Can Time Series Foundation Models Forecast Nocturnal Hypoglycemia? A Benchmarking Study in Type 1 Diabetes

Christopher Risi
School of Computer Science
University of Waterloo
cjrisi@uwaterloo.ca

Tony Chan
Department of ECE
University of Waterloo
t3chan@uwaterloo.ca

Shivam Jindal
School of Computer Science
University of Waterloo
s6jindal@uwaterloo.ca

Alyssa D'Souza
School of Computer Science
University of Waterloo
a46dsouz@uwaterloo.ca

Arohi Gopal
School of Computer Science
University of Waterloo
a5gopal@uwaterloo.ca

Divyansh Bhandari
School of Computer Science
University of Waterloo
d4bhanda@uwaterloo.ca

Kirpa Chandok
School of Computer Science
University of Waterloo
kchandok@uwaterloo.ca

Jonathan Gong
School of Computer Science
University of Waterloo
email@email.com

***Abstract*—Nocturnal hypoglycemia in type 1 diabetes (T1D) accounts for 5–6% of T1D mortality and imposes significant psychological burden on patients and caregivers. Continuous glucose monitors (CGMs) have improved real-time glycemic awareness, yet clinically actionable long-horizon forecasting remains unsolved. Recent time series foundation models (TSFMs) have shown strong generalization across temporal domains, but their application to blood glucose forecasting is critically underexplored. Here we benchmark six state-of-the-art TSFMs—`Chronos2`, `Sundial`, `TiDE`, `TimesFM`, `TimeGrad`, and `TinyTimeMixer`—in zero-shot and fine-tuned settings across four public T1D CGM datasets, targeting 8-hour nocturnal forecasting horizons. Attention-based architectures consistently outperform MLP-based counterparts in capturing blood glucose trajectory morphology, with `TimesFM` achieving best-in-class performance. These findings challenge the assumed competitiveness of MLP-based TSFMs for physiologically complex time series and establish a reproducible benchmark advancing the feasibility of deploying TSFMs for nocturnal hypoglycemia prevention. Code is available at https://github.com/Blood-Glucose-Control/nocturnal-hypo-gly-prob-forecast/.**

I. Introduction

Nocturnal hypoglycemia is a dangerous condition that can occur in individuals with type 1 diabetes, where blood glucose levels drop dangerously low during sleep. Hypoglycemia can lead to immediate severe consequences, including seizures, loss of consciousness, and even death [1]–[3]. Death due to nocturnal hypoglycemia is often referred to as *dead in bed* syndrome and accounts for 5%–6% of mortality cases in individuals with type 1 diabetes [3].

Living with this risk is a significant burden for individuals with type 1 diabetes and their caregivers, leading to anxiety and fear around sleep, and can significantly impact their quality of life [4], [5]. Fear of nocturnal hypoglycemia (FoNH) is a leading cause of diabetes distress (DD); both FoNH and DD are associated with lower levels of self-care, glycemic control, and emotional well-being [6].

Even if a type 1 diabetic does not experience FoNH, and manages to avoid the severe complications due to nocturnal hypoglycemia, they are burdened with frequent sleep disruptions [7]. Exploratory data analysis with continuous glucose monitor (CGM) data used in our study indicate that nearly 40% of nights are affected by sleep disruptions due to hypoglycemia [8]. CGMs facilitate nearly real-time estimates of blood glucose via measurements of interstitial glucose, which estimate blood glucose levels and transfer this information via Bluetooth to mobile devices.

With the advent of CGMs, we expect that frequency of hypoglycemia complications may now be decreasing, however hypoglycemia complications can still frequently occur due to events such as device malfunctions or not waking from mobile alarms. Even with these advancements, the fear and distress around nocturnal hypoglycemia remains prevalent [6]. Person-reported outcome (PRO) studies using the Hypoglycemia Fear Survey II (HFS-II) and T1 Diabetes Distress Scale (T1-DDS) indicate that even two hour forecasts would help alleviate fear and distress around nocturnal hypoglycemia, and would be a significant improvement in quality of life for individuals with type 1 diabetes [6].

Unfortunately, modern machine learning and artificial intelligence techniques have yet to demonstrate the ability to forecast nocturnal hypoglycemia with a two hour forecast horizon (fh) that would be practical in clinical settings. And given that our data analysis shows 24% of nights are affected by sleep disruptions due to hypoglycemia between the hours of 4 AM and 7 AM, there is a clear need for methods capable of much longer forecasting horizons.

A. Motivation

Over the past decade, two significant developments have made long-horizon forecasting in diabetes worth exploring. First, the advent of CGMs has led to a significant increase

