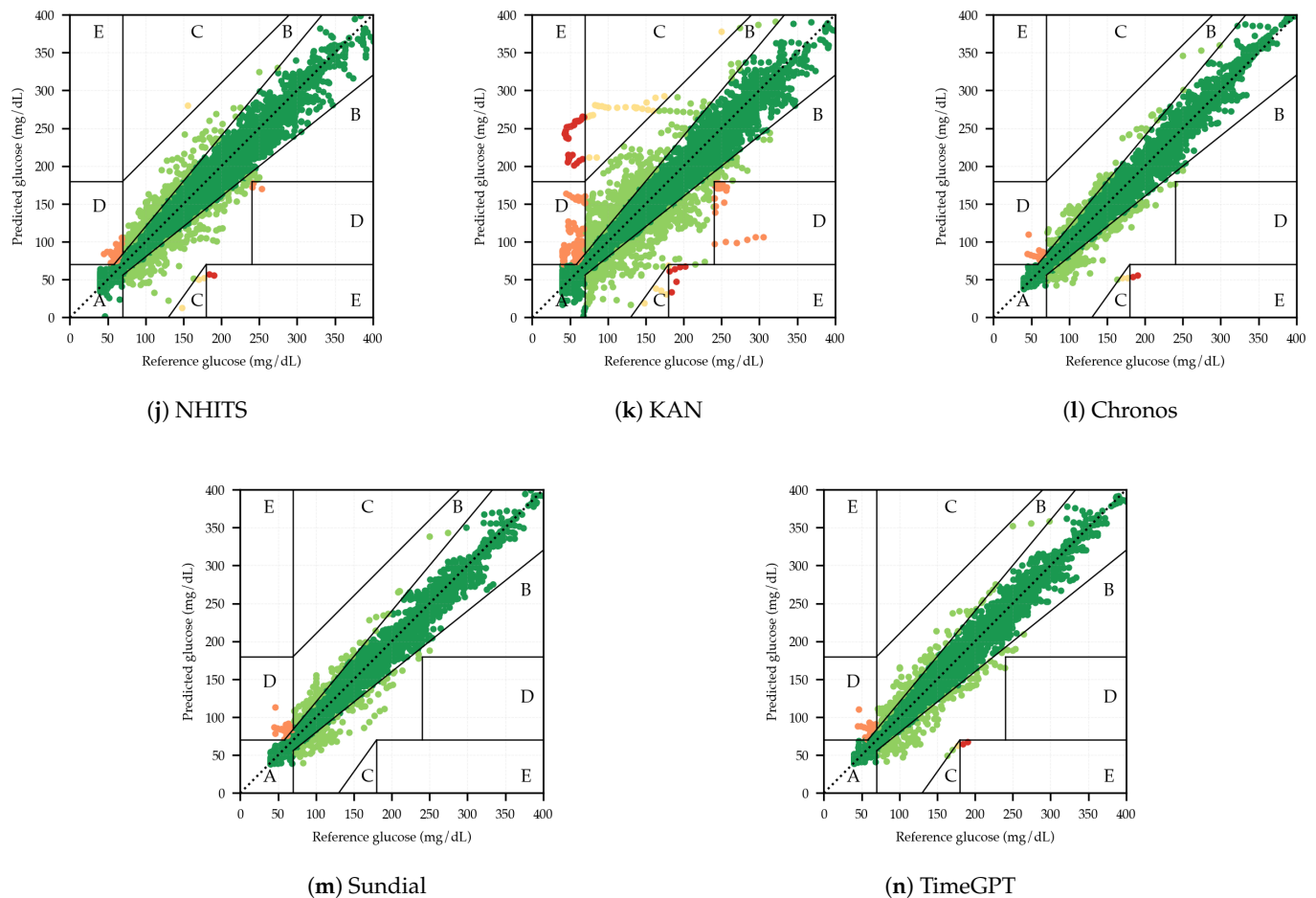


Figure 6. Clarke error grid analysis for the evaluated forecasting models.

Table 6. Percentage of predictions in each Clarke error grid zone for all models.

Model	A (%)	B (%)	C (%)	D (%)	E (%)
Sundial	96.91	2.72	0.00	0.38	0.00
Chronos	96.88	2.81	0.03	0.25	0.03
LightGBM	96.29	3.07	0.07	0.55	0.03
ARIMA	96.26	3.40	0.04	0.27	0.03
XGBoost	96.09	3.57	0.01	0.32	0.00
TimeGPT	95.95	3.54	0.01	0.46	0.03
CatBoost	95.62	3.78	0.03	0.55	0.03
ETS	95.13	4.73	0.00	0.14	0.00
NHITS	92.37	6.59	0.06	0.95	0.03
NBEATSx	90.35	8.36	0.00	1.27	0.01
MLP	84.22	13.04	0.07	2.65	0.03
KAN	84.18	12.10	0.42	2.63	0.67
LSTM	81.62	16.09	0.15	2.14	0.00
GRU	80.05	17.24	0.06	2.66	0.00

In contrast, several deep learning architectures showed a less favorable clinical distribution. The GRU and LSTM yielded the lowest proportions of Zone A (80.05% and 81.62%, respectively), together with larger shifts towards Zone B and higher contributions in Zone D. The MLP and KAN followed a similar pattern, with values in Zone A near 84%; however, the KAN also showed the highest proportions in Zone C (0.42%) and Zone E (0.67%), indicating a higher incidence of clinically discordant predictions.

