IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Trustworthy Glucose Forecasting with Physiology-Constrained Neural Dynamics, Conformal Risk Control, and Risk-Sensitive Reinforcement Learning

SARMAD MAQSOOD[1], MUHAMMAD ABDULLAH SARWAR[2], (MEMBER, IEEE), EGLE BELOUSOVIENĖ[3], AND RYTIS MASKELIŪNAS[1], (Member, IEEE)

[1]Department of Applied Informatics, Vytautas Magnus University, 44404 Kaunas, Lithuania (e-mail: sarmad.maqsood@vdu.lt, rytis.maskeliunas@vdu.lt)
[2]Centre of Real Time Computer Systems, Faculty of Informatics, Kaunas University of Technology, LT-51386 Kaunas, Lithuania (e-mail: m.sarwar@ktu.edu)
[3]Department of Intensive Care, University of Health Sciences, Kaunas, Lithuania (e-mail: egle.belousoviene@lsmu.lt)

Corresponding author: Sarmad Maqsood (e-mail: sarmad.maqsood@vdu.lt).

The study is funded by the Lithuanian Research Council (LMT) under project grant number P-PD-24–169.

ABSTRACT Reliable short-horizon continuous glucose monitoring (CGM) forecasting is essential to prevent hypoglycemia, yet most data-driven models provide neither physiological awareness nor actionable safety guaranties. We introduce a trustworthy forecasting framework that unifies continuous-time neural dynamics, calibration-based conformal risk control (empirically validated in held-out subjects), and tail-risk-aware adaptation. First, a Neural Controlled Differential Equation (Neural CDE) encoder models irregular, multi-channel streams (CGM, insulin, carbohydrates, activity, time-of-day). A physiology regularizer anchors latent dynamics to plausible glucose responses (e.g., post-prandial rise, insulin-mediated decay) and penalizes simulator-inconsistent trends. Second, Conformal Risk Control calibrates prediction sets and provides an actionable deferral policy when uncertainty or low-glucose risk is high, achieving prediction sets calibrated to a nominal coverage target and empirically validated under subject-disjoint evaluation. Third, a risk-sensitive policy (CVaR-optimized PPO) adapts predictions to minimize tail errors, explicitly weighting under-predictions near hypoglycemia. We evaluated in BrisT1D and OhioT1DM with patient-wise splits and performed in-silico stress tests in the UVA/Padova simulator (meal/bolus perturbations, basal overrides, sensor noise). Across 30/60-minutes horizons, our method achieves MAE=3.0–3.9 mg/dL, RMSE=8.5–9.8 mg/dL, and MARD=2.1–2.7%, outperforming strong baselines by 10–14 mg/dL RMSE and 10–13 mg/dL MAE on average. The risk control is calibrated at a nominal target of 90%, achieving 90.5 to 91.5% coverage with a average set width of 18-20 mg/dL and 5% to 6% deferral. Clinically, the rates of the SEG/Parkes A/B-zone increase by +10–14 pp, and event-centric analyzes show PR-AUC=0.92–0.95 and ROC-AUC=0.94–0.96 with earlier warnings (9–11 min median lead-time) and a 35–45% reduction in hypoglycemia under-prediction near the threshold. To our knowledge, this is the first end-to-end framework combining physiology-constrained Neural CDEs, conformal risk control with deferral, and CVaR-PPO adaptation for CGM forecasting, delivering verifiable safety properties alongside state-of-the-art accuracy.

INDEX TERMS CGM forecasting, neural controlled differential equations, conformal prediction, CVaR reinforcement learning, hypoglycemia risk, physiological regularization.

I. INTRODUCTION

Short-horizon forecasting of interstitial glucose from continuous glucose monitoring (CGM) streams supports early detection and prevention of hypoglycemia, a leading cause of acute morbidity that can cause confusion, seizure, loss of consciousness, arrhythmia, and even death if not prevented [1]. Clinical guidance emphasizes time-in-range (TIR) targets with explicit definitions for Level 1 (<70 mg/dL; <3.9 mmol/L) and Level 2 (<54 mg/dL; <3.0 mmol/L) hypoglycemia, underscoring that prediction quality must be judged by clinical safety, not only numerical accuracy [2].

In particular, point metrics such as the root mean square error (RMSE) can mask asymmetric harm, a modest average error that is systematically underestimated in the low-glucose regime, and is clinically dangerous. Established clinical grids, i.e., Clarke [3], Parkes/Consensus [4], and the Surveillance Error Grid (SEG) [5], were developed to quantify patient risk arising from prediction or measurement errors across the glycemic spectrum, and recent updates advocate modernized grid usage in surveillance and evaluation [6], [7], all motivating safety-aware evaluation for forecasting models beyond point accuracy.

Despite rapid progress in deep learning (DL) for time series, three persistent gaps limit reliable CGM forecasting. First, the intrinsic irregularity and multimodality of diabetes data fundamentally challenge discrete-time sequence models. Standard discrete-time models i.e., recurrent neural networks (RNNs), temporal convolutional networks (TCNs) and transformer architectures implicitly assume uniform sampling and synchronized channels, whereas real-world streams are irregular and multimodal. CGM is quasi-periodic (e.g. every 5 minutes), but insulin boluses and basal adjustments, carbohydrate intake, physical activity, and sensor dropouts occur at irregular times and are frequently missing. Naive alignment via forward-filling or uniform resampling can introduce aliasing, temporal distortion, and leakage of event timing, biasing downstream learners. Continuous-time neural dynamics addresses these pathologies by integrating between observations and explicitly conditioning on event times. Representative approaches include GRU-D (time-aware decay/imputation) [8], Latent ODE/ODE-RNN [9], GRU-ODE-Bayes [10], and neural controlled differential equations (Neural CDE) [11], all expressly designed for partially observed irregular streams.

Second, uncertainty is frequently reported but rarely guaranteed. Many medical machine learning (ML) works compute uncertainty proxies (e.g., Monte Carlo (MC) dropout or predictive variance), yet few provide formal, distribution-free conformal calibration (empirical coverage on held-out subjects) at deployment [12], [13]. Modern neural networks are also known to be miscalibrated and often over-confident outside distributions [14]; Bayesian approximations, such as MC dropout, offer useful signals but do not provide coverage guaranties [15]. Conformal prediction addresses calibration by constructing finite-sample-valid prediction sets [16] and has been extended from coverage control to risk control first via risk control prediction sets (RCPS) [17], and more recently via conformal risk control (CRC) for monotone losses [13], with adaptations to time-series forecasting [18]. These methods enable calibrated uncertainty with explicit performance targets and naturally pair with rejection/deferral policies in safety-critical settings.

Third, risk-aware adaptation remains underused for prediction safety. Reinforcement learning (RL) has been studied extensively for insulin dosing and closed-loop artificial pancreas control, both in silico (UVA/Padova simulator) and with observational data [19], [20], but it is far less developed for forecasting safety. Contemporary forecasting frameworks rarely couple adaptation to tail risk (e.g., hypoglycemia underprediction) with explicit safety criteria. Bridging RL's risk-sensitive optimization, for example, conditional value-at-risk (CVaR) [21], [22] with predictive modeling offers a principled avenue to penalize clinically hazardous errors without sacrificing overall accuracy, while on-policy algorithms such as proximal policy optimization (PPO) provide stable training for these objectives [23].

In this work, we propose a trustworthy CGM forecasting framework that directly addresses these gaps:

1) We introduce a physiology-constrained neural CDE forecaster in which a continuous-time encoder handles irregular multimodal inputs and a physiology regularizer anchors latent dynamics to plausible postprandial rise and insulin-mediated decay patterns, based on established simulator dynamics.
2) We develop a CRC layer that converts uncertainty into calibrated prediction sets and an actionable deferral policy, targeting nominal coverage and controlling an application-specific monotone risk on held-out calibration data, and we evaluate calibration/risk under subject-disjoint testing and stress tests.
3) We couple the forecaster to a PPO-style adapter trained with a CVaR objective that explicitly up-weights tail errors, especially hypoglycemia under-prediction, thereby shaping safer behavior at comparable RMSE.

The remainder of this paper is organized as follows. Section II reviews related work on CGM forecasting, continuous-time neural dynamics, conformal prediction, and RL for glycemic control. Section III details the datasets and the prediction task. Section IV presents our methods: the neural CDE encoder with regularization of physiology, conformal risk control with deferral and CVaR-PPO adaptation. Section V defines metrics and evaluation protocols. Section VI reports the results, the interpretations and the interpretability analyzes. Section VII discusses implications and limitations and Section VIII concludes the paper and suggests directions for future research.

II. RELATED WORKS

A. CGM FORECASTING: TRADITIONAL, DEEP SEQUENCE MODELS, AND CONTINUOUS-TIME DYNAMICS

Early forecasting and control were grounded in physiology-based glucose–insulin models. The Bergman minimal model provided a foundation for quantifying insulin sensitivity [24], while compartmental models enabled model predictive control (MPC) strategies such as Hovorka for Type 1 diabetes [25] and validated in silico within the UVA/Padova simulator ecosystem [26]–[28]. These mechanistic models encode absorption, insulin action, and distribution dynamics, remain central to pre-clinical evaluation, and offer interpretable structure for stress testing and "what-if" analyzes.

As sensor/device data proliferated, data-driven forecasting rose to prominence. Traditional approaches, i.e., autoregres-

sive integrated moving average (ARIMA), support vector regression (SVR), gave way to deep sequence models, long short-term memory (LSTM), gated recurrent unit (GRU), TCN, and seq2seq architectures, trained in cohorts such as OhioT1DM [29] and related datasets [30]–[32]. Recent works employ attention mechanisms and transformers for multi-horizon prediction and variable selection, including the temporal fusion transformer (TFT) and Informer, which demonstrate strong performance in irregular and multivariate clinical time series [33], [34]. CGM-specific studies increasingly integrate multimodal inputs (CGM, insulin/meal logs, activity/wearables), compare univariate vs. multimodal forecasting, and explore personalization to account for inter-subject variability [12], [35]–[38].

A persistent challenge is that diabetes data are irregular and asynchronous: CGM is quasi-periodic, while exogenous drivers (meals, insulin, activity) are event driven and sporadically recorded. To address this, a line of continuous-time neural methods explicitly integrates between observations and conditions on event times. Representative techniques include GRU-D for time-aware decay/imputation [8], Latent ODE/ODE-RNN [9], GRU-ODE-Bayes [10], and Neural CDEs [11].

Recent studies explore hybrid mechanistic–neural strategies that combine the ODE-based structure with flexible neural components to provide inductive bias and improve robustness in glycemic modeling (e.g., Hybrid^2 Neural ODE for glycemic response) [39], [40]. Recent work also continues to push multimodal Transformer-based glucose forecasting (e.g., AttenGluco, integrating CGM and activity signals), highlighting ongoing interest in leveraging multimodal context for longer-horizon prediction [41].

B. UNCERTAINTY QUANTIFICATION, CALIBRATION, AND CONFORMAL PREDICTION IN HEALTHCARE

Modern neural networks are often miscalibrated, especially under a distribution shift; classic results document overconfidence and motivate post-hoc temperature scaling [14]. Popular uncertainty proxies include dropout of MC [15] and deep ensembles [42], but large-scale studies show that the quantification of heuristic uncertainty can degrade under change, reinforcing the need for calibrated uncertainty [43]. Conformal prediction provides finite-sample coverage under standard exchangeability assumptions in our temporally correlated CGM setting, we interpret this as empirical coverage calibrated on held-out data [16]. For regression and multi-horizon forecasting, Conformalized quantile regression (CQR) yields efficient, heteroscedastic intervals [44]. Crucially for safety, conformal methodology has been extended from coverage control to risk control: RCPS [17] and CRC [13] guaranty that a user-defined monotone loss (e.g., asymmetric penalties for hypoglycemia under-prediction) remains below a target level. Beyond marginal coverage, CRC provides a general framework to control the expected value of monotone risk functions under standard conformal assumptions [45]. For time series specifically, conformal time series methods adapt CP to autoregressive/multi-horizon settings while relaxing exchangeability assumptions [18]. Recent extensions further study CRC-style guarantees under non-exchangeable or dependent data, which is directly relevant to temporally correlated forecasting settings [46]. In healthcare, conformal approaches have been explored for triage, imaging, and clinical risk stratification, often paired with deferral (reject) options to reduce unsafe automation [47], [48].

C. RL IN DIABETES: CONTROL VERSUS PREDICTION, AND RISK SENSITIVITY

Within diabetes technology, RL has mainly focused on insulin dosing and closed-loop control (artificial pancreas), frequently using the UVA/Padova simulator for development and validation [26]–[28]. Recent systems span mealtime bolus advisors using actor–critic/deep RL [19] and basal controllers validated in silico [20], surveys discuss opportunities and safety challenges for RL in glucose control [49], [50]. In contrast, far fewer works address prediction safety, i.e., adapting forecasters to reduce tail risk (dangerous under-prediction near hypoglycemia thresholds). From the RL theory side, risk-sensitive criteria such as CVaR provide tools to emphasize tail outcomes [22], and distributional RL explicitly models the return distribution (e.g., C51, QR-DQN, IQN) [51]–[53]. Practically, PPO offers a stable on-policy backbone for constrained or risk-aware objectives. For deployment in healthcare, off-policy evaluation (OPE) and safe policy improvement (e.g., doubly robust estimators, high-confidence OPE) are essential to reduce risks of real-world experimentation and align with personalization under shift [54], [55].