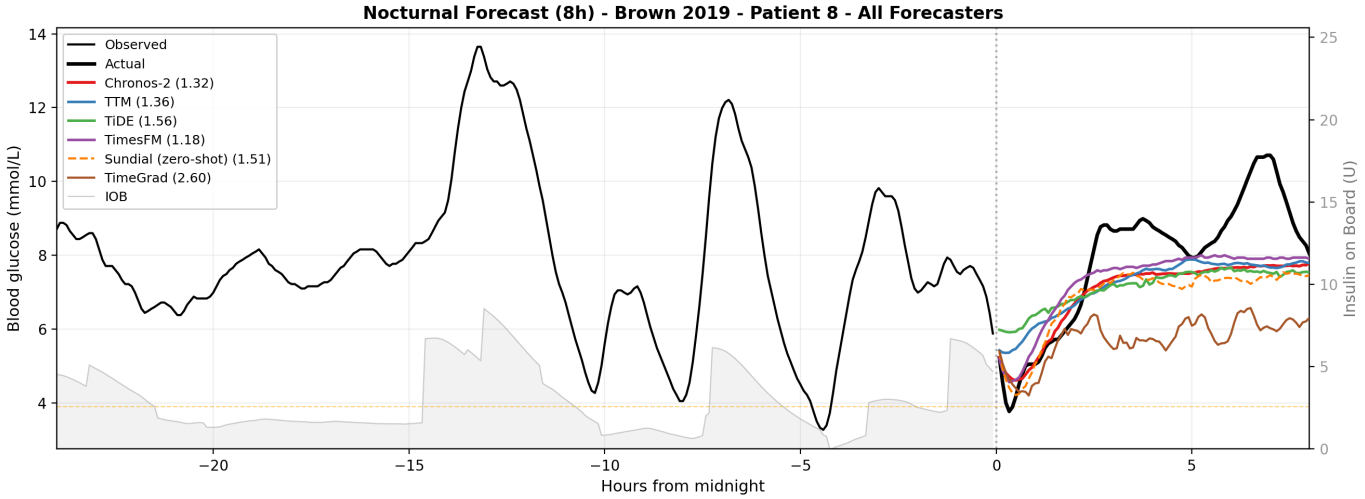

Fig. 1. Forecast of blood glucose concentrations for a single episode anchored at 12 AM, with a forecasting horizon of 8 hours. The black line represents the actual CGM values, and also shows the IOB that the model observes in the context window. We presented all of our forecasters, which have different architectural paradigms, with the same episode to provide a visual comparison of their forecasts. **Note:** the MLP-based architectures like `TiDE` and `TTM` fail to capture the shape of the blood glucose curve, they tend to minimize RMSE by cutting through the middle of the noisy CGM curve. In contrast, architectures like `Chronos2`, `Sundial`, and `TimesFM` are able to capture the shape of the blood glucose curve much better, which is likely due to their attention-based architectures that can capture complex temporal interactions between time steps.

in the amount of data available for analysis and forecasting in type 1 diabetes. The four datasets that we are using in this study contain CGM data from hundreds of patients, with measurements taken every 5 minutes, resulting in with over 50 million CGM readings that can be used to train and evaluate machine learning models for forecasting nocturnal hypoglycemia. Second, there have been significant advancements in AI/ML techniques, particularly in the area of time series foundation modelling. Time series foundation models (TSFMs) have shown promise in forecasting complex temporal patterns across various domains [9]–[11]. Their potential for forecasting nocturnal hypoglycemia in type 1 diabetes therefore warrants thorough investigation. There are at least 42 known factors that impact blood glucose concentrations in type 1 diabetes [12]. TSFMs are likely necessary for forecasting nocturnal hypoglycemia accurately because the dozens of factors that impact blood glucose concentrations are complex, non-linear, violate stationarity assumptions, and involve interaction effects between static categorical features, dynamic categorical features, and many of the time series covariates have known interaction effects that vary in time scale.

For example, basal insulin affects blood glucose over 24–36 hours while bolus insulin acts over 2–5 hours, and these effects are modulated by factors such as meal composition, exercise, and sleep quality—creating complex multi-scale interactions.

Nocturnal hypoglycemia forecasting represents one of the most challenging paradigms in time series prediction, requiring 8-hour (96 time-step) forecasting horizons, very long—42 hour, 512 time-step—context windows, dynamic long-range and short-range covariate interactions, and complete violation of any stationarity assumptions that would make traditional forecasting methodologies feasible.

The challenge with long-horizon forecasting in our problem, is that it is necessary to eliminate non-stationarity assumptions. Otherwise these lead to spurious regressions that negatively impact short-horizon model performance (i.e., first hours of sleep with highest volatility), however, this elimination then hinders long-horizon forecasting, i.e., forecasting until the (4 AM–8 AM) window because it hurts our ability to capture long-horizon covariate interactions [13].

Supporting context windows of at least 42 hours requires architectures with massive receptive fields and memory mechanisms that do not suffer from vanishing gradients or catastrophic forgetting. To add to this, long horizon forecasting has also been shown in Zeng *et al.* to be challenging for Transformer-based architectures because permutation-invariant self-attention mechanisms typically result in temporal information loss [14]. This issue motivated us to include a variety of SoTA architectures in our benchmarking study that have different architectural paradigms, including MLP-based, diffusion-based, and attention-based, to evaluate which architectural paradigms might be more successful.

This long context window also leads to necessary practical considerations around computational complexity of the TSFMs with regards to training and but especially inference. If the long term goal is to use these models at scale in real-time clinical settings with millions of patients, then inference time is a critical consideration that must be taken into account when designing and evaluating the models. Models that cannot be run on edge devices and require cloud-based inference with long latency times are unlikely to be adopted in real-world production settings due to populations needing inference roughly around the same time, creating expensive peak-to-

trough computation demand. Therefore we must consider the computational complexity of the models as a critical factor in their design and evaluation.