NHITS and NBEATSx occupied an intermediate position. Most of their predictions remained within the clinically acceptable Zones A and B, but their Zone A coverage was

clearly lower than that of the strongest foundation, statistical and tree-based models, and their Zone D proportions were also higher. This suggests a partial improvement relative to standard deep learning architectures, but not the same level of clinical reliability achieved by the leading approaches.

3.3. Relevant Cases: Severe Hypoglycemia and Extreme Hyperglycemia

To complement the aggregate analyses, we examined three clinically challenging cases in greater detail: HUPA0002P, HUPA0006P, and HUPA0018P. These patients were intentionally selected because, within the last 24 h evaluation window, they concentrated the most clinically relevant glycemic extremes in the cohort. In particular, HUPA0002P and HUPA0018P showed a high burden of severe hypoglycemia, whereas HUPA0006P combined marked hyperglycemic exposure with the highest observed glucose peak. These conditions are especially informative for short-horizon forecasting because clinically consequential errors tend to arise during abrupt or extreme glycemic transitions, even when global average metrics remain favorable. Table 7 summarizes the clinical characteristics of the three selected patients.

Table 7. Clinical and glycemic summary of the three patients selected for local case analysis.

Patient	Gender	Age	Weight	Height	Therapy	Min	Max
HUPA0002P	Male	48	82	186	CSII	40	274
HUPA0006P	Male	22	71	170	CSII	47	438
HUPA0018P	Female	32	57	167	CSII	40	207

HUPA0002P and HUPA0018P were characterized by sustained exposure to clinically relevant hypoglycemia, spending 32% and 34% of the time below 54 mg/dL, respectively, and 46% and 44% below 70 mg/dL. In contrast, HUPA0006P showed a markedly different profile, reaching a maximum CGM value of 438 mg/dL, with 28.7% of readings above 250 mg/dL. These differences are also evident in Figures 7 and 8. The stacked bars show that the last 24 h of HUPA0002P and HUPA0018P were dominated by hypoglycemic states, with severe hypoglycemia accounting for roughly one third of the evaluation window in both cases, while HUPA0006P combined prolonged normoglycemia with a substantial fraction of severe hyperglycemia. The same contrast is visible in the boxplots, confirming that the selected patients represent distinct but equally demanding forecasting conditions.

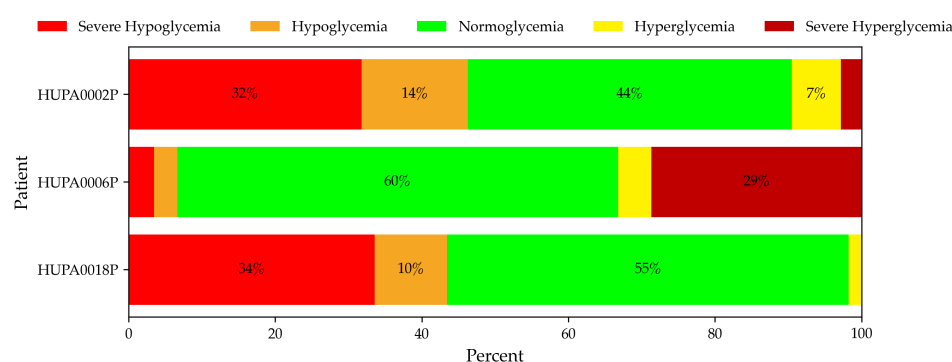


Figure 7. Percentage of glucose observations in each clinical range for the three selected patients over the evaluation window. Ranges were defined as severe hypoglycemia (<54 mg/dL), hypoglycemia (54–69 mg/dL), normoglycemia (70–180 mg/dL), hyperglycemia (181–250 mg/dL), and severe hyperglycemia (>250 mg/dL).

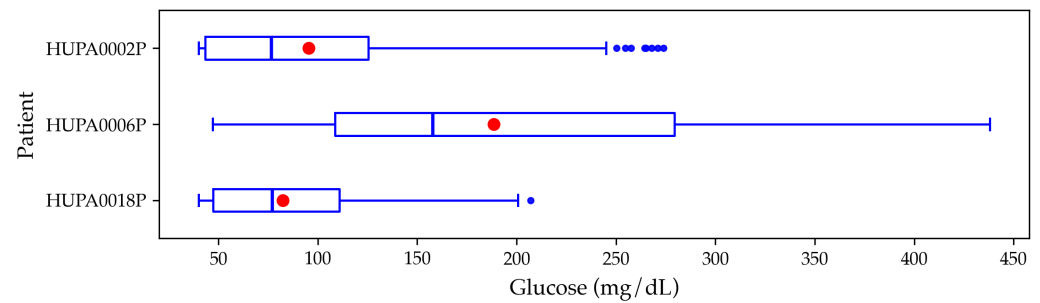


Figure 8. Distribution of CGM glucose values for the three selected patients. Red dots indicate the mean glucose level in each case.