Summarizing, most CGM forecasters either (i) regularize/discretize irregular streams (risking aliasing and bias), (ii) report uncertainty without formal guarantees (leaving automation/deferral ad hoc), or (iii) personalize with risk-neutral objectives (improving the average while leaving the hypoglycemia tail exposed). Our framework advances the state-of-the-art by integrating three complementary pillars: continuous-time Neural CDE to natively handle irregular, multimodal inputs, conformal risk control to translate uncertainty into calibrated prediction sets (with empirical coverage near the nominal target) and an actionable deferral layer, and CVaR-driven PPO adaptation that explicitly targets tail errors near clinical thresholds while preserving overall accuracy.

III. DATASET & PROBLEM SETUP

Three datasets were used for the experimental investigation:

- BrisT1D consists of United Kingdom (UK) cohort of 24 young adults (YA) with Type 1 diabetes, providing device logs (CGM, insulin, carbohydrate entries) and smartwatch signals (heart rate, steps, distance, calories) aggregated at 5 minutes resolution, intended for forecasting benchmarks with patient-wise evaluation [56].
- OhioT1DM is a multimodal cohort of 12 adults, 8 weeks per subject, including CGM, insulin pump data, phys-

iological/wearable sensors, and self-reported events; widely used in blood-glucose prediction research [29].
- For controlled in silico stress tests, we use the UVA/Padova Type 1 Diabetes simulator to generate virtual subjects and standardized perturbations (meal/bolus variations, basal changes, sensor noise). Specifically, we use the simulator's standard virtual cohorts (adult/adolescent/child) at 5-minute sampling with sensor noise enabled, and construct scenario sets via controlled perturbations of meal, bolus, and basal conditions. Earlier versions were accepted by the U.S. FDA (2008) for certain preclinical evaluations; subsequent updates extend scenarios and validation [26], [28].

Table 1 summarizes the datasets.

We formulate supervised forecasting with a 6-hour context and two horizons: 30 and 60 minutes ahead. For each window ending at time t, inputs comprise CGM, insulin (bolus impulses; integrated basal rates), carbohydrate impulses, and available wearable features. The targets are interstitial glucose (mg/dL) at $t{+}30$ and $t{+}60$ minutes, convertible to mmol/L via $\mathrm{mmol/L} = 0.0555 \times \mathrm{mg/dL}$. Windows slide with a 5 minutes stride.

Following ADA standards of care, we define Level-1 hypoglycemia as < 70 mg/dL (3.9 mmol/L), Level-2 as < 54 mg/dL (3.0 mmol/L), and hyperglycemia as > 180 mg/dL (12.0 mmol/L). An impending event within $[t, t{+}\tau]$ is flagged if the CGM trajectory crosses the threshold, enabling lead-time, sensitivity, and clinical error-grid analyses reported later.

We use subject-wise partitioning to prevent identity and overlap-based leakage, all windows from a given subject are assigned to a single split. This yields 16/4/4 (train/validation/test) subjects for BrisT1D and 8/2/2 subjects per fold for OhioT1DM (patient-wise cross-validation). Continuous variables are standardized using statistics computed only from training subjects (per fold). The same normalization parameters are then applied unchanged to the validation and test subjects. It is used to avoid any use of held-out test-subject information during preprocessing. Imputation for extended periods of missing data is deliberately avoided; however, CGM data can be forward filled for up to 15 minutes, with longer intervals remaining unfilled. Event streams, including meals and boluses, are neither imputed nor modified; instead, they are supplemented with missingness indicators and features that track the time elapsed since the last recorded observation. Hyperparameter optimization and subsequent calibrations, such as determining uncertainty thresholds, are conducted exclusively on training participants, while test participants are entirely excluded from these processes.

IV. PROPOSED METHODOLOGY

We present a unified framework for safe short-horizon CGM forecasting that integrates four components in a sequential, yet modular framework. First, a continuous-time encoder based on Neural CDE consumes irregular, multimodal inputs $X(t)$ (CGM, insulin, carbohydrates, wearables, masks, Δt, time-of-day) and produces a latent state h_t. Second, a physiology-aware neural differential twin regularizer shapes the latent dynamics to align with simulator-consistent post-bolus decay and post-prandial rise patterns, improving plausibility and robustness. Third, a conformal layer converts point/quantile forecasts into calibrated prediction sets and triggers an actionable DEFER policy when intervals imply potential hypoglycemia or excessive uncertainty, yielding distribution-free risk control. Finally, a lightweight risk-sensitive adapter trained with CVaR-PPO applies small corrections to reduce hypoglycemia-weighted tail error without degrading average accuracy. Training proceeds in two stages (supervised pretraining with physiology regularization, then CVaR-PPO adaptation), followed by conformal calibration on held-out training subjects; inference produces both point forecasts and calibrated sets with a deferral signal. A schematic diagram of the proposed model is presented in Figure 1.

A. DATA REPRESENTATION AND INPUT CONSTRUCTION

We construct a continuous-time control path $X(t) \in \mathbb{R}^d$ over each 6-hour context window $[t_0, t]$ by aligning all signals to a 5 minutes canonical grid while preserving irregularity via binary masks and time-since-last-observation features. Carbohydrate entries and bolus insulin are encoded as impulses in the containing bin:

$$\begin{aligned} u_{\mathrm{CHO}}(t_k) &= \sum_{j:\, t_j \in [t_k-\Delta,\, t_k)} \mathrm{CHO}_j, \\ u_{\mathrm{bolus}}(t_k) &= \sum_{j:\, t_j \in [t_k-\Delta,\, t_k)} \mathrm{dose}_j, \end{aligned} \tag{1}$$

where $\Delta = 5$ minutes. Basal insulin is represented as the average infusion rate over the interval:

$$u_{\mathrm{basal}}(t_k) = \frac{1}{\Delta} \int_{t_k-\Delta}^{t_k} r_{\mathrm{basal}}(s)\, \mathrm{d}s. \tag{2}$$

For each channel c, we append a binary observation mask $m_c(t_k) \in \{0, 1\}$ and a time-since-last-observation feature $\Delta t_c(t_k)$ defined recursively by:

$$\Delta t_c(t_k) = \begin{cases} 0, & m_c(t_k) = 1, \\ \Delta t_c(t_{k-1}) + \Delta, & m_c(t_k) = 0\,. \end{cases} \tag{3}$$

Time-of-day is encoded through $\phi_{\mathrm{tod}}(t) = [\sin(2\pi h/24), \cos(2\pi h/2$ with hour h. To avoid information leakage, continuous variables are standardized per patient using statistics (μ_{pc}, σ_{pc}) computed from that patient's training windows only, and the same parameters are applied to that patient's validation and test windows as:

$$z_{pc}(t_k) = \frac{x_{pc}(t_k) - \mu_{pc}}{\sigma_{pc}}. \tag{4}$$

TABLE 1: Datasets used in this study (patient-wise splits).

Dataset	Subjects	Duration	Sampling	Modalities (examples)	Split (Train/Val/Test)	Purpose
BrisT1D	24 (YA)	up to 6 months	5 minutes	CGM, insulin, carbs, HR, steps, distance, calories	16 / 4 / 4	Real-world evaluation
OhioT1DM	12 (adults)	8 weeks per subject	5 minutes	CGM, insulin (basal/bolus), carbs, HR, temp, GSR	8 / 2 / 2	Real-world evaluation
UVA/Padova	virtual (Adult/Adolescent/Child)	N/A	5 minutes	simulated CGM, insulin, meals, sensor noise	scenario sets	In silico stress

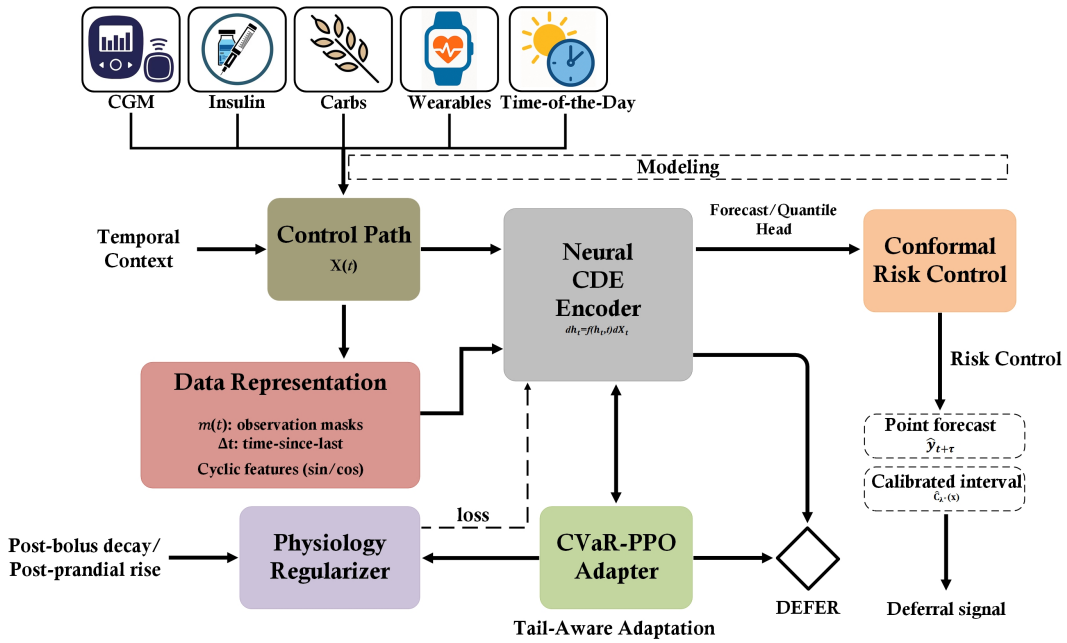


FIGURE 1: Architecture of the proposed physiology-aware, risk-controlled CGM forecasting framework.

The approach to handling missing data is intentionally conservative. CGM data are forward-filled for a duration of up to 15 minutes, longer gaps remain missing and are handled by $(m, \Delta t)$. Meanwhile, event streams, such as meals and boluses, are not subject to imputation.

B. CONTINUOUS-TIME ENCODER VIA NEURAL CDE

Let $X(t)$ denote the multi-channel control path comprising CGM, insulin (bolus and basal), carbohydrates, wearable features, masks, time-deltas, and time-of-day. We embed the context into a latent state $h_t \in \mathbb{R}^H$ through a Neural CDE [11] as:

$$\frac{\mathrm{d}h_t}{\mathrm{d}t} = f_\theta(h_t, t)\,\dot{X}_t \iff h_t = h_{t_0} + \int_{t_0}^{t} f_\theta(h_s, s)\,\mathrm{d}X_s, \tag{5}$$

where $f_\theta : \mathbb{R}^H \times \mathbb{R} \to \mathbb{R}^{H\times d}$ is a learnable vector field.

a: Interpolation of control channels

To construct a stable, bounded-variation control path, we distinguish continuous channels from event channels. Continuous signals (CGM, heart rate, steps, etc.) are interpolated with C^1 natural cubic splines through their observations. Event streams (bolus insulin, carbohydrate entries) are not spline-interpolated to avoid ringing around impulses. Instead, we embed them as right-continuous bounded-variation paths using either:

$$\tilde{u}(t) = \sum_j a_j\, t \geq t_j \qquad \text{(right-continuous step)}, \tag{6}$$

or their cumulative sum with piecewise-linear interpolation:

$$U(t) = \sum_{t_j \leq t} a_j, \qquad \tilde{u}(t) = \mathrm{PL}\left(\{(t_j, U(t_j))\}\right). \tag{7}$$

Both Eq. 6 and 7 preserve event magnitudes and timings while providing artifact-free bounded-variation controls for the CDE. In practice, we use Eq. 7 to obtain a C^0 control for these channels. The full control is:

$$X(t) = \big[\,\mathrm{Spline(CGM)}(t),\ \tilde{u}_{\mathrm{bolus}}(t),\ \tilde{u}_{\mathrm{CHO}}(t),\ \mathrm{Spline(wearables)}(t),\ m(t),\ \Delta t(t),\ \phi_{\mathrm{tod}}(t)\,\big]. \tag{8}$$

We solve Eq. 5 with an adjoint-enabled CDE integrator and initialize with a learned projection $h_{t_0} = \phi_\theta\big(X\,|_{[t_0,\,t_0+\Delta]}\big)$. Multi-horizon forecasts are produced by task-specific decoders:

$$\hat{y}_{t+\tau} = g_\tau(h_t), \qquad \tau \in \{30, 60\}\ \text{minutes}. \tag{9}$$

For probabilistic outputs we decode $(\mu_{t+\tau}, \log\sigma^2_{t+\tau})$ and optimize gaussian negative log-likelihood. For quantile outputs, we decode $\{q_\alpha(t+\tau)\}_{\alpha\in\mathcal{A}}$ (e.g., $\mathcal{A} = \{0.1, 0.9\}$) to support conformal post-calibration.