None of the existing SoTA-TSFM architectures are likely to solve all of these challenges, our goal with this benchmarking study is to provide a introductory evaluation of the current SoTA-TSFM architectures, and to provide an assessment of the feasibility of solving this unique forecasting problem. This study provides valuable insights into what architectural paradigms are currently most successful for forecasting nocturnal hypoglycemia in type 1 diabetes, and provides guidance on future directions for research in this area.

TABLE I
DATA SUMMARY AND HOLDOUT CONFIGURATIONS

Dataset	Patients	Duration	Training	Held-out
Aleppo et al [8]	225	6 months	11.1M	2.6M
Brown et al [15]	168	6 months	8.0M	1.8M
Lynch et al [16]	440	13 weeks	7.9M	1.8M
Tamborlane et al [17]	451	6 months	14.8M	3.4M

B. Related Works

There have been a limited number of studies that use TSFM architectures in the context of diabetes. These studies show early promising results typically improving over traditional time-series models, but still lack meaningful clinical relevance due to the design of their forecasting problems, i.e., short horizons (15 minute–120 minute forecasts), small datasets (0.225M CGM readings–1.6M CGM readings), and often training on non-diabetic populations [18]–[21]. Contrasting with our study, which trains exclusively with T1D patients, via a sliding window over ∼40M CGM readings and then evaluates numerous models by anchoring our evaluation to begin at 12 AM with 8-hour, 96 time-step forecasts, on ∼10M CGM readings.

Rancati *et al.* fine-tuned TimeGPT, one of the earliest TSFMs, on short-horizon 15 minute–30 minute forecasts of blood glucose concentrations in a small dataset of 0.225M CGM readings from 15 patients [18], [22]. Like us, they also compared the performance of TimeGPT to other SoTA time-series architectures that were not foundation models like the encoder-deocder model Time-series Dense Encoder (TiDE) [23], and the MLP-based Time-Series Mixer (TSMixer) [24].

CGM-Large Sensor Model (CGM-LSM) used a Transformer decoder-based architecture trained on 1.6M CGM readings focused on on (0.5 h–2 h) forecasts [19].

GluFormer is an interesting generative foundation model using self-supervised representation learning but was used to forecast clinical measures like A1c, visceral adipose tissue, and liver function rather than blood glucose itself [20].

DiabLLM like previous papers focuses on forecasting (0.5 h–0.75 h) horizons, and fine-tunes two existing architectures Time-LLM and the first version of chronos [21].

C. Problem Definition

The problem we aim to solve is to minimize the loss function for forecasting blood glucose concentrations in individuals with type 1 diabetes using time series foundation models (TSFMs). We present our problem definition using common notation found in popular TSFM studies like TimesFM and TiDE [23], [25]. We evaluate our forecasters over a context window L time-steps and a forecasting horizon of F time-steps. The CGM values for the context window are denoted $\mathbf{y}_{1:L} := \{y_1, y_2, \ldots, y_L\}$ and the target CGM values are denoted $\mathbf{y}_{L+1:L+F} := \{y_{L+1}, y_{L+2}, \ldots, y_{L+F}\}$. Forecasted CGM values are denoted $\hat{\mathbf{y}}_{L+1:L+F} := \{\hat{y}_{L+1}, \hat{y}_{L+2}, \ldots, \hat{y}_{L+F}\}$.

The task for our TSFMs is to learn the function that maps the context window to the forecasting horizon:

$$f : \mathbf{y}_{1:L} \rightarrow \hat{\mathbf{y}}_{L+1:L+F}$$

Loss function: For this study we focus on point forecasting so we use Root Mean Square Error (RMSE) as our loss function, which is defined as:

$$\texttt{EvalLoss} = \frac{1}{N}\sum_{j=1}^{N} RMSE_j(\hat{\mathbf{y}}_{L+1:L+F} - \mathbf{y}_{L+1:L+F})$$

Where N is the number of midnight anchored episodes in the holdout set, and $RMSE_j$ is the RMSE for episode j, which is the square root of the average element-wise squared differences between the forecasted and actual CGM values for that episode.

II. METHODOLOGY

A. Data Preparation

We began our study by sourcing publicly available datasets curated by the Awesome-CGM GitHub repository [26]. We selected our four datasets to focus on T1D patients with 5-minute frequency CGM data, and with a large number of patients and CGM readings.

The CGM data is often irregularly sampled and not in a format suitable for direct training, therefore it required extensive cleaning, unit conversion (mmol/L), and coercing the data to regular 5-minute intervals, with many 5-minute windows not having a reading. When carbohydrate and insulin dosing data was available, we performed feature engineering to create time series features that capture the following:

- **Insulin on Board (IOB)/Insulin Activation:** The estimated remaining insulin to be absorbed into the blood stream and the expected currently active insulin in the blood stream at each time step, which is calculated based on the insulin dosing data and the pharmacokinetics of insulin.
- **Carbohydrates on Board (COB)/Carb Activation:** The estimated remaining carbohydrates to be absorbed into the blood stream and the expected current carbohydrates active in the blood stream at each time step, which is calculated based on the carbohydrate intake data and the digestion kinetics of carbohydrates.