To assess the behavior of the local model under these conditions, we compared the best-performing model from each forecasting family. Table 8 summarizes the corresponding local metrics. ARIMA yielded the lowest MAE and RMSE for HUPA0002P and HUPA0006P, while LightGBM achieved the highest R^2 and the lowest RMSE for HUPA0018P. Sundial, however, retained one of the smallest bias factors in the two cases, indicating a more balanced local error structure.

Table 8. Local forecasting performance for the three selected cases using the best-performing model from each family. Best values within each patient are highlighted in blue and worst values in red. For bias factor, the best value corresponds to the estimate closest to zero in absolute terms.

Patient	Model	R^2	MAE (mg/dL)	RMSE (mg/dL)	Bias Factor (%)
HUPA0002P	ARIMA	0.9788	4.66	8.68	0.16
	Sundial	0.9781	5.39	8.81	−0.16
	LightGBM	0.9718	5.94	10.01	1.25
	NHITS	0.9667	7.12	10.87	−0.90
	TimeGPT	0.9647	6.57	11.20	0.47
HUPA0006P	ARIMA	0.9877	6.53	11.81	0.11
	Sundial	0.9873	7.13	12.01	0.55
	TimeGPT	0.9809	9.25	14.70	1.50
	LightGBM	0.9708	10.81	18.19	0.18
	NHITS	0.9555	16.04	22.45	1.04
HUPA0018P	LightGBM	0.9592	4.60	7.33	0.58
	ARIMA	0.9478	3.97	8.29	−1.08
	Sundial	0.9454	4.55	8.48	−0.09
	TimeGPT	0.9388	4.76	8.98	−1.09
	NHITS	0.9076	7.11	11.03	−0.86

Figures 9 and 10 provide a more detailed view of local dynamics. The observed and predicted trajectories show that all models capture stable glycemic segments reasonably well, whereas the main deterioration occurs around abrupt excursions, where the differences between methods become more pronounced. This behavior is especially evident in the residual series, where the spread of errors increases near peaks, steep drops, and late rebounds.

The three cases illustrate different forecasting difficulties. HUPA0002P is dominated by prolonged hypoglycemia followed by a sharp midday excursion, where residual dispersion increases around the peak, indicating difficulty in matching both timing and amplitude. HUPA0006P is the most demanding case because it combines extreme hyperglycemia

with a steep subsequent drop; here, the variance of the residuals increases markedly around the turning point, and large deviations are observed across all models. HUPA0018P, in turn, remains comparatively smooth for most of the window, but ends with a late upward excursion, which produces predominantly negative residuals and reveals delayed or attenuated responses by several methods.

Across the three cases, the same local pattern is preserved. Errors remain small during relatively smooth segments but expand around clinically relevant transitions. In that setting, Sundial, ARIMA, and LightGBM exhibit the most reliable local behavior, whereas NHITS is more sensitive to abrupt changes and shows a wider residual dispersion.