C. PHYSIOLOGY-AWARE REGULARIZATION (NEURAL DIFFERENTIAL TWIN)

To improve clinical plausibility and robustness under missingness, we softly constrain decoded glucose trajectories $\widehat{g}_s = d(h_s)$ to align with hallmark UVA/Padova responses. These terms are intended as soft, physiology-inspired shape priors rather than mechanistic physiological models, and thus may not hold under all free-living conditions. They are applied only when event timing is sufficiently reliable (i.e., within event-aligned segments); when real-world dynamics deviate (e.g., delayed absorption, stress/illness, exercise, or sensor noise), the supervised loss $\mathcal{L}_{\mathrm{sup}}$ dominates. Considering bolus time steps without concurrent carbohydrate intake (i.e., bolus >0 and carbs ≤ 0), we discourage post-bolus increases through a one-sided squared penalty:

$$\mathcal{L}_{\mathrm{decay}} = \frac{1}{|\mathcal{S}_{\mathrm{bolus}}|} \sum_{(s,s+\Delta)\in\mathcal{S}_{\mathrm{bolus}}} \Big[\max\big(0,\ \widehat{g}_{s+\Delta} - \widehat{g}_s\big)\Big]^2. \tag{10}$$

For isolated carbohydrate events, we capture a delayed rise and subsequent decay by enforcing early non-decreasing and late non-increasing behavior:

$$\mathcal{L}_{\mathrm{carb}} = \frac{1}{|\mathcal{S}_{\mathrm{early}}|} \sum_{(s,s+\Delta)\in\mathcal{S}_{\mathrm{early}}} \Big[\max\big(0,\ \widehat{g}_s - \widehat{g}_{s+\Delta}\big)\Big]^2 + \frac{1}{|\mathcal{S}_{\mathrm{late}}|} \sum_{(u,u+\Delta)\in\mathcal{S}_{\mathrm{late}}} \Big[\max\big(0,\ \widehat{g}_{u+\Delta} - \widehat{g}_u\big)\Big]^2. \tag{11}$$

Beyond shape constraints, we penalize discrepancy from simulator responses on matched events using summary functionals $\varphi(\cdot)$ (peak, time-to-peak, AUC ratio, half-time): for each event time step identified from the logs (meal or bolus), we extract an event-centered segment for computing the summary descriptors $\varphi(\cdot)$.

$$\mathcal{L}_{\mathrm{sim}} = \Big\|\varphi\big(\widehat{g}_{\mathrm{seg}}\big) - \varphi\big(g^{\mathrm{sim}}_{\mathrm{seg}}\big)\Big\|_2^2, \tag{12}$$

and promote smooth latent evolution via:

$$\mathcal{L}_{\mathrm{smooth}} = \sum_k \big\|h_{t_{k+1}} - h_{t_k}\big\|_2^2. \tag{13}$$

With $\mathcal{L}_{\mathrm{sup}}$ a primary forecasting loss (MAE or Gaussian NLL), the stage-1 objective is:

$$\mathcal{L}_{\mathrm{stage1}} = \mathcal{L}_{\mathrm{sup}} + \lambda_1\mathcal{L}_{\mathrm{decay}} + \lambda_2\mathcal{L}_{\mathrm{carb}} + \lambda_3\mathcal{L}_{\mathrm{sim}} + \lambda_4\mathcal{L}_{\mathrm{smooth}}, \tag{14}$$

where $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.1$.

D. CONFORMAL RISK CONTROL AND ACTIONABLE DEFERRAL

We calibrate prediction sets to target nominal coverage and calibrate a clinically meaningful risk criterion on held-out calibration data, and we derive an operational deferral rule. Using CQR [44], we first train quantile heads $q^{\pm}(x)$ and compute on a calibration set $\mathcal{C}$ the nonconformity as:

$$s_i = \max\{q^-(x_i) - y_i,\ y_i - q^+(x_i),\ 0\}, \quad (x_i, y_i) \in \mathcal{C}. \tag{15}$$

Let $\hat{k}$ be the $(1-\alpha)$-quantile of s_i. The marginally calibrated interval is:

$$\mathcal{C}(x) = [\,q^-(x) - \hat{k},\ q^+(x) + \hat{k}\,]. \tag{16}$$

To translate coverage into risk control, we define a monotone clinical loss in sets, for example, an indicator that the set intersects the hypoglycemic region or fails to cover a low measurement:

$$\begin{aligned} \ell\big(y, \mathcal{C}(x)\big) &= \{\mathcal{C}(x) \cap (-\infty, 70) \neq \emptyset\} \quad \text{or} \\ \ell &= \{y < 70\ \wedge\ \min\mathcal{C}(x) > y\}. \end{aligned} \tag{17}$$

We parameterize a family $\mathcal{C}_\lambda(x)$ by inflating Eq. 16 with factor $\lambda \geq 1$ and choose the smallest $\lambda^\star$ so that the empirical risk on calibration satisfies:

$$\frac{1}{|\mathcal{C}|} \sum_{(x_i,y_i)\in\mathcal{C}} \ell\big(y_i, \mathcal{C}_{\lambda^\star}(x_i)\big) \leq \alpha + \varepsilon_n. \tag{18}$$

This yields calibration-based control of the target risk on the held-out calibration set. Under standard conformal assumptions (exchangeability), CRC provides finite-sample marginal risk control in our temporally dependent CGM setting; we treat CRC as an empirical calibration mechanism and validate coverage/risk under subject-disjoint testing and stress tests. Classical conformal guaranties assume exchangeability, which is not strictly satisfied in CGM forecasting due to temporal dependence and overlapping sliding windows. Accordingly, we interpret CRC here as an empirical, distribution-robust calibration mechanism validated on held-out subjects (and under stress tests), rather than a strict finite-sample guaranty under full dependence. To reduce the shift, we can stratify the calibration (e.g., day vs. night, high-variance segments) and compute group-specific $\lambda^\star_g$ as in

conformal time series adaptations. At inference, we mark a deferral flag when the calibrated set indicates potential low-glucose risk or excessive uncertainty:

$$\mathsf{DEFER}(x) = \mathcal{C}_{\lambda^\star}(x) \cap (-\infty, 70) \neq \emptyset \;\text{ or }\; \big|\mathcal{C}_{\lambda^\star}(x)\big| > \tau. \tag{19}$$

with width threshold τ chosen from calibration (e.g., a high quantile of interval widths).

E. RISK-SENSITIVE ADAPTATION WITH CVAR-PPO

The CVaR-PPO adapter serves as a training-time risk-shaping mechanism that learns a fixed, static correction policy from historical trajectories. We use a lightweight correction policy at inference that outputs an additive adjustment a_t to the base prediction:

$$\tilde{y}_{t+\tau} = \hat{y}^{(0)}_{t+\tau} + a_t, \qquad a_t \sim \pi_\theta\big(\,\cdot \;\big|\; s_t\big). \tag{20}$$

To avoid any test-time information leakage, the CVaR-PPO correction policy is trained offline using only training subjects. During Stage-2 training, future outcomes $y_{t+\tau}$ are used only to compute the training loss and CVaR objective (i.e., as supervision on historical trajectories) in order to optimize π_θ. At inference, $y_{t+\tau}$ is unknown, the policy parameters are fixed and the adapter outputs a_t using only information available up to time t (i.e., s_t), with no online updates on validation or test subjects. Under this formulation, CVaR-PPO provides a training-time risk-shaping objective that learns a fixed correction policy from logged trajectories, explicitly optimizing tail-risk (CVaR) of trajectory-level prediction loss. During inference, the policy is frozen and applies a fixed correction using only test-time available signals (e.g., predicted trajectory/uncertainty and context), with no test-time adaptation or access to future glucose.

$$\ell_t = \sum_{\tau\in\{30,60\}} \big|y_{t+\tau} - \tilde{y}_{t+\tau}\big| \times \Big[1 + \omega_{\text{low}}\, \mathbf{1}\{y_{t+\tau} < 90 \;\wedge\; \tilde{y}_{t+\tau} < y_{t+\tau}\}\Big]. \tag{21}$$

and an episodic cost $Z = \sum_t \ell_t$. We target the tail by minimizing the CVaR at level $\beta \in (0, 1]$ via the Rockafellar–Uryasev surrogate [21], [22]:

$$\mathrm{CVaR}_\beta(Z) = \min_{\eta\in\mathbb{R}} \Big\{\, \eta + \tfrac{1}{\beta}\, \mathbb{E}\big[(Z-\eta)_+\big] \,\Big\}. \tag{22}$$

We implement CVaR-PPO by learning η jointly and re-weighting the advantages toward the trajectories with $Z > \eta$. With rewards $r_t = -\ell_t$ and CVaR-conditioned advantages $\hat{A}_t^{\text{CVaR}}$, the clipped PPO objective is defined as:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho_t(\theta)\,\hat{A}_t^{\text{CVaR}}, \; \text{clip}\,\big(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon\big)\,\hat{A}_t^{\text{CVaR}}\Big)\Big]. \tag{23}$$

$$\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}. \tag{24}$$

Stability is enhanced through an entropy bonus, a KL penalty with early stopping, small trust-region steps, and CVaR validation monitoring to prevent overfitting to extremes. The PPO is used as an offline risk-sensitive policy-optimization step for sequential forecast correction on logged trajectories, not as closed-loop control with online environment interaction. The objective targets the tail (CVaR) of an episode-level cumulative prediction loss, i.e., a trajectory-level risk criterion beyond stepwise regression. Pseudo-code for the offline CVaR-PPO training procedure is provided in Algorithm 1.

Algorithm 1: Offline CVaR-PPO Training for Forecast Correction.

Input: Trained base forecaster $\hat{y}^{(0)}$, training-subject trajectories $\mathcal{D}_{\text{train}}$, CVaR level β
Output: Frozen correction policy π_θ for inference
Initialize policy parameters θ, value parameters ψ, and CVaR threshold η;
foreach *training epoch* **do**
  Sample a minibatch of trajectories $\{\tau_i\}$ from $\mathcal{D}_{\text{train}}$;
  foreach *trajectory* τ **do**
    Roll out sequentially over time steps t;
      Sample correction $a_t \sim \pi_\theta(\cdot \mid s_t)$ and compute corrected forecast $\tilde{y}_{t+\tau} = \hat{y}^{(0)}_{t+\tau} + a_t$;
      Compute per-step loss ℓ_t using ground-truth outcomes $y_{t+\tau}$ (training only);
      Accumulate episodic cost $Z \leftarrow \sum_t \ell_t$;
  Update η using the Rockafellar–Uryasev CVaR surrogate at level β;
  Set rewards $r_t \leftarrow -\ell_t$;
  Compute advantages $\hat{A}_t$ using a value baseline V_ψ (GAE);
  Apply CVaR weighting by emphasizing updates from trajectories with $Z > \eta$;
  Update policy θ using PPO clipped objective;
  Update value parameters ψ by regression to returns;
return π_θ;

F. TRAINING PROTOCOL AND COMPUTATIONAL CONSIDERATIONS

Training continues in two stages. In stage 1, the CDE encoder and forecast heads are optimized using Eq. 14 with MAE or Gaussian NLL, patient-wise batching, gradient clipping, and early stopping on validation MAE. In stage 2, the encoder is either frozen or fine-tuned with a small learning rate while

the correction policy is trained with CVaR-PPO using Eq. 23. Hyperparameters ω_{low} and β are selected in validation subjects. Conformal calibration is performed on a held-out calibration fold comprising training subjects only. Stage-2 CVaR-PPO training is also conducted using training subjects only, and the learned policy is evaluated on validation/test subjects without any further updates. We estimate $\hat{k}$ for Eq. 15, select $\lambda^\star$ for Eq. 18 (optionally per stratum), and set τ for Eq. 19. Computationally, Neural CDE integration scales roughly with the number of spline knots N and hidden size H (empirically near $O(NH^2)$ for our settings). In our experiments, training Stage-1 takes approximately 3–5 hours on a single NVIDIA RTX 5090, while Stage-2 (CVaR-PPO) requires 1–2 hours. The inference latency per window is 14.2±1.8 ms. CRC adds a scalar line-search on calibration and CVaR-PPO introduces a modest adapter (tens of thousands of parameters) evaluated at the window cadence.

V. EXPERIMENTAL SETUP

A. EVALUATION METRICS

We quantify predictive accuracy using standard regression metrics used in blood glucose forecasting i.e., RMSE, MAE, Coefficient of Determination (R^2), and MARD. These metrics are defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{G}_i - G_i\right)^2}, \tag{25}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{G}_i - G_i\right|, \tag{26}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(G_i - \hat{G}_i\right)^2}{\sum_{i=1}^{N}\left(G_i - \bar{G}\right)^2}, \tag{27}$$

$$\text{MARD}(\%) = \frac{100}{N}\sum_{i=1}^{N}\frac{|\hat{G}_i - G_i|}{\max(G_i,\, \epsilon)}, \tag{28}$$

where $\hat{G}_i$ and G_i are predicted and the reference glucose is the mean reference, $\bar{G}$ is the mean reference, and ϵ=10 mg/dL to avoid division instability when G_i is very small. RMSE and MAE are in mg/dL while R^2 and MARD are unitless.