Table I provides data summary and holdout configurations for each dataset, which were then generated for reproducibility and to ensure that all models were trained and evaluated on the same data splits. We then followed a holdout design nearly identical to Luo *et al.*'s CGM-LSM, where the data is split into

TABLE II
ZERO–SHOT (ZS) AND FINE-TUNED (FT) RESULTS ON THE FOUR BENCHMARK DATASETS.

	Chronos2		Sundial		TiDE		TimesFM		TimeGrad		TTM	
Dataset	ZS	FT	ZS	FT	ZS	FT	ZS	FT	ZS	FT	ZS	FT
Aleppo 2017	2.756	2.733	2.799	-	-	2.790	2.760	**2.731**	-	3.655	2.911	2.756
Brown 2019	2.650	**2.505**	2.679	-	-	2.543	2.667	2.525	-	3.160	2.843	2.566
Lynch 2022	4.490	3.979	4.233	-	-	4.115	4.247	**3.939**	-	4.550	4.698	4.085
Tamborlane 2008	3.353	3.266	3.306	-	-	3.291	3.312	**3.249**	-	3.867	3.424	3.258

training and holdout sets based both on temporal and patient dimensions [19].

For each dataset, we begin by holding out 10% of patients that the models will never be trained on, and then with the remaining 90% of patients we split on the final 10% of temporal data. This design represents a realistic clinical scenarios where we forecast for patients that the model has been trained on but used in the future and forecasting for new patients that the model has never seen before.

Finally, the patients from the training datasets were combined and shuffled into a single training set to avoid any model biasing performance to a specific dataset.

B. Model Selection and Training

We designed our github repository to be modular and extensible, with a clear separation between data processing, model training, and evaluation. This design allows us to easily add new models and evaluation procedures as we continue to iterate on this project, and also allows other researchers to easily reproduce our work and build upon it.

All architectures were sourced from publically available implementations (AutoGluon, sktime, HuggingFace) and inherit from a common base forecasting class, ensuring consistency and ease of integration. For this study we selected the our architectures based on their performance on other benchmarks, whether they provided architectural variety, customizability and the ease of integration into our codebase.

All models were configured to have a context window of 42-hours ($cw = 512$ time steps) and a forecasting horizon of 8-hours ($fh = 96$ time steps). We settled on a 42-hour context window because we wanted to provide the models with context around the entire day-night cycle, capturing both diurnal patterns and potential anomalies. The exact 42-hour context window was selected to help adhere to some pre-trained architectures like TinyTimeMixer which do not faciliate varying context lengths. Where possible we trained multiple versions of the same architecture with different data ablation setups (CGM-only, CGM+Insulin, CGM+Insulin+Carbohydrates) we only report the best performing models in Table II, but the setup is documented in our GitHub repository.

We then performed a sliding window training procedure that shifts forward in time by 1 time step, this design avoids data leakage. The training data was panneled by patient to avoid erroneous mixing of patient data that would hinder model performance.

Notably, our models are trained on all time steps rather than only nocturnal windows, enabling generalization to any time of day. However, daytime forecasting performance is expected to degrade because the model cannot observe meal events that occur during the forecasting horizon. We concede that this design likely hinders model performance and slows down training. In future work we will need to design a training procedure that avoids training evaluations on forecasting horizons where the model cannot observe important covariates like large meals and exercise that impact blood glucose concentrations.

The following architectures were included in our benchmarking study:

- `TinyTimeMixer` (TTM) [27]: A small full-supervised pretrained *MLP-based* TSFM with patch tokenization and a novel time-mixing architecture that captures temporal interactions between time steps.
- `TiDE` [23]: A self-supervised generative encoder-decoder *MLP-based* architecture designed for long-term time-series forecasting.
- `Chronos2` [28]: A large pretrained TSFM with patch tokenization with *group attention* between multivariate time-series.
- `TimesFM` [25]: A pretrained decoder-only *attention-based* model with input patching.
- `Sundial` [29]: A *generative* TSFM that is trained with a novel *continuous* tokenization procedure and a novel *TimeFlow Loss* based on flow matching.
- `TimeGrad` [30]: An autoregressive *diffusion-based* TSFM.

All models were trained on two NVIDIA RTX PRO 6000 Blackwell GPUs with 98GB of VRAM, and training time varied from a few minutes to a few days depending on the model and the dataset. Even with this setup, we were computationally constrained for some architectures which limited our ability to perform extensive hyperparameter tuning, and therefore we typically used the default hyperparameters provided by the original implementations of models, with modest adjustments.

All model configurations, training procedures, and hyperparameters are documented in our GitHub repository for reproducibility and transparency.

C. Evaluation

Our main evaluation of concern was on nocturnal hypoglycemia, therefore we designed our evaluation procedure to anchor the forecasts at 12 AM, and then evaluate the forecasts over an 8-hour forecasting horizon until 8 AM, which is the window where we see the highest frequency of nocturnal

hypoglycemia and sleep disruptions due to hypoglycemia in our exploratory data analysis.

We evaluated models in both zero-shot (ZS) and fine-tuned (FT) settings using a two-level holdout strategy to assess generalization. First, patients were partitioned into a 90% training cohort and a 10% held-out cohort that models never observe during initial training. Second, within each cohort, we perform a temporal split: the first 90% of each patient's time series is used for training or fine-tuning, while the final 10% serves as the evaluation window. In the ZS setting, models are evaluated on both the temporal and patient-level held-out data without further adaptation. In the FT setting, models are trained on the 90% training cohort (first 90% of timesteps) and evaluated on both held-out cohorts without further adaptation.