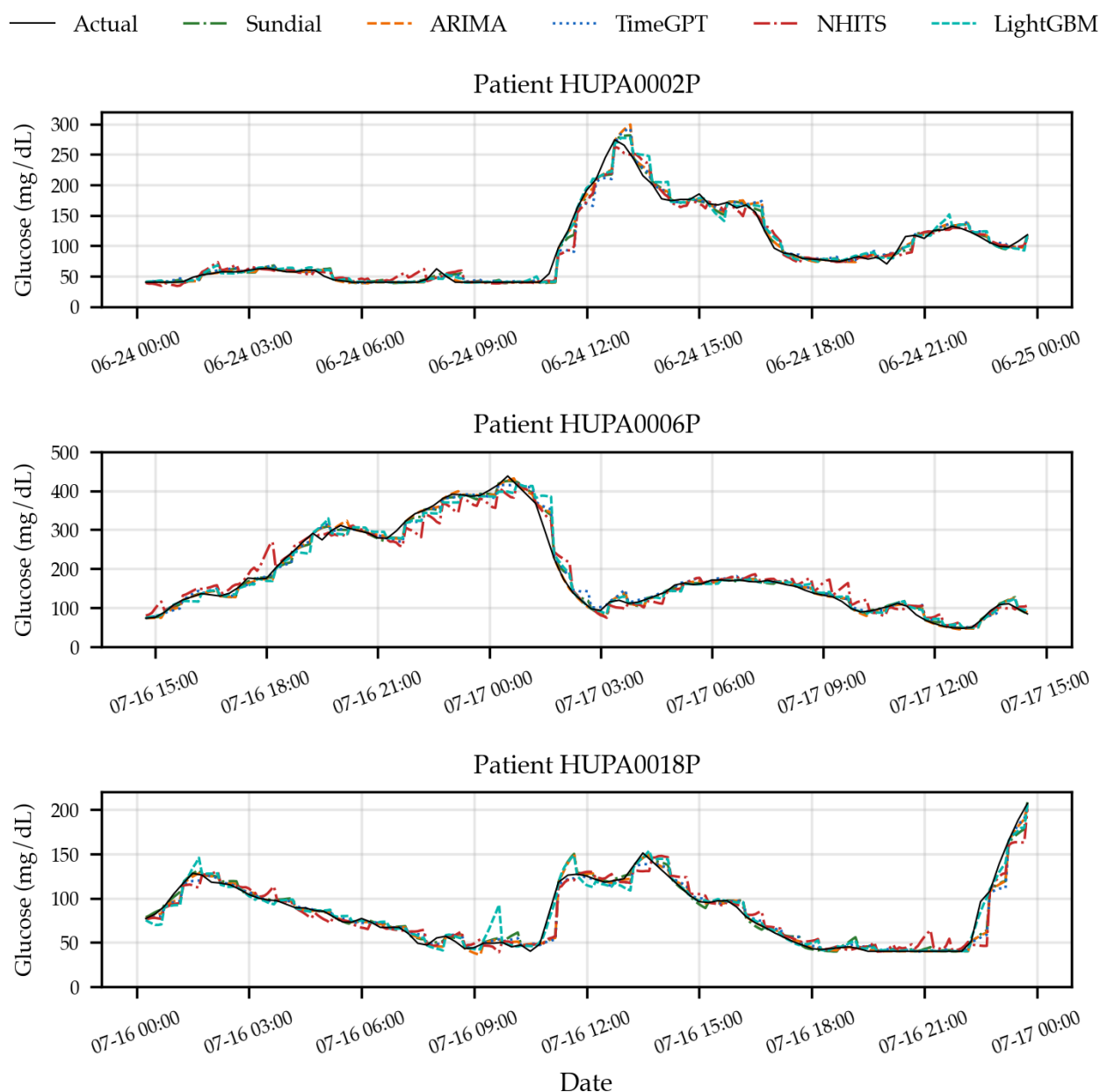


Figure 9. Observed CGM glucose trajectories and model forecasts for the three selected patients using Sundial, ARIMA, TimeGPT, LightGBM, and NHITS.