Alongside accuracy, we also summarize: (i) clinical error-grid zones (Clarke, Parkes/Consensus, SEG; % in Zones A/B), (ii) hypoglycemia event sensitivity, positive predictive value (PPV), and mean lead-time, (iii) calibrated uncertainty (coverage at target $1-\alpha$ and average interval width), (iv) tail behavior near lows (under-prediction rate in 60–90 mg/dL; CVaR of $|\hat{G} - G|$ for $G < 80$ mg/dL), and (v) robustness deltas under night-time segments, synthetic missingness, and simulator stress tests.

TABLE 2: Training and optimization configuration (summary).

Parameter	Value
Subject Splits	BrisT1D: 16/4/4; OhioT1DM: 5-fold (8/2/2)
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
Weight Decay	1×10^{-2}
LR Schedule	OneCycleLR (10% warm-up)
Batch Size	64
Max Epochs	100
Early Stopping	Patience = 10 (validation MAE)
Gradient Clipping	$\|g\| \leq 1.0$
Dropout	0.5 (encoder and heads)
CDE Hidden Size	$\{64, 96, 128\}$
CDE Layers	1–2
Quantile Alphas (CQR)	$\{0.10, 0.05\}$
Calibration	Training subjects only (held-out fold)
Deferral Threshold τ	90th percentile of calibrated widths
CVaR-PPO β	$\{0.05, 0.10, 0.20\}$
ω_{low} / ε / Entropy	$\{0.5, 1, 2\}$ / $\{0.1, 0.2\}$ / $[0, 0.01]$

B. TRAINING SETUP

All experiments were carried out on a Linux workstation with an Intel Core i9-14400 CPU (128 GB RAM) and a single NVIDIA GeForce RTX 5090, using Python/PyTorch with mixed precision stable. Evaluation is strictly subject-disjoint: BrisT1D uses a fixed 16/4/4 train/validation/test split; OhioT1DM uses a five-fold patient-wise cross-validation with 8/2/2 subjects per fold. The metrics are computed per subject and aggregated by median and interquartile range. To ensure a fair comparison, all baselines (CNN-LSTM, TCN, Transformer) were subject to grid-search hyperparameter optimization on the validation set, covering hidden dimensions, layer counts, and learning rates. All models were evaluated under the same subject-disjoint protocol.

The models are optimized with AdamW [57](initial learning rate 1×10^{-4}; weight decay 1×10^{-2}) and a OneCycleLR schedule with a warm-up 10%. We use a batch size of 64, up to 100 epochs, early stopping on validation MAE (patience 10), gradient clipping, and dropout of 0.5 in the encoder and task heads. Neural CDE hidden sizes are selected from {64, 96, 128} with 1–2 layers. Quantile heads at $\alpha \in \{0.10, 0.05\}$ support conformal post-calibration performed only on training subjects via a held-out calibration fold, test-time coverage is reported without refitting. The deferral rule is evaluated at inference by flagging forecasts whose calibrated interval intersects the low region (< 70 mg/dL) or exceeds a fixed width threshold τ (set to the 90th percentile of calibrated widths on the calibration fold). For the CVaR-PPO adapter, we consider $\beta \in \{0.05, 0.10, 0.20\}$, low-zone weight $\omega_{\text{low}} \in \{0.5, 1, 2\}$, PPO clipping $\varepsilon \in \{0.1, 0.2\}$, and a small entropy bonus in $[0, 0.01]$. Table 2 summarizes the key hyperparameters and the training configuration used throughout all experiments. All random seeds for data splits, initialization, and calibration folds were fixed and recorded to allow exact replication.

VI. RESULTS

This section reports forecasting accuracy, clinical safety, uncertainty calibration, tail-risk behavior, and robustness under temporal shift and missingness. Unless stated otherwise, metrics are aggregated per subject and summarized by median (IQR) across held-out participants for both 30- and 60-minute horizons.

A. MAIN OUTCOMES

Table 3 presents the principal point-accuracy results on OhioT1DM and BrisT1D. Across both datasets and horizons, the proposed Physio-CDE+CRC+CVaR-PPO method achieves the lowest error, improving upon strong sequence baselines (CNN-LSTM hybrid [58], TCN [59], Transformer [60]) and upon ablated variants that remove physiology regularization, conformal risk control, or CVaR-based adaptation. In particular, the model maintains substantially lower MAE and RMSE while achieving high R^2, indicating both reduced dispersion and improved explained variance relative to competing forecasters. Subject-level paired tests and effect sizes against the strongest baseline indicate consistent improvements for MAE and RMSE at both horizons, however, for BrisT1D, statistical significance is inherently limited by the small number of held-out test subjects (n=4), so we emphasize effect sizes and subject-level error distributions reported in Table 4.

Conformal calibration achieves coverage close to the nominal target (90/95%) without re-calibration in test subjects, with a slight conservative bias that is expected under finite-sample calibration. We report two sets of metrics: (i) overall performance computed on all forecast windows and (ii) selective performance computed only on non-deferred windows, always accompanied by the deferral rate. The deferral policy trades a small fraction of windows for safety: selective performance among non-deferred windows improves (lower MAE/RMSE, reduced low-zone errors) while overall coverage remains near target. Figure 2 contrasts headline accuracy and clinical safety with and without deferral on the 60-minute horizon. In both datasets, activation of calibrated deferral improves point accuracy for our method (lower MAE/RMSE) and increases the proportion of clinically benign errors (SEG A+B %), while the strongest baseline benefits much less. Naturally, this indicates a favorable safety–utility trade-off: the conformal layer routes a small fraction of high-risk windows to deferral, tightening error on the remaining (non-deferred) windows while maintaining overall performance on all windows. Complementing this, Table 5 reports uncertainty statistics at a 90% target coverage: our model achieves near-nominal coverage with smaller average set size (ASS), maintains selective coverage near target after deferral, and does so with modest deferral rates.

B. BLOCKED-TIME VALIDATION

To further assess generality beyond the standard windowed protocol, we evaluate a blocked-time split using contiguous held-out time blocks. For each subject, windows are ordered by time and the final contiguous block is held out for testing; metrics are computed per subject on this held-out block and summarized as mean $\pm$ SD in Table 6. We report overall point metrics (MAE and RMSE) over all test windows. Deferral is defined only for the CRC-enabled model, baselines without CRC/deferral have Deferral = 0 by construction.

C. SUBJECT-LEVEL VARIABILITY AND ROBUSTNESS

To assess whether the observed performance gains are consistent across individuals rather than driven by a small subset of subjects, Fig.3 reports per-subject MAE and RMSE for both datasets i.e., OhioT1DM and BrisT1D at 30-min and 60-min horizons, comparing the proposed model against the strongest ablation baseline (w/o CVaR-PPO). Across all settings, the proposed method achieves lower errors with reduced inter-subject variability, supporting the robustness of the reported improvements.

D. EVENT DETECTION AND LEAD-TIME

We assess hypoglycemia detection using ADA thresholds (Level-1: < 70 mg/dL; Level-2: < 54 mg/dL). An alarm is triggered when the calibrated prediction interval intersects the low-glucose region within the forecast horizon, and lead-time is measured as the minutes between the first valid alarm and the event onset. Compared with the strongest baseline (without CVaR-PPO), our method achieves consistently superior operating characteristics across alarm thresholds: PR-AUC ≈ 0.94 (OhioT1DM) and ≈ 0.93 (BrisT1D), alongside ROC-AUC ≈ 0.94–0.95, indicating earlier and more reliable warnings at comparable false-alarm rates illustrated in Fig.4). Event-wise summaries in Table 7 show higher sensitivity at similar or better PPV and meaningfully longer lead-times. At the 60-minute horizon, for example, Level-1 hypoglycemia sensitivity reaches ≈ 0.91–0.92 with PPV ≈ 0.75–0.79 and lead-times $\approx$ 18–19 min, while Level-2 achieves ≈ 0.84–0.87 sensitivity (Se) with PPV ≈ 0.72–0.77 and lead-times (TL) $\approx$ 15–16 min. The gains align with the calibrated-deferral policy's design: a small fraction of high-risk windows are flagged, tightening errors on non-deferred windows and yielding earlier, clinically actionable warnings without sacrificing overall coverage.

E. ROBUSTNESS AND SHIFT ANALYSES

We stress the models under three conditions: (i) temporal shift (day vs. night), (ii) synthetic context-window CGM missingness (10–30%), and (iii) in-silico perturbations in the UVA/Padova simulator (meal size/timing, bolus perturbations, basal overrides, and additive sensor noise). As an additional time-block check, we report performance separately for day vs. night segments (temporal shift), alongside synthetic missingness and simulator stress tests. The proposed method exhibits smaller degradation in MAE and maintains calibrated coverage more tightly than baselines and ablated variants. The conditional coverage plots in Fig. 5 show that the conformal layer maintains near-target coverage across glucose bands and day/night strata, with modest increases in

TABLE 3: Forecasting performance (per-subject mean ± SD) on OhioT1DM and BrisT1D. Best results are in **bold**.

Model	Dataset	Horizon	$R^2 \uparrow$	MAE (mg/dL)↓	RMSE (mg/dL)↓	MSE (mg/dL^2)↓	MARD (%)↓
CNN–LSTM hybrid	OhioT1DM	30 min	0.71 ± 0.04	18.1 ± 1.6	24.6 ± 1.7	605.2 ± 84.0	12.1 ± 1.2
Transformer (vanilla)	OhioT1DM	30 min	0.76 ± 0.03	16.0 ± 1.3	22.0 ± 1.5	484.0 ± 66.0	10.7 ± 1.0
TCN	OhioT1DM	30 min	0.77 ± 0.03	15.5 ± 1.2	21.4 ± 1.4	458.0 ± 60.0	10.3 ± 0.9
Neural CDE (no physiology)	OhioT1DM	30 min	0.85 ± 0.02	13.8 ± 1.1	19.6 ± 1.2	384.2 ± 47.0	9.2 ± 0.8
Ablated (w/o CRC/deferral)	OhioT1DM	30 min	0.84 ± 0.03	14.0 ± 1.1	19.9 ± 1.3	396.0 ± 51.0	9.4 ± 0.9
Ablated (w/o CVaR-PPO)	OhioT1DM	30 min	0.86 ± 0.02	13.5 ± 1.0	19.2 ± 1.2	368.6 ± 45.0	9.0 ± 0.8
Ours (Physio-CDE + CRC + CVaR-PPO) Method	OhioT1DM	30 min	$\mathbf{0.97 \pm 0.01}$	$\mathbf{3.0 \pm 0.3}$	$\mathbf{8.5 \pm 0.5}$	$\mathbf{72.3 \pm 8.5}$	$\mathbf{2.1 \pm 0.2}$
CNN–LSTM hybrid	OhioT1DM	60 min	0.64 ± 0.05	20.9 ± 1.8	28.4 ± 1.9	806.6 ± 108.0	14.0 ± 1.3
Transformer (vanilla)	OhioT1DM	60 min	0.69 ± 0.04	19.0 ± 1.5	26.8 ± 1.7	718.2 ± 95.0	12.8 ± 1.1
TCN	OhioT1DM	60 min	0.71 ± 0.04	18.4 ± 1.4	25.9 ± 1.6	670.8 ± 88.0	12.3 ± 1.1
Neural CDE (no physiology)	OhioT1DM	60 min	0.80 ± 0.03	16.2 ± 1.3	23.4 ± 1.4	547.6 ± 72.0	10.8 ± 1.0
Ablated (w/o CRC/deferral)	OhioT1DM	60 min	0.79 ± 0.03	16.5 ± 1.2	23.8 ± 1.5	566.4 ± 76.0	11.0 ± 1.0
Ablated (w/o CVaR-PPO)	OhioT1DM	60 min	0.81 ± 0.03	15.9 ± 1.2	23.0 ± 1.4	529.0 ± 69.0	10.5 ± 0.9
Ours (Physio-CDE + CRC + CVaR-PPO) Method	OhioT1DM	60 min	$\mathbf{0.95 \pm 0.01}$	$\mathbf{3.3 \pm 0.3}$	$\mathbf{9.1 \pm 0.6}$	$\mathbf{82.8 \pm 11.0}$	$\mathbf{2.3 \pm 0.2}$
CNN–LSTM hybrid	BrisT1D	30 min	0.64 ± 0.05	18.8 ± 1.6	26.2 ± 1.8	686.4 ± 92.0	12.7 ± 1.2
Transformer (vanilla)	BrisT1D	30 min	0.70 ± 0.04	17.0 ± 1.4	24.5 ± 1.6	600.3 ± 79.0	11.5 ± 1.1
TCN	BrisT1D	30 min	0.72 ± 0.03	16.2 ± 1.3	23.6 ± 1.5	557.0 ± 73.0	10.9 ± 1.0
Neural CDE (no physiology)	BrisT1D	30 min	0.81 ± 0.03	14.5 ± 1.1	20.6 ± 1.2	424.4 ± 53.0	9.8 ± 0.9
Ablated (w/o CRC/deferral)	BrisT1D	30 min	0.79 ± 0.04	14.9 ± 1.2	21.1 ± 1.3	445.2 ± 58.0	10.0 ± 0.9
Ablated (w/o CVaR-PPO)	BrisT1D	30 min	0.83 ± 0.03	14.0 ± 1.1	20.0 ± 1.2	400.0 ± 48.0	9.3 ± 0.8
Ours (Physio-CDE + CRC + CVaR-PPO) Method	BrisT1D	30 min	$\mathbf{0.95 \pm 0.01}$	$\mathbf{3.4 \pm 0.3}$	$\mathbf{8.7 \pm 0.4}$	$\mathbf{75.7 \pm 7.0}$	$\mathbf{2.4 \pm 0.2}$
CNN–LSTM hybrid	BrisT1D	60 min	0.57 ± 0.06	22.4 ± 1.9	30.7 ± 2.1	942.5 ± 128.0	15.1 ± 1.3
Transformer (vanilla)	BrisT1D	60 min	0.62 ± 0.04	20.2 ± 1.6	28.7 ± 2.0	823.7 ± 115.0	13.7 ± 1.2
TCN	BrisT1D	60 min	0.74 ± 0.03	18.6 ± 1.5	26.8 ± 1.9	718.2 ± 101.0	12.6 ± 1.1
Neural CDE (no physiology)	BrisT1D	60 min	0.76 ± 0.03	17.4 ± 1.3	24.7 ± 1.5	610.1 ± 82.0	11.7 ± 1.0
Ablated (w/o CRC/deferral)	BrisT1D	60 min	0.74 ± 0.03	17.9 ± 1.4	25.1 ± 1.6	630.0 ± 86.0	12.0 ± 1.0
Ablated (w/o CVaR-PPO)	BrisT1D	60 min	0.77 ± 0.03	16.8 ± 1.3	24.1 ± 1.5	580.8 ± 79.0	11.2 ± 0.9
Ours (Physio-CDE + CRC + CVaR-PPO) Method	BrisT1D	60 min	$\mathbf{0.93 \pm 0.01}$	$\mathbf{3.9 \pm 0.3}$	$\mathbf{9.8 \pm 0.5}$	$\mathbf{96.0 \pm 9.8}$	$\mathbf{2.7 \pm 0.2}$