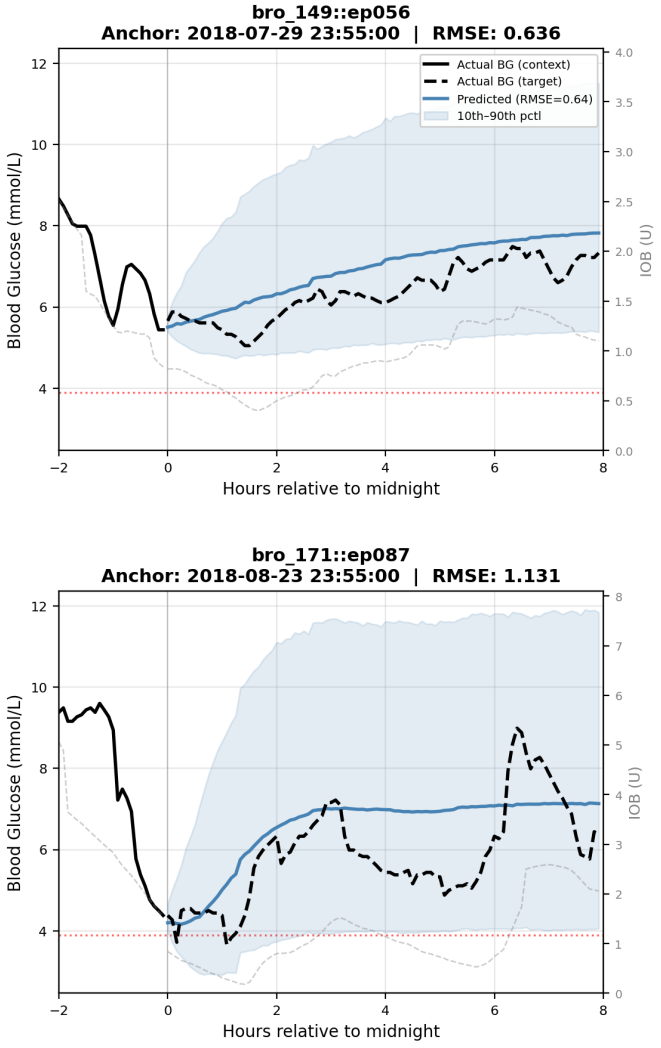

Fig. 2. Example of probabilistic forecasting. Probabilistic forecasts provide a distribution over possible future values rather than a single point estimate, which allows for uncertainty quantification. In the context of nocturnal hypoglycemia forecasting, this is likely the more appropriate and meaningful way to forecast for patients. It facilitates more informed clinical decision-making and accounts for the extremely volatile nature of blood glucose concentrations, as well as instances of poor data quality and untracked covariates.

III. Results

Table II presents RMSE scores across all six architectures in both zero-shot (ZS) and fine-tuned (FT) settings over the four benchmark datasets. We discuss the key findings below.

Attention-based architectures outperform MLP-based counterparts. Across all datasets, attention-based architectures—`TimesFM`, `Chronos2`, and `Sundial`—consistently achieve lower RMSE than MLP-based architectures `TiDE` and `TTM`. As illustrated qualitatively in Figure 1, MLP-based architectures exhibit a pronounced tendency to revert to the mean of the blood glucose trajectory rather than tracking its morphological shape. This mean-reversion behaviour minimizes squared error in expectation, but produces clinically misleading forecasts that fail to capture the dynamic rises and falls in blood glucose that are most critical for nocturnal hypoglycemia detection. This observation motivates future adoption of shape-aware loss functions during training. Differentiable approximations to trajectory similarity metrics, such as Soft-DTW [31] or DILATE loss [32], are strong candidates as they penalize temporal misalignment and shape error rather than purely element-wise deviation. While discrete Fréchet distance is a well-suited evaluation metric for measuring trajectory similarity post-hoc, it is not differentiable and therefore unsuitable as a training objective; differentiable relaxations such as Soft-DTW are preferred for backpropagation [33].

Fine-tuning consistently improves performance. Among models that support both ZS and FT evaluation, fine-tuning yields consistent improvements across datasets, suggesting that pre-trained TSFMs encode broadly useful temporal representations that can be meaningfully adapted to the physiological characteristics of patient cohorts. `TimesFM` achieves best-in-class performance after fine-tuning on three of four datasets, with RMSE ranging from 2.731 to 3.939 mmol/L. Notably, `Chronos2` performance was nearly on par with `TimesFM` on all evaluations and better on the Brown 2019 holdout cohort. These differences may grow with more extensive hyperparameter investigation, and different covariate setups.

Data quality issues in the Lynch 2022 dataset. The Lynch 2022 dataset exhibits substantially higher RMSE across all architectures compared to the other three datasets. Upon inspection, this dataset contains frequent and prolonged data gaps in both CGM readings and covariate streams, likely attributable to sensor dropout and irregular recording practices in the original study protocol. These gaps propagate noise into the sliding window training procedure and distort the distributional evaluation at the 12 AM anchor point. Addressing these data quality issues through imputation strategies or gap-aware training objectives is an important direction for future work before this dataset can be used for fair comparative evaluation.