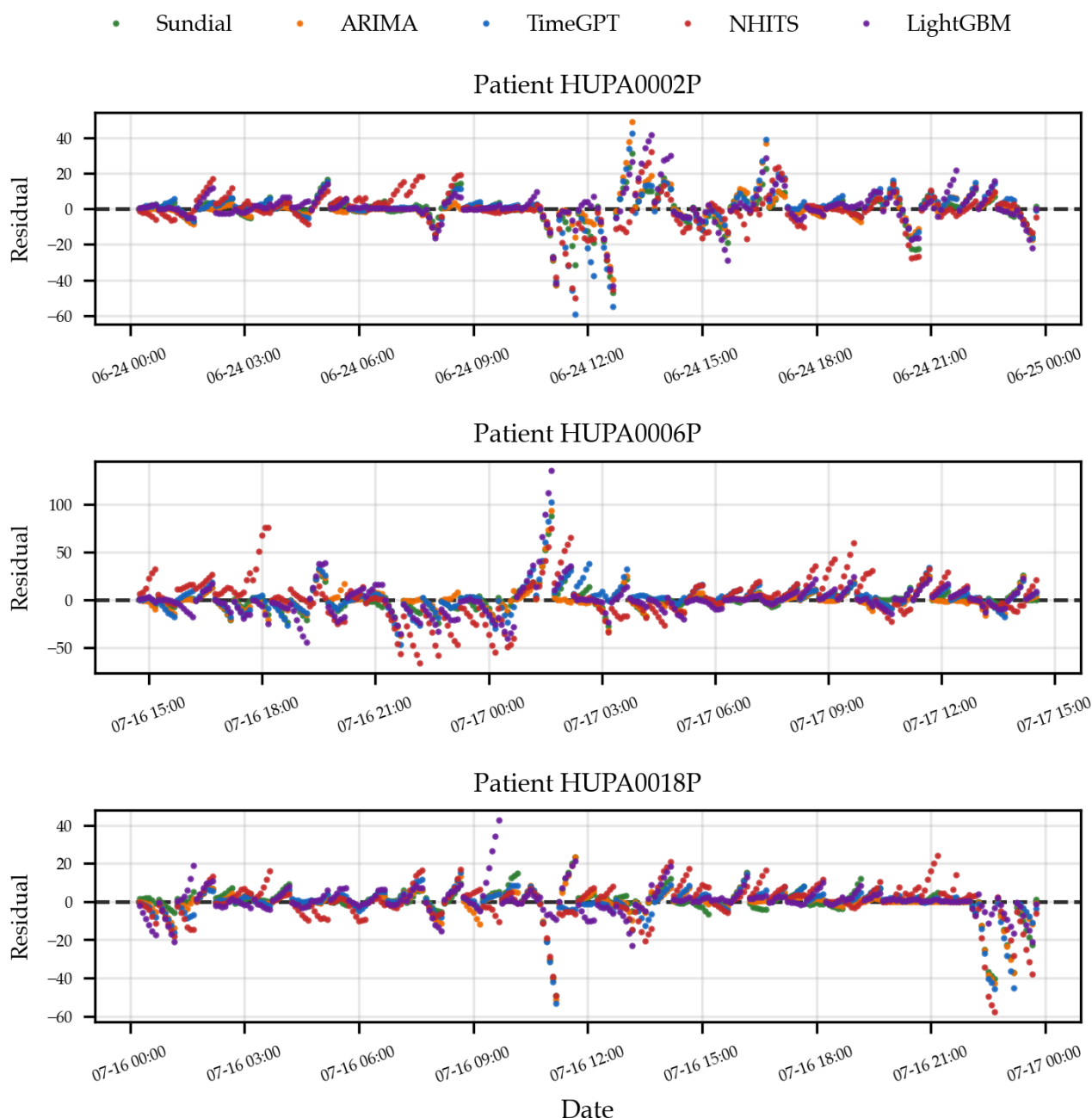


Figure 10. Residual series for the three selected patients using the best-performing model for Sundial, ARIMA, TimeGPT, LightGBM, and NHITS.

4. Conclusions and Future Work

This study presented a systematic evaluation of foundation time-series models for 30 min ahead glucose forecasting in adults with T1DM, using only CGM signals from the HUPA–UCM dataset. Statistical, machine learning, and deep learning models were included as comparators under the same local CGM-only setting. Beyond aggregate metrics, the analysis considered cohort-level performance, patient-level variability, and clinically relevant glycemic dynamics.

The results show that foundation models are a strong alternative for short-horizon CGM forecasting in adults with T1DM. Sundial achieved the lowest global error, while TimeGPT and Chronos remained among the most competitive models, indicating that large-scale pretraining can be useful even when only recent glucose history is available.

Under the adopted experimental protocol, foundation models also outperformed the conventional deep learning architectures evaluated in this study. However, their advantage was not uniform across patients. ARIMA remained highly competitive and achieved the best local performance in selected challenging cases, while LightGBM was also effective for specific individuals. These findings suggest that foundation models should be viewed as part of a patient-specific forecasting strategy rather than as universal replacements for classical methods.

From a clinical perspective, the findings are relevant because useful short-term forecasts were obtained without manually recorded inputs such as meals, insulin doses, or physical activity. In real-world settings, these variables are often incomplete, delayed, or inconsistently reported, whereas CGM remains the most systematic source of temporal information. Under these conditions, a univariate and patient-centered evaluation provides a practical baseline for assessing the intrinsic predictive value of CGM history.

Several limitations should be acknowledged. First, the analysis was restricted to a single public dataset of 25 adults with T1DM, which may limit external validity and generalizability to broader populations and settings, including pediatric patients, older adults, different treatment regimens, CGM devices, and clinical environments. Second, the study followed a strictly CGM-only design; therefore, the potential contribution of exogenous variables such as insulin, meals, physical activity, and wearable-derived signals was not assessed. Consequently, the present results should not be interpreted as a direct comparison against state-of-the-art multimodal glucose forecasting approaches, nor as an estimate of the maximum performance attainable when richer clinical and behavioral information is available. This choice also preserved a comparable input setting across heterogeneous forecasting families, since not all evaluated models or implementations provide equivalent mechanisms for incorporating exogenous variables under the same experimental conditions. Third, all models were evaluated locally, with each patient modeled independently, so global multi-patient approaches capable of learning shared glycemic patterns while preserving individual adaptation remain to be explored.

In addition, the evaluation focused on point forecasts and did not include predictive uncertainty estimates, such as prediction intervals or calibrated probabilistic forecasts. This is clinically relevant because uncertainty information may help to distinguish confident short-term predictions from forecasts that require greater caution, especially near hypo- or hyperglycemic thresholds.

This study also did not perform exhaustive model-specific hyperparameter optimization. This decision preserved a standardized and reproducible comparison across heterogeneous forecasting families under the same local CGM-only protocol. However, it may have affected the relative performance of ML and DL baselines, particularly models that are more sensitive to input length, learning rate, network size, regularization, early stopping, and other training choices. A fully nested tuning procedure within the walk-forward design would have substantially increased the experimental burden, given the repeated evaluation across patients, models, and temporal splits. Therefore, the results should be interpreted as a comparison under common experimental conditions rather than as the maximum attainable performance of each model family.

Finally, computational efficiency and inference time were not quantitatively evaluated. Although all models were assessed under the same patient-specific walk-forward validation protocol, the forecasting families differ substantially in implementation, execution environment, and deployment requirements. In particular, locally executed models and externally hosted foundation models may not be directly comparable in runtime because API latency, server-side infrastructure, hardware resources, and software implementation can influence measured execution times. Therefore, the present results should be inter-

puted primarily in terms of predictive accuracy and clinical error behavior rather than computational efficiency.

Future studies should evaluate foundation models that explicitly support exogenous covariates, including recent TimeGPT-2 variants and emerging time-series foundation models such as Moirai, TimesFM, FlowState, TiRex, Toto, and TabPFN-TS. These models should be tested against CGM-only baselines to determine when additional clinical, behavioral, or wearable-derived information provides meaningful gains. Ongoing work should also investigate global foundation-based forecasting strategies trained across multiple patients, incorporating patient-level descriptors and time-varying covariates to learn shared temporal patterns while preserving individual adaptation. Future benchmarks should also include calibrated prediction intervals, coverage and interval-width metrics, budget-matched or nested hyperparameter tuning, and standardized runtime evaluation, including training time, inference latency, memory consumption, and deployment feasibility for real-time or near-real-time CGM forecasting applications. External validation on larger, independent, and more diverse CGM datasets, including pediatric cohorts and data collected across different clinical environments, will be necessary for clarifying when foundation models provide the greatest benefit and when simpler local models remain preferable.