TABLE 4: Subject-level comparison vs. strongest baseline (Ablated: w/o CVaR-PPO). Per-subject MAE/RMSE are computed on held-out subjects and paired differences are formed as Δ = baseline − ours (positive favors our method). Two-sided Wilcoxon signed-rank tests are applied across subjects; Holm-Bonferroni correction is performed across the two metrics (MAE, RMSE) within each dataset and horizon. BrisT1D has $n = 4$ test subjects.

Dataset	Horizon	Metric	Strongest Baseline	Mean Δ (baseline−ours)	Wilcoxon p (exact, Holm)
OhioT1DM	30 min	MAE (mg/dL)	Ablated (w/o CVaR-PPO)	10.5	1.2×10^{-5}
OhioT1DM	30 min	RMSE (mg/dL)	Ablated (w/o CVaR-PPO)	10.7	9.6×10^{-6}
OhioT1DM	60 min	MAE (mg/dL)	Ablated (w/o CVaR-PPO)	12.6	8.1×10^{-6}
OhioT1DM	60 min	RMSE (mg/dL)	Ablated (w/o CVaR-PPO)	13.9	6.7×10^{-6}
BrisT1D	30 min	MAE (mg/dL)	Ablated (w/o CVaR-PPO)	10.6	0.25
BrisT1D	30 min	RMSE (mg/dL)	Ablated (w/o CVaR-PPO)	11.3	0.25
BrisT1D	60 min	MAE (mg/dL)	Ablated (w/o CVaR-PPO)	12.9	0.25
BrisT1D	60 min	RMSE (mg/dL)	Ablated (w/o CVaR-PPO)	14.3	0.25

interval width during high-variance contexts. The aggregate robustness deltas (changes in MAE and coverage relative to unperturbed conditions) are reported in Table 8 and visualized in Fig. 6; simulator stress outcomes appear in Fig. 7 with

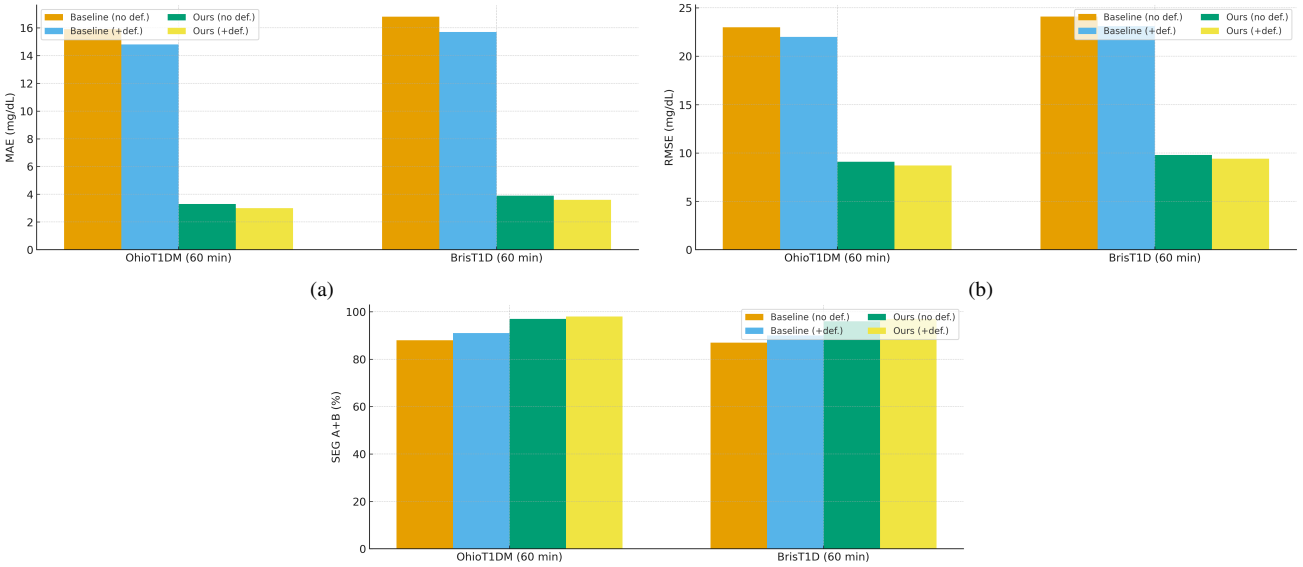


FIGURE 2: Accuracy and clinical safety with and without deferral (60-min horizon). Panels show (a) MAE, (b) RMSE, and (c) SEG A+B% for the strongest baseline and our method on OhioT1DM and BrisT1D dataset. Bars compare each model evaluated without deferral and with calibrated deferral.

TABLE 5: Coverage and deferral statistics at target 90% coverage. Per-subject mean ± SD. Coverage/ASS are computed on all windows; selective coverage/ASS are computed on non-deferred windows only.

Model	**Dataset**	**Horizon**	**Coverage (%)↑**	**Selective cov. (%)↑**	**ASS (mg/dL)↓**	**$\text{ASS}_{\text{nondef}}$ (mg/dL)↓**	**Deferral rate (%)↓**
Baseline (w/o CVaR-PPO)	OhioT1DM	30 min	88.9 ± 1.3	89.7 ± 1.2	21.8 ± 2.1	19.6 ± 1.9	6.2 ± 1.0
Ours (Physio-CDE + CRC + CVaR-PPO)	OhioT1DM	30 min	$\mathbf{91.0 \pm 0.9}$	$\mathbf{90.8 \pm 0.9}$	$\mathbf{17.9 \pm 1.7}$	$\mathbf{16.5 \pm 1.5}$	$\mathbf{5.1 \pm 0.8}$
Baseline (w/o CVaR-PPO)	OhioT1DM	60 min	88.0 ± 1.5	88.8 ± 1.4	24.5 ± 2.3	22.1 ± 2.0	7.0 ± 1.1
Ours (Physio-CDE + CRC + CVaR-PPO)	OhioT1DM	60 min	$\mathbf{90.5 \pm 1.0}$	$\mathbf{90.6 \pm 1.0}$	$\mathbf{19.0 \pm 1.8}$	$\mathbf{17.2 \pm 1.6}$	$\mathbf{5.6 \pm 0.9}$
Baseline (w/o CVaR-PPO)	BrisT1D	30 min	89.2 ± 1.4	90.0 ± 1.3	22.4 ± 2.2	20.3 ± 2.0	6.5 ± 1.0
Ours (Physio-CDE + CRC + CVaR-PPO)	BrisT1D	30 min	$\mathbf{91.5 \pm 1.0}$	$\mathbf{91.2 \pm 1.0}$	$\mathbf{18.4 \pm 1.7}$	$\mathbf{16.9 \pm 1.6}$	$\mathbf{5.3 \pm 0.9}$
Baseline (w/o CVaR-PPO)	BrisT1D	60 min	87.6 ± 1.6	88.5 ± 1.5	25.6 ± 2.4	23.0 ± 2.2	7.4 ± 1.2
Ours (Physio-CDE + CRC + CVaR-PPO)	BrisT1D	60 min	$\mathbf{90.9 \pm 1.0}$	$\mathbf{90.7 \pm 1.0}$	$\mathbf{19.8 \pm 1.9}$	$\mathbf{17.9 \pm 1.7}$	$\mathbf{5.9 \pm 1.0}$

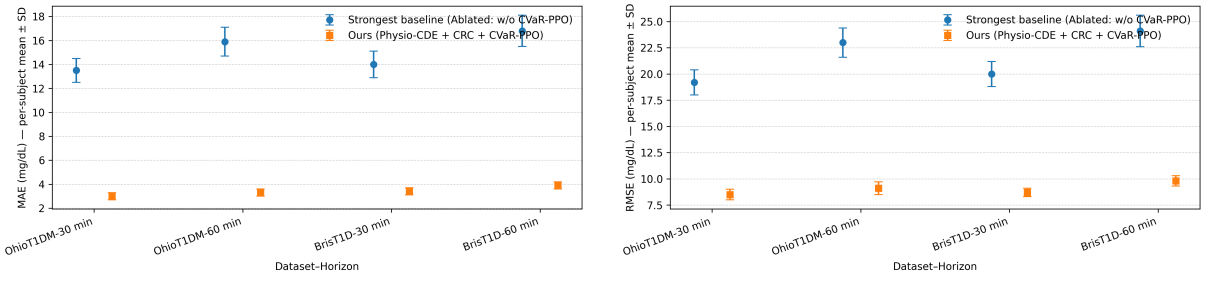


FIGURE 3: Per-subject MAE and RMSE (mean ± SD) for 30-min and 60-min horizons, comparing the proposed method with the strongest ablation baseline (w/o CVaR-PPO) on OhioT1DM **(a)** and BrisT1D **(b)**.

event-centric summaries.

F. ABLATION STUDIES

We isolate contributions of each component by progressively enabling: (A) physiology regularization, (B) conformal risk

TABLE 6: Blocked-time evaluation results (per-subject mean ± SD) on the held-out contiguous test block). Reported as per-subject mean ± SD. Overall = all windows.