Probabilistic forecasting as a more appropriate clinical paradigm. While our current evaluation framework uses point forecasts and RMSE as the primary metric, we argue that probabilistic forecasting is likely a more clinically appropriate and practically tractable formulation of this problem. As illus-

trated in Figure 2, probabilistic forecasts provide a distribution over future blood glucose trajectories, enabling uncertainty quantification that is essential for clinical decision support. Given the extreme volatility of blood glucose, the prevalence of untracked covariates such as stress and unlogged meals, and frequent sensor noise, a single point estimate is unlikely to be sufficient for safe nocturnal hypoglycemia prevention in practice. Probabilistic forecasts also relax the requirement for morphologically perfect trajectory reconstruction; instead of requiring the model to predict the exact shape of a volatile signal, the model need only assign sufficient probability mass to the hypoglycemic region of the forecast distribution, which is a substantially less demanding and more clinically meaningful objective.

IV. Conclusion

This study presents the first comprehensive benchmarking of time series foundation models for long-horizon nocturnal hypoglycemia forecasting in type 1 diabetes. Evaluating six architectures across four public CGM datasets comprising over 50 million readings, we demonstrate that attention-based TSFMs—particularly `TimesFM`—consistently outperform MLP-based counterparts in both zero-shot and fine-tuned settings. Our results reveal that MLP-based architectures suffer from mean-reversion behaviour that, while minimizing RMSE, fails to capture the clinically critical morphological structure of blood glucose trajectories. Fine-tuning yields consistent performance gains across most model-dataset combinations, though we note that intra-patient temporal distribution shift—arising from the months-long gap between fine-tuning and evaluation windows—may limit the utility of patient-specific adaptation in practice, as physiological drift can cause the population prior to generalize better than stale patient-level models.

The gap between current model performance and clinical utility remains substantial; RMSE values of 2.5–4.0 mmol/L over 8-hour horizons indicate that point forecasting alone is unlikely to provide sufficiently reliable predictions for safe nocturnal hypoglycemia prevention. We believe that probabilistic forecasting, shape-aware loss functions, and larger-scale training datasets represent the most promising paths toward clinically deployable systems. This benchmark establishes a reproducible evaluation framework and provides actionable guidance for future TSFM development in this challenging physiological forecasting domain.

A. Future Work

Several directions remain open for extending this benchmarking study. First, we intend to incorporate probabilistic forecasting adapters to move beyond point estimates, enabling uncertainty quantification that is critical for clinical decision support in nocturnal hypoglycemia prevention. Second, our evaluation will be extended to the Gluroo dataset, comprising over 100,000 patients and 10 billion CGM readings, which would enable training TSFMs from scratch rather than fine-tuning, while also providing more accurate sleep/wake time anchoring and improved recording of hypoglycemia corrections for fairer distributional evaluation. Third, a more rigorous data ablation study is warranted to assess the marginal contribution of covariates such as IOB/COB, activity levels, and sleep/wake times on forecast accuracy.

Beyond data, several model-level improvements are planned. We will expand the benchmark to include additional TSFM architectures and evaluate models on clinically motivated loss metrics that penalize trajectory shape error rather than mean squared error alone. Horizon-specific models targeting 2, 4, 6, and 8-hour forecasting windows will be developed and compared against the current unified 8-hour forecaster, with ensemble combination strategies explored. Subgroup analyses stratified by patient demographics (e.g., insulin delivery method, age, sex), insulin sensitivity factors, carbohydrate ratios, and time-of-night will be conducted to assess model equity and robustness across clinically relevant populations. Finally, calibration analysis and Clarke Error Grid Analysis [34] will be incorporated to evaluate model confidence and clinical safety in alignment with established diabetes care standards.

V. Acknowledgements

We disclose that authors Christopher Risi and Tony Chan were employed by *Gluroo Imaginations Incorporated* during the writing of this paper. We would like to thank both of Gluroo's Walker Payne, Data Scientist and Greg Badros, CEO for the insights feedback they provided into formulating this problem. We would also like to thank Franz Kiraly and the *sktime* community for their guidance in design this problem. We would also like to thank Professor Jesse Hoey, *University of Waterloo* for his guidance on this project. We were also provided extensive funding for our GPUs from the *Waterloo.AI Data and Artificial Intelligence Institute*. Finally we must thank the incredible volunteers that helped build the student-run *WAT.ai* organization, without them we would never have been able to build our team of talented undergraduate researchers.

REFERENCES

[1] E. A. Davis, B. Keating, G. C. Byrne, M. Russell, and T. W. Jones, "Hypoglycemia: incidence and clinical predictors in a large population-based sample of children and adolescents with IDDM," *Diabetes Care*, vol. 20, no. 1, pp. 22–25, Jan. 1997.

[2] B. Buckingham, D. M. Wilson, T. Lecher, R. Hanas, K. Kaiserman, and F. Cameron, "Duration of nocturnal hypoglycemia before seizures," *Diabetes Care*, vol. 31, no. 11, pp. 2110–2112, Nov. 2008.