Author Contributions: Conceptualization, R.C.D.-V., J.A.G.-R. and E.R.-V.; methodology, R.C.D.-V., A.G.-O. and E.R.-V.; software, R.C.D.-V., A.G.-O. and J.I.C.-C.; validation, R.C.D.-V., A.G.-O., J.A.G.-R., J.I.C.-C. and E.R.-V.; formal analysis, R.C.D.-V. and A.G.-O.; investigation, R.C.D.-V., J.A.G.-R., J.I.C.-C. and E.R.-V.; resources, J.A.G.-R.; data curation, J.A.G.-R. and J.I.C.-C.; writing—original draft preparation, R.C.D.-V. and A.G.-O.; writing—review and editing, A.G.-O., J.A.G.-R., J.I.C.-C. and E.R.-V.; visualization, R.C.D.-V., A.G.-O. and J.I.C.-C.; supervision, J.A.G.-R. and E.R.-V.; project administration, E.R.-V.; funding acquisition, E.R.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This study is a secondary analysis of the publicly available HUPA-UCM Diabetes Dataset and did not involve new recruitment of participants, direct patient contact, or access to identifiable personal information by the authors. According to the dataset descriptor, the original data collection was approved by the Ethics Committee of Hospital Universitario Príncipe de Asturias (Protocol No. EC/11/2018; approval date: 12 December 2018), and all participants provided informed consent. Therefore, no additional ethical approval was required for the present secondary analysis of anonymized public data.

Data Availability Statement: The dataset used in this study is publicly available through Mendeley Data as the HUPA-UCM Diabetes Dataset: <https://data.mendeley.com/datasets/3hbcscwz44/1>, (accessed on 24 May 2025) DOI: <https://doi.org/10.17632/3hbcscwz44.1>.

Acknowledgments: The authors acknowledge the financial support provided by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through scholarship grants awarded to Roberto Carlos Diaz-Velazco (No. 4045886), Alberto Gudiño-Ochoa (No. 4074738) and Jorge Iván Cuevas-Chávez (No. 4045862). Additionally, during the preparation of this manuscript, the authors used ChatGPT (Model 5.5, OpenAI) for the purposes of assisting in the design of the Graphical Abstract and formatting Table 2 in Section 2.2. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CGM	Continuous glucose monitoring
T1DM	Type 1 diabetes mellitus
DM	Diabetes mellitus
ML	Machine learning
DL	Deep learning
RNNs	Recurrent neural networks
LSTM	Long short-term memory
GRUs	Gated recurrent units
LLM	Large language model
MLP	Multilayer perceptron
CSII	Continuous subcutaneous insulin infusion
MDI	Multiple doses of insulin
NHITS	Neural hierarchical interpolation for time series
NBEATSx	Neural basis expansion analysis
MAE	Mean absolute error
KANs	Kolmogorov-Arnold networks
RMSE	Root mean squared error
DM	Diebold–Mariano