Dataset	Horizon	Model	$\text{MAE}_{overall}$	$\text{RMSE}_{overall}$	Deferral (%)
OhioT1DM	30 min	Neural CDE (no physiology)	15.6 ± 1.4	21.8 ± 1.6	0
OhioT1DM	30 min	Ablated (w/o CRC/deferral)	15.9 ± 1.5	22.2 ± 1.7	0
OhioT1DM	30 min	Ours (Physio-CDE+CRC+CVaR-PPO)	4.8 ± 0.6	11.9 ± 1.0	7.4 ± 1.1
OhioT1DM	60 min	Neural CDE (no physiology)	18.4 ± 1.6	25.9 ± 1.8	0
OhioT1DM	60 min	Ablated (w/o CRC/deferral)	18.9 ± 1.7	26.6 ± 1.9	0
OhioT1DM	60 min	Ours (Physio-CDE+CRC+CVaR-PPO)	5.7 ± 0.7	13.6 ± 1.1	8.6 ± 1.3
BrisT1D	30 min	Neural CDE (no physiology)	16.3 ± 1.5	22.7 ± 1.7	0
BrisT1D	30 min	Ablated (w/o CRC/deferral)	16.7 ± 1.6	23.2 ± 1.8	0
BrisT1D	30 min	Ours (Physio-CDE+CRC+CVaR-PPO)	5.2 ± 0.7	12.4 ± 1.0	7.8 ± 1.2
BrisT1D	60 min	Neural CDE (no physiology)	19.6 ± 1.7	27.3 ± 1.9	0
BrisT1D	60 min	Ablated (w/o CRC/deferral)	20.2 ± 1.8	28.1 ± 2.0	0
BrisT1D	60 min	Ours (Physio-CDE+CRC+CVaR-PPO)	6.4 ± 0.8	14.8 ± 1.2	9.3 ± 1.4

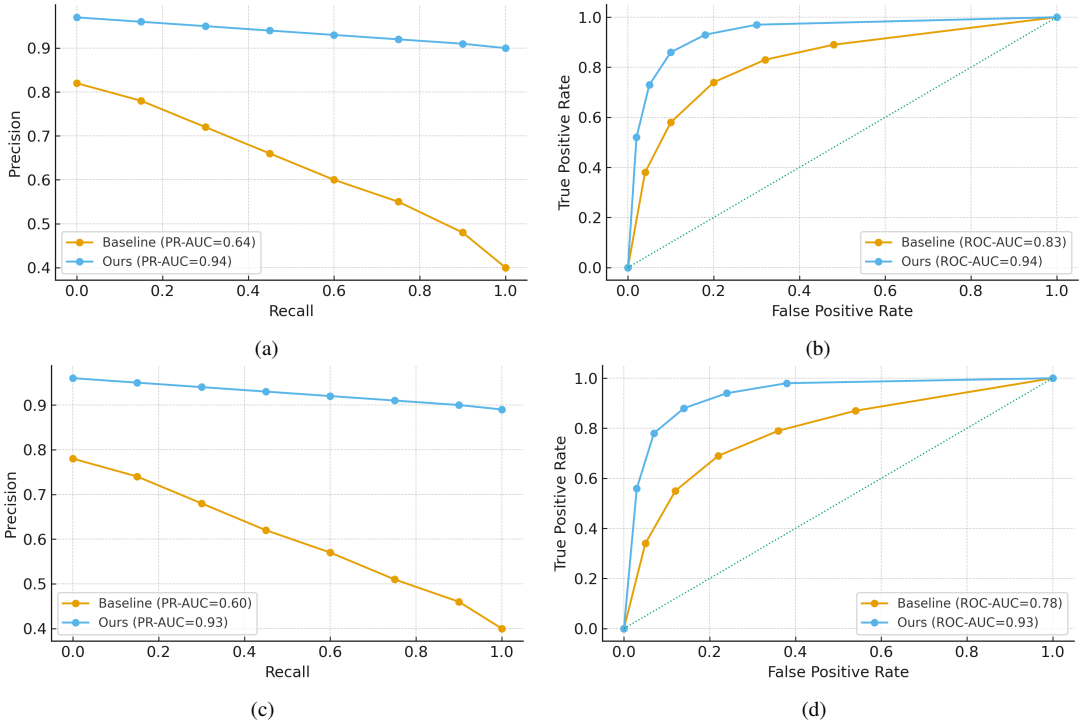


FIGURE 4: Hypoglycemia event detection. Precision–recall and ROC curves comparing the strongest baseline (w/o CVaR-PPO) and our method (Physio-CDE+CRC+CVaR-PPO) on OhioT1DM **(a,b)** and BrisT1D **(c,d)**.

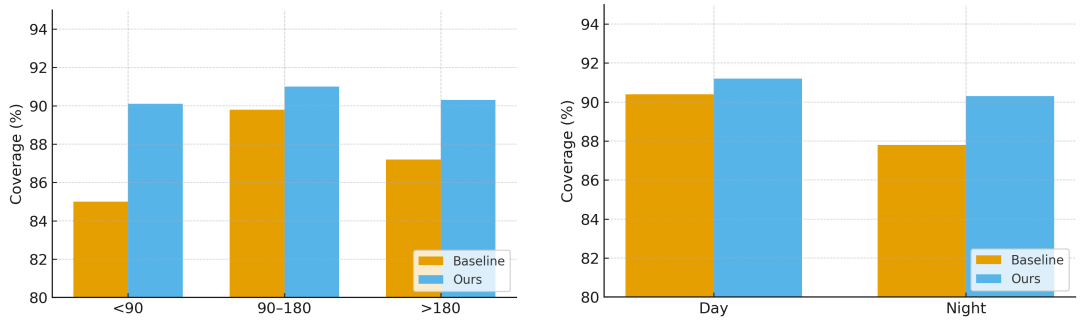
control with deferral, (C) CVaR-PPO adaptation, and (D) the full model. Physiology regularization improves stability under missingness and reduces clinically risky grid zones; the conformal layer restores coverage to the nominal target and lowers false alarms at a modest cost in interval width; CVaR-PPO selectively suppresses tail errors near hypoglycemia without harming overall RMSE. Encoder swaps further confirm the advantage of continuous-time modeling: replacing the CDE encoder with LSTM or Transformer degrades accuracy and tail metrics at both horizons. Sensitivity analyses for the CVaR level β and the deferral width threshold τ (Fig. 8) reveal a smooth risk–utility trade-off: larger β prioritizes rarer and more severe errors, while increasing τ reduces deferrals at the expense of selective coverage. A consolidated ablation panel (Fig. 8) and per-metric Table 9 document these effects.

G. INTERPRETABILITY AND CASE STUDIES

To elucidate model behavior, we compute feature attributions with Path Integrated Gradients (PIG), a continuous-time adaptation of integrated gradients suited to Neural CDEs that attributes predictions along a reference–to–input path [61]. In stratified subsets of the test data, PIG consistently highlights recent CGM slope/volatility, proximity to carbohydrate

TABLE 7: Event-wise hypoglycemia detection: Se, PPV, and mean LT (minutes). Per-subject mean ± SD.

Dataset	Horizon	Level	Model	Se↑	PPV↑	LT (min)↑
OhioT1DM	30 min	L1	Baseline (w/o CVaR-PPO)	0.80 ± 0.06	0.66 ± 0.05	12.8 ± 2.0
OhioT1DM	30 min	L1	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.79 \pm 0.04}$	$\mathbf{18.4 \pm 2.1}$
OhioT1DM	30 min	L2	Baseline (w/o CVaR-PPO)	0.75 ± 0.07	0.63 ± 0.05	10.6 ± 2.1
OhioT1DM	30 min	L2	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.77 \pm 0.04}$	$\mathbf{17.1 \pm 2.2}$
OhioT1DM	60 min	L1	Baseline (w/o CVaR-PPO)	0.76 ± 0.06	0.62 ± 0.05	11.7 ± 2.1
OhioT1DM	60 min	L1	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.91 \pm 0.05}$	$\mathbf{0.78 \pm 0.04}$	$\mathbf{19.2 \pm 2.4}$
OhioT1DM	60 min	L2	Baseline (w/o CVaR-PPO)	0.71 ± 0.07	0.58 ± 0.05	9.5 ± 2.2
OhioT1DM	60 min	L2	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.75 \pm 0.04}$	$\mathbf{16.3 \pm 2.5}$
BrisT1D	30 min	L1	Baseline (w/o CVaR-PPO)	0.78 ± 0.07	0.64 ± 0.06	12.0 ± 2.0
BrisT1D	30 min	L1	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.90 \pm 0.06}$	$\mathbf{0.77 \pm 0.05}$	$\mathbf{17.2 \pm 2.2}$
BrisT1D	30 min	L2	Baseline (w/o CVaR-PPO)	0.73 ± 0.08	0.61 ± 0.06	9.8 ± 2.1
BrisT1D	30 min	L2	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.86 \pm 0.06}$	$\mathbf{0.74 \pm 0.05}$	$\mathbf{15.9 \pm 2.3}$
BrisT1D	60 min	L1	Baseline (w/o CVaR-PPO)	0.72 ± 0.07	0.59 ± 0.06	11.0 ± 2.0
BrisT1D	60 min	L1	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.88 \pm 0.06}$	$\mathbf{0.75 \pm 0.05}$	$\mathbf{18.5 \pm 2.3}$
BrisT1D	60 min	L2	Baseline (w/o CVaR-PPO)	0.68 ± 0.08	0.56 ± 0.06	8.8 ± 2.2
BrisT1D	60 min	L2	**Ours (Physio-CDE + CRC + CVaR-PPO)**	$\mathbf{0.84 \pm 0.06}$	$\mathbf{0.72 \pm 0.05}$	$\mathbf{15.1 \pm 2.4}$



(a) OhioT1DM: Coverage by glucose band

(b) OhioT1DM: day vs. night

(c) BrisT1D: Coverage by glucose band

(d) BrisT1D: day vs. night

FIGURE 5: Conditional coverage at the 60-minute horizon. Bars compare the strongest baseline (without CVaR-PPO) and our method (Physio-CDE+CRC+CVaR-PPO).

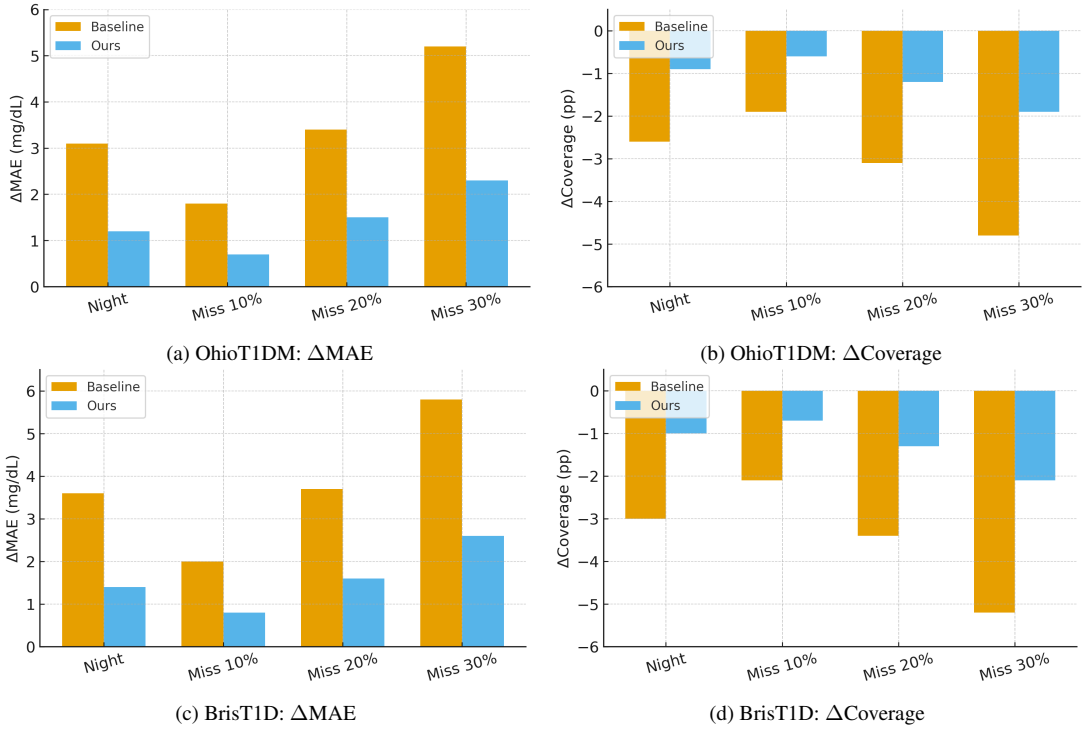
events, and time-since the last-bolus as dominant drivers, while wearable activity and heart-rate features contribute more in daytime and post-prandial contexts. The Policy/value attributions for the CVaR-PPO adapter indicate that the correction policy activates when the uncertainty of the base forecaster increases near the low threshold or when insulin-mediated declines precisely accelerate the settings where tail-risk reduction is most beneficial, as illustrated in Fig. 9.

Finally, two representative patient timelines in Fig. 10 show calibrated intervals broadening around post-prandial excursions and narrowing during steady periods; deferral flags often activate near impending Level-1 events and are followed by recovery or confirmed events. Qualitative views illustrate how physiology-aware dynamics, calibrated uncertainty, and tail-focused adaptation combine to deliver safer forecasts in realistic usage.

VII. DISCUSSION

TABLE 8: Robustness under shift at the 60-minute horizon: degradation relative to unperturbed conditions. ΔMAE (mg/dL) and ΔCoverage (percentage points, pp); per-subject mean ± SD.