[3] D. Koltin and D. Daneman, "'Dead-in-bed' syndrome - a diabetes nightmare," *Pediatric Diabetes*, vol. 9, no. 5, pp. 504–507, 2008, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1399-5448.2008.00404.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1399-5448.2008.00404.x

[4] K. Barnard, S. Thomas, P. Royle, K. Noyes, and N. Waugh, "Fear of hypoglycaemia in parents of young children with type 1 diabetes: a systematic review," *BMC pediatrics*, vol. 10, p. 50, Jul. 2010.

[5] K. Barnard, J. James, D. Kerr, P. Adolfsson, A. Runion, and G. Serbedzija, "Impact of Chronic Sleep Disturbance for People Living With T1 Diabetes," *Journal of Diabetes Science and Technology*, vol. 10, no. 3, pp. 762–767, May 2016. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1932296815619181

[6] D. Ehrmann, L. Laviola, L.-S. Priesterroth, N. Hermanns, N. Babion, and T. Glatzer, "Fear of Hypoglycemia and Diabetes Distress: Expected Reduction by Glucose Prediction," *Journal of Diabetes Science and Technology*, vol. 18, no. 5, pp. 1027–1034, Sep. 2024. [Online]. Available: https://doi.org/10.1177/19322968241267886

[7] K. M. Perez, E. R. Hamburger, M. Lyttle, R. Williams, E. Bergner, S. Kahanda, E. Cobry, and S. S. Jaser, "Sleep in Type 1 Diabetes: Implications for Glycemic Control and Diabetes Management," *Current Diabetes Reports*, vol. 18, no. 2, p. 5, Feb. 2018. [Online]. Available: http://link.springer.com/10.1007/s11892-018-0974-8

[8] G. Aleppo, K. J. Ruedy, T. D. Riddlesworth, D. F. Kruger, A. L. Peters, I. Hirsch, R. M. Bergenstal, E. Toschi, A. J. Ahmann, V. N. Shah, M. R. Rickels, B. W. Bode, A. Philis-Tsimikas, R. Pop-Busui, H. Rodriguez, E. Eyth, A. Bhargava, C. Kollman, R. W. Beck, and for the REPLACE-BG Study Group, "REPLACE-BG: A Randomized Trial Comparing Continuous Glucose Monitoring With and Without Routine Blood Glucose Monitoring in Adults With Well-Controlled Type 1 Diabetes," *Diabetes Care*, vol. 40, no. 4, pp. 538–545, Feb. 2017. [Online]. Available: https://doi.org/10.2337/dc16-2482

[9] P. Trirat, Y. Shin, J. Kang, Y. Nam, J. Na, M. Bae, J. Kim, B. Kim, and J.-G. Lee, "Universal Time-Series Representation Learning: A Survey," Aug. 2024, arXiv:2401.03717 [cs]. [Online]. Available: http://arxiv.org/abs/2401.03717

[10] Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen, "Foundation Models for Time Series Analysis: A Tutorial and Survey," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6555–6565. [Online]. Available: https://dl.acm.org/doi/10.1145/3637528.3671451

[11] S. R. K. Kottapalli, K. Hubli, S. Chandrashekhara, G. Jain, S. Hubli, G. Botla, and R. Doddaiah, "Foundation Models for Time Series: A Survey," Apr. 2025, arXiv:2504.04011 [cs]. [Online]. Available: http://arxiv.org/abs/2504.04011

[12] Adam Brown, "42 Factors That Affect Blood Glucose?! A Surprising Update," Feb. 2018. [Online]. Available: https://diatribe.org/diabetes-management/42-factors-affect-blood-glucose-surprising-update

[13] P. Liu, B. Wu, Y. Hu, N. Li, T. Dai, J. Bao, and S.-t. Xia, "TimeBridge: Non-Stationarity Matters for Long-term Time Series Forecasting," Oct. 2024, arXiv:2410.04442 [cs] version: 1. [Online]. Available: http://arxiv.org/abs/2410.04442

[14] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 11 121–11 128, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26317

[15] S. A. Brown, B. P. Kovatchev, D. Raghinaru, J. W. Lum, B. A. Buckingham, Y. C. Kudva, L. M. Laffel, C. J. Levy, J. E. Pinsker, R. P. Wadwa, E. Dassau, F. J. Doyle, S. M. Anderson, M. M. Church, V. Dadlani, L. Ekhlaspour, G. P. Forlenza, E. Isganaitis, D. W. Lam, C. Kollman, and R. W. Beck, "Six-Month Randomized, Multicenter Trial of Closed-Loop Control in Type 1 Diabetes," *New England Journal of Medicine*, vol. 381, no. 18, pp. 1707–1717, Oct. 2019, _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1907863. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1907863

[16] J. Lynch, L. G. Kanapka, S. J. Russell, E. R. Damiano, F. H. El-Khatib, K. J. Ruedy, C. Balliro, P. Calhoun, and R. W. Beck, "The Insulin-Only Bionic Pancreas Pivotal Trial Extension Study: A Multi-Center Single-Arm Evaluation of the Insulin-Only Configuration of the Bionic Pancreas in Adults and Youth with Type 1 Diabetes," *Diabetes Technology & Therapeutics*, vol. 24, no. 10, pp. 726–736, Oct. 2022. [Online]. Available: https://journals.sagepub.com/doi/full/10.1089/dia.2022.0341