References

1. International Diabetes Federation. *IDF Diabetes Atlas 11th Edition*. Available online: <https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/> (accessed on 24 May 2026).
2. ElSayed, N.A.; Aleppo, G.; Aroda, V.R.; Bannuru, R.R.; Brown, F.M.; Bruemmer, D.; Collins, B.S.; Gaglia, J.L.; Hilliard, M.E.; Isaacs, D.; et al. 2. Classification and diagnosis of diabetes: Standards of care in diabetes—2023. *Diabetes Care* **2023**, *46*, S19–S40. [[PubMed](#)]
3. Addissouky, T.A.; Ali, M.M.; El Sayed, I.E.T.; Wang, Y. Type 1 diabetes mellitus: Retrospect and prospect. *Bull. Natl. Res. Cent.* **2024**, *48*, 42. [[CrossRef](#)]
4. Nwokolo, M.; Hovorka, R. The artificial pancreas and type 1 diabetes. *J. Clin. Endocrinol. Metab.* **2023**, *108*, 1614–1623. [[CrossRef](#)] [[PubMed](#)]
5. Fabris, C.; Kovatchev, B. The closed-loop artificial pancreas in 2020. *Artif. Organs* **2020**, *44*, 671–679. [[PubMed](#)]
6. Fan, W.; Deng, C.; Xu, R.; Liu, Z.; Leslie, R.D.; Zhou, Z.; Li, X. Efficacy and safety of automated insulin delivery systems in patients with type 1 diabetes mellitus: A systematic review and meta-analysis. *Diabetes Metab. J.* **2025**, *49*, 235–251. [[CrossRef](#)] [[PubMed](#)]
7. Katsarou, D.N.; Georga, E.I.; Christou, M.A.; Christou, P.A.; Tigas, S.; Papaloukas, C.; Fotiadis, D.I. Optimizing hypoglycaemia prediction in type 1 diabetes with Ensemble Machine Learning modeling. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 46. [[CrossRef](#)] [[PubMed](#)]
8. Kozinetz, R.M.; Berikov, V.B.; Semenova, J.F.; Klimontov, V.V. Machine learning and deep learning models for nocturnal high-and low-glucose prediction in adults with type 1 diabetes. *Diagnostics* **2024**, *14*, 740. [[CrossRef](#)] [[PubMed](#)]
9. Singh, J.; Singh, B.P.; Dandotiya, M.; Ekvitayavetchanukul, P.; Rana, M.; Singh, B. Predicting Blood Glucose Levels in Type 1 Diabetes Using Deep Learning and Regression Techniques. In *2025 International Conference on Intelligent Control, Computing and Communications (IC3)*; IEEE: Piscataway, NJ, USA, 2025; pp. 86–91.
10. Cinar, B.; van den Boom, L.; Maleshkova, M. Review of Machine Learning Models in Short-and Long-term Glucose Forecasting and Hypoglycemia Classification. *Inform. Med. Unlocked* **2025**, *60*, 101723.
11. Gudiño-Ochoa, A.; Ruiz-Velázquez, E.; García-Rodríguez, J.A.; Ochoa-Ornelas, R.; Uribe-Toscano, S. Non-Invasive Blood Glucose Estimation from Exhaled Breath: Patient-Level Validation of a Compact Electronic Nose Approach. *AI* **2026**, *7*, 213. [[CrossRef](#)]
12. Nemat, H.; Khadem, H.; Elliott, J.; Benaissa, M. Data-driven blood glucose level prediction in type 1 diabetes: A comprehensive comparative analysis. *Sci. Rep.* **2024**, *14*, 21863. [[CrossRef](#)] [[PubMed](#)]
13. Basile, I.; Sannino, G. Blood glucose level prediction in type 1 diabetes: A comparative analysis of interpretable artificial intelligence approaches. *Results Eng.* **2025**, *25*, 103681.