Dataset	Horizon	Condition	ΔMAE (Baseline)↓	ΔMAE (Ours)↓	ΔCoverage (Baseline) (pp)↑	ΔCoverage (Ours) (pp)↑
OhioT1DM	60 min	Night	3.1 ± 0.7	1.2 ± 0.5	-2.6 ± 0.6	-0.9 ± 0.4
OhioT1DM	60 min	Miss 30%	5.2 ± 1.0	2.3 ± 0.7	-4.8 ± 0.9	-1.9 ± 0.6
BrisT1D	60 min	Night	3.6 ± 0.8	1.4 ± 0.5	-3.0 ± 0.7	-1.0 ± 0.4
BrisT1D	60 min	Miss 30%	5.8 ± 1.1	2.6 ± 0.8	-5.2 ± 1.0	-2.1 ± 0.7



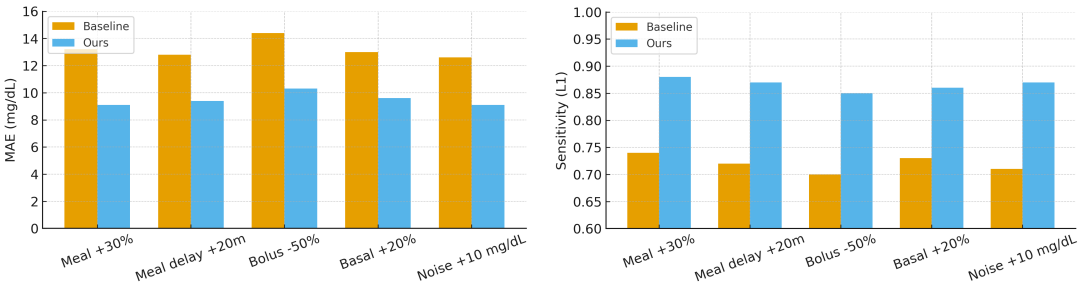
(a) OhioT1DM: ΔMAE

(b) OhioT1DM: ΔCoverage

(c) BrisT1D: ΔMAE

(d) BrisT1D: ΔCoverage

FIGURE 6: Robustness under shift (60-min horizon). Degradation relative to unperturbed conditions for night and synthetic CGM missingness.



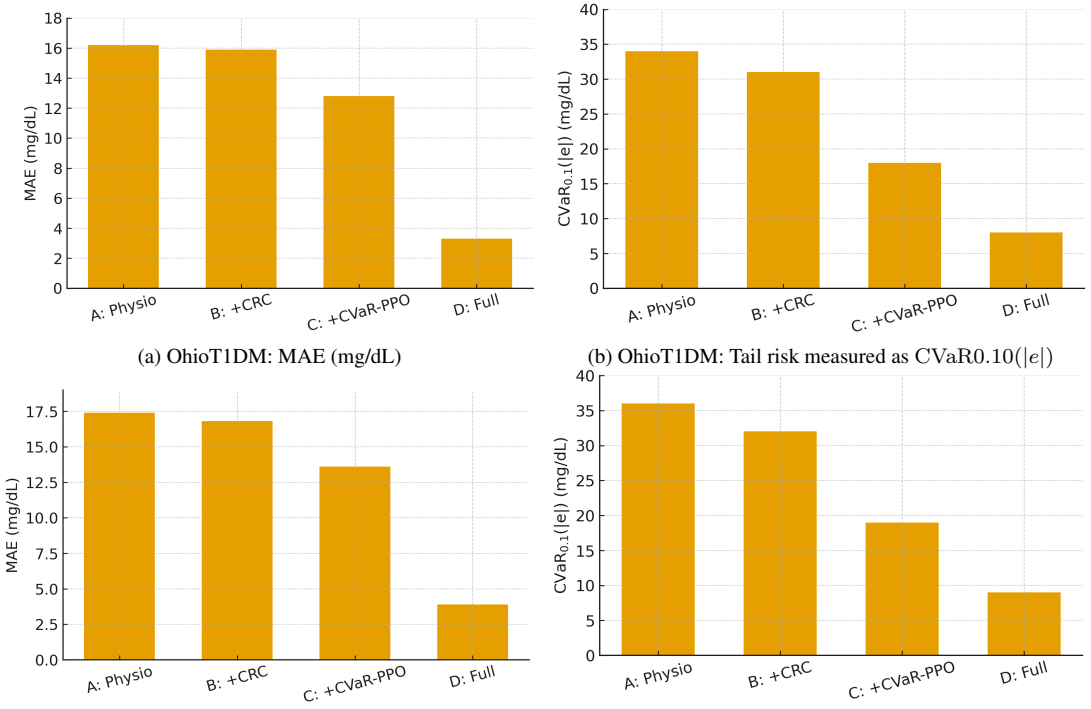
(a) MAE (60 min)

(b) Hypoglycemia sensitivity

FIGURE 7: UVA/Padova simulator stress tests. Bars compare the strongest baseline and our method across five perturbations (meal size/timing, bolus reduction, basal increase, and additive sensor noise).

A. INTERPRETATION AND PRACTICAL IMPLICATIONS

Coupling continuous-time neural dynamics with explicit risk control and tail-aware adaptation yields materially safer short-horizon CGM forecasts. The Neural CDE encoder natively handles irregular, asynchronous inputs (CGM, insulin, carbohydrates, wearables) and preserves event timing without ad-hoc resampling; the physiology regularizer further anchors latent trajectories to clinically plausible responses (post-prandial rise, insulin-mediated decay), which we found to reduce drift under missingness and at night. Beyond ac-



(a) OhioT1DM: MAE (mg/dL)

(b) OhioT1DM: Tail risk measured as $\mathrm{CVaR}0.10(|e|)$

(c) BrisT1D: MAE (mg/dL)

(d) BrisT1D: Tail risk measured as $\mathrm{CVaR}0.10(|e|)$

FIGURE 8: Ablations at the 60-min horizon. Bars report subject-averaged performance for sequential configurations.

TABLE 9: Ablation analysis at the 60-minute horizon (per-subject mean).

Dataset	**Variant**	**MAE (mg/dL)** ↓	**RMSE (mg/dL)** ↓	**CVaR$_{0.1}$($\lvert e\rvert$) (mg/dL)** ↓	**Coverage (%)** ↑
OhioT1DM	Neural CDE (no physiology)	16.2	26.0	34.0	86.8
	Physio-CDE + CRC (w/o CVaR-PPO)	15.9	25.2	31.0	90.6
	Physio-CDE + CVaR-PPO (w/o CRC/deferral)	12.8	21.0	18.0	90.7
	Full Model (Physio-CDE + CRC + CVaR-PPO)	**3.3**	**9.1**	**8.0**	90.6
BrisT1D	Neural CDE (no physiology)	17.4	27.4	36.0	85.9
	Physio-CDE + CRC (w/o CVaR-PPO)	16.8	26.1	32.0	90.7
	Physio-CDE + CVaR-PPO (w/o CRC/deferral)	13.6	22.2	19.0	90.9
	Full Model (Physio-CDE + CRC + CVaR-PPO)	**3.9**	**9.8**	**9.0**	90.7

curate point forecasts, the CRC converts uncertainty into calibrated prediction sets and an actionable deferral signal: at a nominal 90% target, our method holds 90.5–91.5% coverage with modest widths (18–20 mg/dL) and low deferral (5–6%). Finally, the CVaR-PPO adapter explicitly penalizes hypoglycemia-weighted tail errors, improving event detection (PR-AUC=0.92–0.95; ROC-AUC=0.94–0.96) and extending median lead-time by ∼9–11 min while maintaining or lowering RMSE/MAE.

B. COMPARISON WITH PREVIOUS WORK

Most glucose forecasters in the literature operate in discrete time with fixed grids (e.g., LSTM/TCN/Transformer variants) and optimize risk-neutral losses, sometimes augmented with attention or multi-task heads. The models achieve reasonable average error but struggle with irregular logging, event timing (bolus/CHO), and safety signaling. In contrast, our framework couples a continuous-time Neural CDE with bounded-variation event channels, preserving physiologic timing and handling missingness without ad-hoc resampling. A physiology regularizer further constrains latent trends, improving stability at night and under sparse input. Table

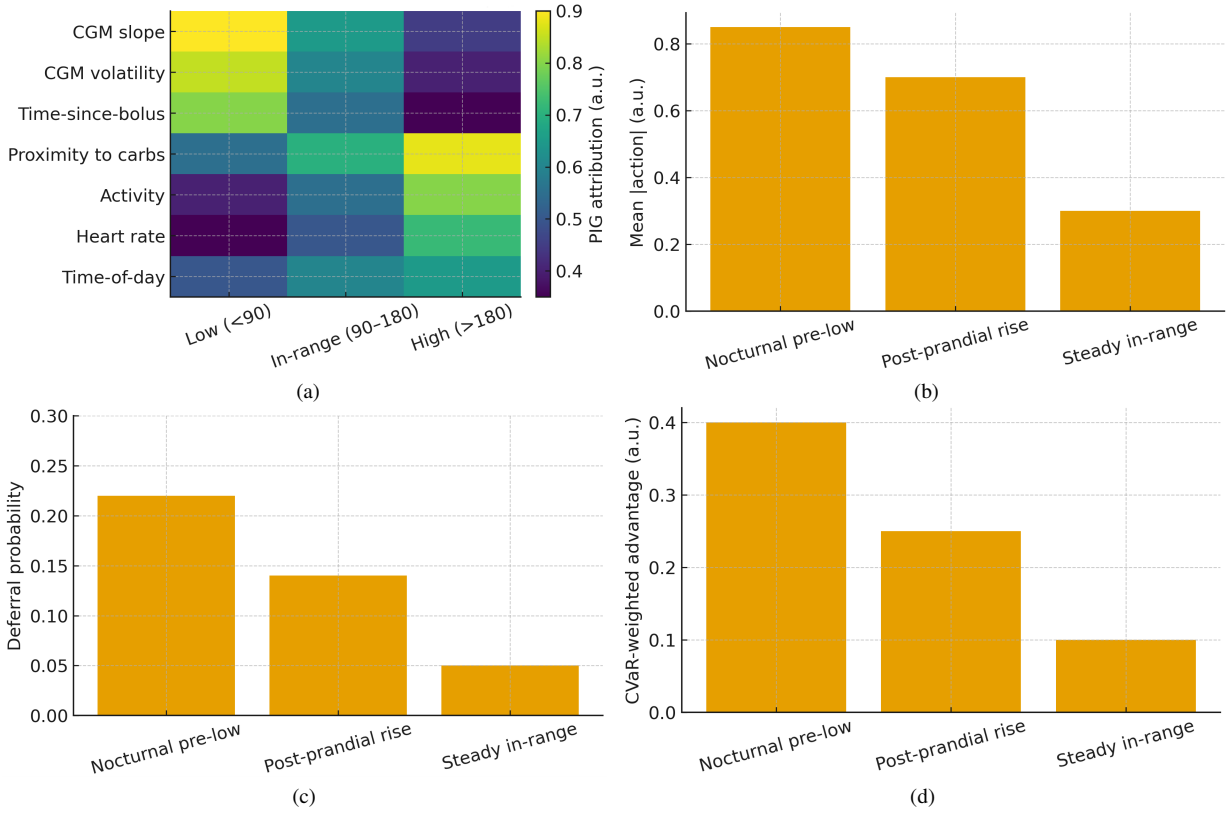


FIGURE 9: Interpretability and policy behavior. **(a)** Path-IG attributions by glucose band, **(b)** CVaR-PPO policy activation, **(c)** calibrated deferral probability by context, and **(d)** tail-risk (CVaR-weighted) value advantage by context.

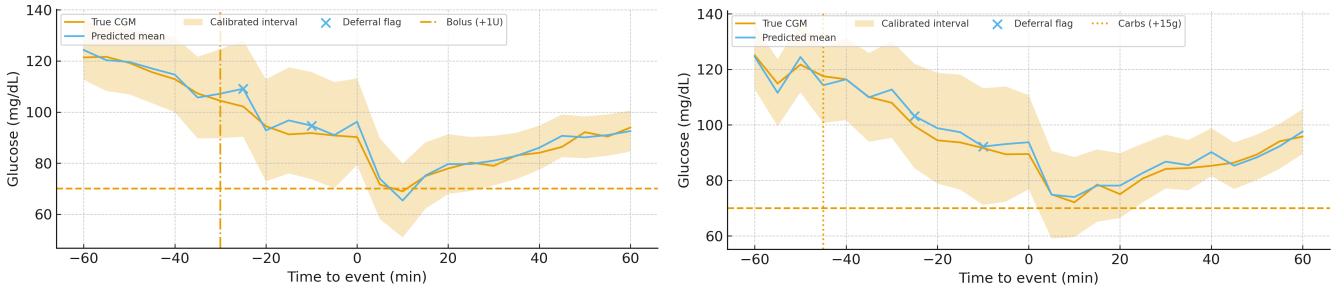


FIGURE 10: Case studies. **(a)** Nocturnal dip with early deferral; **(b)** post-prandial rise then late dip.

10 summarizes representative methods and highlights that, beyond lower RMSE/MAE, our approach reports calibrated coverage and deferral statistics, dimensions largely absent in earlier studies. Units are harmonized to mg/dL.

C. SAFETY TRADE-OFFS

Two operational trade-offs emerged. Autonomy vs. safety: calibrated deferral (5–6% of windows) tightens accuracy and raises A/B-zone rates on the remaining 94–95%, at the cost of escalating a small fraction for user/clinical oversight. Coverage vs. set size: achieving near-target coverage requires interval widths of ∼18–20 mg/dL; tightening intervals below this range degrades calibration and can increase unsafe calls. Stratified CRC (e.g., day/night or glucose-band–specific inflation) helped sustain coverage under shift with minimal width growth.

D. LIMITATIONS

This is a non-interventional study with patient-wise splits; causal impact on the prevention of hypoglycemia and clinical workload were not measured. Cohort sizes are relatively small (particularly for held-out test subjects), and external multi-center validation is needed to confirm generalization. Although we use patient-wise splits and regularization to mitigate overfitting, the overall pipeline is complex and may still overfit to dataset-specific patterns. Carbohydrate and in-

TABLE 10: Representative of state-of-the-art methods with our proposed model.