[17] The Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, "Continuous Glucose Monitoring and Intensive Treatment of Type 1 Diabetes," *New England Journal of Medicine*, vol. 359, no. 14, pp. 1464–1476, Oct. 2008. [Online]. Available: http://www.nejm.org/doi/abs/10.1056/NEJMoa0805017

[18] S. Rancati, P. Bosoni, R. Schiaffini, A. Deodati, P. A. Mongini, L. Sacchi, C. Toffanin, and R. Bellazzi, "Exploration of Foundational Models for Blood Glucose Forecasting in Type-1 Diabetes Pediatric Patients," *Diabetology*, vol. 5, no. 6, pp. 584–599, Nov. 2024, company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2673-4540/5/6/42

[19] J. Luo, A. Kumbara, M. Shomali, R. Han, A. Iyer, G. Aleppo, R. Agarwal, and G. Gao, "A large sensor foundation model pretrained on continuous glucose monitor data for diabetes management," *npj Health Systems*, vol. 2, no. 1, p. 35, Sep. 2025. [Online]. Available: https://www.nature.com/articles/s44401-025-00039-y

[20] G. Lutsker, G. Sapir, S. Shilo, J. Merino, A. Godneva, J. R. Greenfield, D. Samocha-Bonet, R. Dhir, F. Gude, S. Mannor, E. Meirom, G. Chechik, H. Rossman, and E. Segal, "GluFormer: Learning Generalizable Representations from Continuous Glucose Monitoring Data," Jul. 2025. [Online]. Available: https://openreview.net/forum?id=yNoiaSVL3y

[21] A. Mahmoudi, G. Farahani, P. Domanski, B. Farahani, F. Firouzi, and K. Chakrabarty, "DiabLLM: An LLM-Based Framework for Blood Glucose Prediction in Type 1 Diabetes," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/11373821

[22] A. Garza, C. Challu, and M. Mergenthaler-Canseco, "TimeGPT-1," May 2024, arXiv:2310.03589 [cs]. [Online]. Available: http://arxiv.org/abs/2310.03589

[23] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu, "Long-term Forecasting with TiDE: Time-series Dense Encoder," Apr. 2024, arXiv:2304.08424 [stat]. [Online]. Available: http://arxiv.org/abs/2304.08424

[24] S.-A. Chen, C.-L. Li, S. O. Arik, N. C. Yoder, and T. Pfister, "TSMixer: An All-MLP Architecture for Time Series Forecast-ing," *Transactions on Machine Learning Research*, Apr. 2023. [Online]. Available: https://openreview.net/forum?id=wbpxTuXgm0

[25] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, Jul. 2024, pp. 10 148–10 167. [Online]. Available: https://proceedings.mlr.press/v235/das24c.html

[26] X. Xu, N. Kok, J. Tan, M. Martin, D. Buchanan, E. Chun, R. Bhat, S. Cass, E. Wang, S. Senthil, and I. Gaynanova, "IrinaStatsLab/Awesome-CGM: Updated release with additional public CGM dataset and enhanced processing," Dec. 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.14541646

[27] V. Ekambaram, A. Jati, P. Dayama, S. Mukherjee, N. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam, "Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series," *Advances in Neural Information Processing Systems*, vol. 37, pp. 74 147–74 181, Dec. 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/874a4d89f2d04b4bcf9a2c19545cf040-Abstract-Conference.html

[28] A. F. Ansari, O. Shchur, J. Küken, A. Auer, B. Han, P. Mercado, S. S. Rangapuram, H. Shen, L. Stella, X. Zhang, M. Goswami, S. Kapoor, D. C. Maddix, P. Guerron, T. Hu, J. Yin, N. Erickson, P. M. Desai, H. Wang, H. Rangwala, G. Karypis, Y. Wang, and M. Bohlke-Schneider, "Chronos-2: From Univariate to Universal

Forecasting," Oct. 2025, arXiv:2510.15821 [cs]. [Online]. Available: http://arxiv.org/abs/2510.15821

[29] Y. Liu, G. Qin, Z. Shi, Z. Chen, C. Yang, X. Huang, J. Wang, and M. Long, "Sundial: A Family of Highly Capable Time Series Foundation Models," Nov. 2025, arXiv:2502.00816 [cs]. [Online]. Available: http://arxiv.org/abs/2502.00816

[30] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting," Feb. 2021, arXiv:2101.12072 [cs]. [Online]. Available: http://arxiv.org/abs/2101.12072

[31] M. Cuturi and M. Blondel, "Soft-DTW: a Differentiable Loss Function for Time-Series," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, Aug. 2017, pp. 894–903. [Online]. Available: https://proceedings.mlr.press/v70/cuturi17a.html

[32] V. LE GUEN and N. THOME, "Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/466accbac9a66b805ba50e42ad715740-Paper.pdf

[33] H. Alt and M. Godau, "COMPUTING THE FRÉCHET DISTANCE BETWEEN TWO POLYGONAL CURVES," *International Journal of Computational Geometry & Applications*, vol. 05, no. 01n02, pp. 75–91, Mar. 1995. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218195995000064

[34] W. L. Clarke, "The Original Clarke Error Grid Analysis (EGA)," *Diabetes Technology & Therapeutics*, vol. 7, no. 5, pp. 776–779, Oct. 2005. [Online]. Available: https://journals.sagepub.com/doi/full/10.1089/dia.2005.7.776