14. Jacobs, P.G.; Herrero, P.; Facchinetti, A.; Vehi, J.; Kovatchev, B.; Breton, M.D.; Cinar, A.; Nikita, K.S.; Doyle, F.J.; Bondia, J.; et al. Artificial intelligence and machine learning for improving glycemic control in diabetes: Best practices, pitfalls, and opportunities. *IEEE Rev. Biomed. Eng.* **2023**, *17*, 19–41. [\[CrossRef\]](#)
15. Langarica, S.; de la Vega, D.; Cariman, N.; Miranda, M.; Andrade, D.C.; Núñez, F.; Rodríguez-Fernandez, M. Deep learning-based glucose prediction models: A guide for practitioners and a curated dataset for improved diabetes management. *IEEE Open J. Eng. Med. Biol.* **2024**, *5*, 467–475. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Nordin, M.S.M.; Mahmud, F. Univariate and multivariate time series blood glucose prediction with lstm deep learning model. *Evol. Electr. Electron. Eng.* **2024**, *5*, 276–285.
17. Zhang, M.; Flores, K.B.; Tran, H.T. Deep learning and regression approaches to forecasting blood glucose levels for type 1 diabetes. *Biomed. Signal Process. Control* **2021**, *69*, 102923. [\[CrossRef\]](#)
18. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Woo, W.L.; Wei, B.; Pardo-Quiles, D.J. A comparison of feature selection and forecasting machine learning algorithms for predicting glycaemia in type 1 diabetes mellitus. *Appl. Sci.* **2021**, *11*, 1742. [\[CrossRef\]](#)
19. Gancotti, R.; Bosoni, P.; Vizza, P.; Tradigo, G.; Gnasso, A.; Guzzi, P.H.; Bellazzi, R.; Irace, C.; Veltri, P. Forecasting glucose values for patients with type 1 diabetes using heart rate data. *Comput. Methods Programs Biomed.* **2024**, *257*, 108438. [\[CrossRef\]](#) [\[PubMed\]](#)
20. García-Rodríguez, J.A.; Diaz-Velazco, R.C.; Gudiño-Ochoa, A.; Navarrete-Guzmán, A. Robust two-degree-of-freedom control: A simulation-based approach for adults population with T1D. *Results Control Optim.* **2024**, *16*, 100454. [\[CrossRef\]](#)
21. Shuvo, M.M.H.; Islam, S.K. Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 1612–1623. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Neumann, A.; Zghal, Y.; Cremona, M.A.; Hajji, A.; Morin, M.; Rekik, M. A data-driven personalized approach to predict blood glucose levels in type-1 diabetes patients exercising in free-living conditions. *Comput. Biol. Med.* **2025**, *190*, 110015. [\[PubMed\]](#)
23. Jaloli, M.; Cescon, M. Long-term prediction of blood glucose levels in type 1 diabetes using a cnn-lstm-based deep neural network. *J. Diabetes Sci. Technol.* **2023**, *17*, 1590–1601.
24. Butunoi, B.P.; Stolojescu-Crisan, C.; Negru, V. Short-term glucose prediction in Type 1 Diabetes. *Procedia Comput. Sci.* **2024**, *238*, 41–48. [\[CrossRef\]](#)
25. Butt, H.; Khosa, I.; Iftikhar, M.A. Feature transformation for efficient blood glucose prediction in type 1 diabetes mellitus patients. *Diagnostics* **2023**, *13*, 340. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Chandran, S.R.; Teh, M.M.; Tan, H.C.; Oo, M.Z.; Chang, A.A.L.; Gardner, D. Identifying high glucose variability using non-glycemic factors in low continuous glucose monitoring use settings. *Prim. Care Diabetes* **2025**, *19*, 624–628. [\[CrossRef\]](#)
27. Sánchez, O.D.; Ruiz-Velázquez, E.; Alanís, A.Y.; Quiroz, G.; Torres-Treviño, L. Parameter estimation of a meal glucose–insulin model for T1DM patients from therapy historical data. *IET Syst. Biol.* **2019**, *13*, 8–15. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Alanis, A.Y.; Sanchez, O.D.; Velázquez, E.R. Deep neuronal network-based glucose prediction for personalized medicine. In *Feedback Control for Personalized Medicine*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 87–108.
29. Lara-Abelenda, F.J.; Chushig-Muzo, D.; Peiro-Corbacho, P.; Wäagner, A.M.; Granja, C.; Soguero-Ruiz, C. Personalized glucose forecasting for people with type 1 diabetes using large language models. *Comput. Methods Programs Biomed.* **2025**, *265*, 108737. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Rancati, S.; Bosoni, P.; Schiaffini, R.; Deodati, A.; Mongini, P.A.; Sacchi, L.; Toffanin, C.; Bellazzi, R. Exploration of foundational models for blood glucose forecasting in type-1 diabetes pediatric patients. *Diabetology* **2024**, *5*, 584–599. [\[CrossRef\]](#)
31. Hidalgo, J.I.; Alvarado, J.; Botella, M.; Aramendi, A.; Velasco, J.M.; Garnica, O. HUPA-UCM diabetes dataset. *Data Brief* **2024**, *55*, 110559. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Shumway, R.H.; Stoffer, D.S. ARIMA models. In *Time Series Analysis and Its Applications: With R Examples*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 75–163.
33. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018. Available online: <https://otexts.com/fpp2/> (accessed on 25 June 2026).
34. Hyndman, R.J.; Koehler, A.B.; Ord, J.K.; Snyder, R.D. A state space framework for automatic forecasting using exponential smoothing methods. *Int. J. Forecast.* **2002**, *18*, 439–454. [\[CrossRef\]](#)
35. Hyndman, R.; Koehler, A.; Ord, K.; Snyder, R. *Forecasting with Exponential Smoothing: The State Space Approach*; Springer: Berlin/Heidelberg, Germany, 2008.
36. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*; NeurIPS: San Diego, CA, USA, 2018; Volume 31.
37. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; NeurIPS: San Diego, CA, USA, 2017; Volume 30.
38. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
39. Chen, T. *XGBoost: A Scalable Tree Boosting System*; Cornell University: New York, NY, USA, 2016.
40. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [\[CrossRef\]](#)

41. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*; Omnipress: Madison, WI, USA, 2010; pp. 807–814.
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
43. Sak, H.; Senior, A.; Beaufays, F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv* **2014**, arXiv:1402.1128.
44. Beauchamp, J.; Bunesco, R.; Marling, C.; Li, Z.; Liu, C. LSTMs and deep residual networks for carbohydrate and bolus recommendations in type 1 diabetes management. *Sensors* **2021**, *21*, 3303. [[CrossRef](#)] [[PubMed](#)]
45. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
46. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 103–111.
47. Challu, C.; Olivares, K.G.; Oreshkin, B.N.; Ramirez, F.G.; Canseco, M.M.; Dubrawski, A. Nhits: Neural hierarchical interpolation for time series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 6989–6997. [[CrossRef](#)]
48. Olivares, K.G.; Challu, C.; Marcjasz, G.; Weron, R.; Dubrawski, A. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *Int. J. Forecast.* **2023**, *39*, 884–900.
49. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv* **2024**, arXiv:2404.19756.
50. Liao, W.; Wang, S.; Yang, D.; Yang, Z.; Fang, J.; Rehtanz, C.; Porté-Agel, F. TimeGPT in load forecasting: A large time series model perspective. *Appl. Energy* **2025**, *379*, 124973.
51. Gonzalez, F.C.; Gudiño-Ochoa, A. Temporal hierarchical forecast reconciliation of photovoltaic power generation from heterogeneous base models. *Meas. Energy* **2026**, *10*, 100094. [[CrossRef](#)]
52. Liu, Y.; Qin, G.; Shi, Z.; Chen, Z.; Yang, C.; Huang, X.; Wang, J.; Long, M. Sundial: A family of highly capable time series foundation models. *arXiv* **2025**, arXiv:2502.00816.
53. Ansari, A.F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S.S.; Arango, S.P.; Kapoor, S.; et al. Chronos: Learning the language of time series. *arXiv* **2024**, arXiv:2403.07815.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.