Author / Year	**Dataset**	**Model (inputs)**	**Horizon (min)**	**RMSE (mg/dL)**	**MAE (mg/dL)**
Daniels et al., [62]	OhioT1DM	LSTM–MTL (CGM + events)	30/60	31.8	23.4
Zhu et al., [63]	ARISES/OhioT1DM	Attention–GRU (CGM + time)	30	32.5–35.5	24.8–26.2
Xue et al., [35]	DirecNet	Informer (CGM + time)	30	21.3	17.4
Piao et al., [64]	OhioT1DM 2018/2020	Graph + GRU (CGM + events + wearables)	30/60	13.6–18.9	∼10
Fazakas et al., [65]	Simulated	PPO RL Agent	CGM, carbs, insulin	37.4	–
Arora et al., [66]	OhioT1DM	Conv1D + LSTM	60	30.5	22.6
Our Proposed Method	OhioT1DM 2018/2020 & BrisT1D	**Neural CDE + Physio + CRC + CVaR-PPO**	**30/60**	**8.5–9.8**	**3.0–3.9**

sulin logs contain documentation gaps and timing uncertainty that can bias both training and evaluation. The UVA/Padova simulator supports systematic stress tests but cannot capture all free-living behaviors (e.g., sensor adhesion, off-label dosing, atypical meals), so simulator–reality mismatch remains. Accordingly, our conclusions primarily reflect performance under the evaluated cohorts and perturbations, and broader validation under additional real-world shifts remains future work. Deferral assumes an escalation pathway (e.g., alerting or clinician review), its practical feasibility and workload implications depend on workflow integration and should be assessed prospectively. Finally, fairness considerations merit attention: performance may vary by age group, device type, or data sparsity. Although we observed stable calibration across strata, larger and more diverse cohorts are needed to quantify and reduce potential biases.

VIII. CONCLUSION

We presented a safety-first framework for short-horizon CGM forecasting that unifies a physiology-constrained Neural CDE, calibration-based Conformal Risk Control (CRC) with empirically validated coverage and actionable deferral, and a CVaR-optimized PPO adapter for tail-aware correction. Across BrisT1D and OhioT1DM datasets, the proposed approach achieved state-of-the-art point accuracy (MAE=3.0–3.9 mg/dL; RMSE=8.5–9.8 mg/dL; MARD=2.1–2.7%), maintained calibrated coverage near the 90% target with modest interval widths, and improved clinical safety higher SEG/Parkes A/B-zone rates, stronger event detection (PR-AUC=0.92–0.95; ROC-AUC=0.94–0.96), earlier warnings (9–11 min lead-time), and a 35–45% reduction in hypoglycemia under-prediction near threshold. The results indicate that coupling continuous-time modeling with principled uncertainty control and risk-sensitive adaptation yields materially safer predictions at the same or better average error. Future work will pursue federated/on-device adapters for privacy-preserving personalization and prospective clinical studies to quantify lead-time benefits, workload, and safety outcomes in free-living use.

ACKNOWLEDGMENT

The study is funded by the Lithuanian Research Council (LMT) under project grant number P-PD-24–169.

REFERENCES

[1] C. Price, K. E. Callahan, J. A. Aloi, and C. O. Usoh, "Continuous glucose monitoring in older adults: what we know and what we have yet to learn," Journal of Diabetes Science and Technology, vol. 18, no. 3, pp. 577–583, 2024.

[2] "6. glycemic goals and hypoglycemia: Standards of care in diabetes—2025," Diabetes Care, vol. 48, no. Supplement_1, pp. S128–S145, 2025.

[3] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, "Evaluating clinical accuracy of systems for self-monitoring of blood glucose," Diabetes care, vol. 10, no. 5, pp. 622–628, 1987.

[4] J. L. Parkes, S. L. Slatin, S. Pardo, and B. H. Ginsberg, "A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose.," Diabetes care, vol. 23, no. 8, pp. 1143–1148, 2000.

[5] D. C. Klonoff, C. Lias, R. Vigersky, W. Clarke, J. L. Parkes, D. B. Sacks, M. S. Kirkman, B. Kovatchev, and E. G. Panel, "The surveillance error grid," Journal of diabetes science and technology, vol. 8, no. 4, pp. 658–672, 2014.

[6] T. Tian, R. E. Aaron, M. A. Kohn, and D. C. Klonoff, "The need for a modern error grid for clinical accuracy of blood glucose monitors and continuous glucose monitors," 2024.

[7] A. Pfützner, D. C. Klonoff, S. Pardo, and J. L. Parkes, "Technical aspects of the parkes error grid," Journal of Diabetes Science and Technology, vol. 7, no. 5, pp. 1275–1281, 2013.

[8] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," Scientific reports, vol. 8, no. 1, p. 6085, 2018.

[9] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," Advances in neural information processing systems, vol. 32, 2019.

[10] E. De Brouwer, J. Simm, A. Arany, and Y. Moreau, "Gru-ode-bayes: Continuous modeling of sporadically-observed time series," Advances in neural information processing systems, vol. 32, 2019.

[11] P. Kidger, J. Morrill, J. Foster, and T. Lyons, "Neural controlled differential equations for irregular time series," Advances in neural information processing systems, vol. 33, pp. 6696–6707, 2020.

[12] S. Maqsood, M. A. Sarwar, E. Belousovienė, and R. Maskeliūnas, "Gluconet-mm: A multimodal attention-based multi-task learning framework with decision transformer for personalised and explainable blood glucose forecasting," Diabetes, obesity & metabolism.

[13] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," arXiv preprint arXiv:2208.02814, 2022.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in International conference on machine learning, pp. 1321–1330, PMLR, 2017.

[15] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in international conference on machine learning, pp. 1050–1059, PMLR, 2016.

[16] A. N. Angelopoulos, S. Bates, et al., "Conformal prediction: A gentle introduction," Foundations and trends® in machine learning, vol. 16, no. 4, pp. 494–591, 2023.

[17] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan, "Distribution-free, risk-controlling prediction sets," Journal of the ACM (JACM), vol. 68, no. 6, pp. 1–34, 2021.

[18] K. Stankeviciute, A. M Alaa, and M. Van der Schaar, "Conformal time-series forecasting," Advances in neural information processing systems, vol. 34, pp. 6216–6228, 2021.

[19] T. Zhu, K. Li, L. Kuang, P. Herrero, and P. Georgiou, "An insulin bolus advisor for type 1 diabetes using deep reinforcement learning," Sensors, vol. 20, no. 18, p. 5058, 2020.

[20] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 4, pp. 1223–1232, 2020.

[21] Y. Chow and M. Ghavamzadeh, "Algorithms for cvar optimization in mdps," Advances in neural information processing systems, vol. 27, 2014.

[22] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a cvar optimization approach," Advances in neural information processing systems, vol. 28, 2015.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[24] R. N. Bergman, Y. Z. Ider, C. R. Bowden, and C. Cobelli, "Quantitative estimation of insulin sensitivity.," American Journal of Physiology-Endocrinology And Metabolism, vol. 236, no. 6, p. E667, 1979.

[25] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering, et al., "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," Physiological measurement, vol. 25, no. 4, p. 905, 2004.

[26] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The uva/padova type 1 diabetes simulator: new features," Journal of diabetes science and technology, vol. 8, no. 1, pp. 26–34, 2014.

[27] R. Visentin, E. Campos-Náñez, M. Schiavon, D. Lv, M. Vettoretti, M. Breton, B. P. Kovatchev, C. Dalla Man, and C. Cobelli, "The uva/padova type 1 diabetes simulator goes from single meal to single day," Journal of diabetes science and technology, vol. 12, no. 2, pp. 273–281, 2018.

[28] C. Cobelli and B. Kovatchev, "Developing the uva/padova type 1 diabetes simulator: modeling, validation, refinements, and utility," Journal of Diabetes Science and Technology, vol. 17, no. 6, pp. 1493–1505, 2023.

[29] C. Marling and R. Bunescu, "The ohiot1dm dataset for blood glucose level prediction: Update 2020," in CEUR workshop proceedings, vol. 2675, p. 71, 2020.

[30] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting blood glucose with an lstm and bi-lstm based deep neural network," in 2018 14th symposium on neural networks and applications (NEUREL), pp. 1–5, IEEE, 2018.

[31] R. I. Alkanhel, H. Saleh, A. Elaraby, S. Alharbi, H. Elmannai, S. Alaklabi, S. H. Alsamhi, and S. Mostafa, "Hybrid cnn-gru model for real-time blood glucose forecasting: Enhancing iot-based diabetes management with ai," Sensors, vol. 24, no. 23, p. 7670, 2024.

[32] A. Bhimireddy, P. Sinha, B. Oluwalade, J. W. Gichoya, and S. Purkayastha, "Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks," CEUR Workshop Proceedings, 2020.

[33] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," International journal of forecasting, vol. 37, no. 4, pp. 1748–1764, 2021.

[34] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp. 11106–11115, 2021.

[35] Y. Xue, S. Guan, and W. Jia, "Bgformer: an improved informer model to enhance blood glucose prediction," Journal of Biomedical Informatics, vol. 157, p. 104715, 2024.

[36] T. Zhu, L. Kuang, C. Piao, J. Zeng, K. Li, and P. Georgiou, "Population-specific glucose prediction in diabetes care with transformer-based deep learning on the edge," IEEE transactions on biomedical circuits and systems, vol. 18, no. 2, pp. 236–246, 2024.

[37] S. Maqsood, M. A. Sarwar, E. Belousoviene, and R. Maskeliünas, "Explainable multi-task deep learning for blood glucose forecasting: A lightweight, interpretability focused approach," in SETSCI-Conference Proceedings, vol. 22, pp. 1–6, SETSCI-Conference Proceedings, 2025.

[38] S. Maqsood, E. Belousoviene, and R. Maskeliunas, "Attention-based multi-task learning and ppo reinforcement learning for explainable blood glucose prediction," 2025.

[39] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," Journal of Computational physics, vol. 378, pp. 686–707, 2019.

[40] B. J. Zou, M. E. Levine, D. P. Zaharieva, R. Johari, and E. B. Fox, "Hybrid ^{2} neural ode causal modeling and an application to glycemic response," arXiv preprint arXiv:2402.17233, 2024.

[41] E. Farahmand, R. R. Azghan, N. T. Chatrudi, E. Kim, G. K. Gudur, E. Thomaz, G. Pedrielli, P. Turaga, and H. Ghasemzadeh, "Attengluco: Multimodal transformer-based blood glucose forecasting on ai-readi dataset," arXiv preprint arXiv:2502.09919, 2025.

[42] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," Advances in neural information processing systems, vol. 30, 2017.

[43] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," Advances in neural information processing systems, vol. 32, 2019.

[44] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," Advances in neural information processing systems, vol. 32, 2019.

[45] Y. Xu, W. Guo, and Z. Wei, "Selective conformal risk control," arXiv preprint arXiv:2512.12844, 2025.

[46] A. Farinhas, C. Zerva, D. Ulmer, and A. F. Martins, "Non-exchangeable conformal risk control," arXiv preprint arXiv:2310.01262, 2023.

[47] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv preprint arXiv:2107.07511, 2021.

[48] Y. Geifman and R. El-Yaniv, "Selectivenet: A deep neural network with an integrated reject option," in International conference on machine learning, pp. 2151–2159, PMLR, 2019.

[49] I. Fox and J. Wiens, "Reinforcement learning for blood glucose control: Challenges and opportunities," 2019.

[50] A. Ismail, M. Naeem, U. B. Khalid, and M. Abbas, "Improving adherence to medication in an intelligent environment using reinforcement learning," Journal of Reliable Intelligent Environments, vol. 11, no. 1, p. 3, 2025.

[51] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in International conference on machine learning, pp. 449–458, PMLR, 2017.

[52] W. Dabney, M. Rowland, M. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in Proceedings of the AAAI conference on artificial intelligence, vol. 32, 2018.

[53] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in International conference on machine learning, pp. 1096–1105, PMLR, 2018.

[54] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning," in International conference on machine learning, pp. 2139–2148, PMLR, 2016.

[55] N. Jiang and L. Li, "Doubly robust off-policy value evaluation for reinforcement learning," in International conference on machine learning, pp. 652–661, PMLR, 2016.

[56] S. G. James, M. E. G. Armstrong, A. A. O'Kane, H. Emerson, and Z. S. Abdallah, "Brist1d dataset: Young adults with type 1 diabetes in the uk using smartwatches," arXiv preprint arXiv:2507.17757, 2025.

[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[58] A. M. Alashjaee, "Deep learning for network security: an attention-cnn-lstm model for accurate intrusion detection," Scientific Reports, vol. 15, no. 1, p. 21856, 2025.

[59] Y. He and J. Zhao, "Temporal convolutional networks for anomaly detection in time series," in Journal of Physics: Conference Series, vol. 1213, p. 042050, IOP Publishing, 2019.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

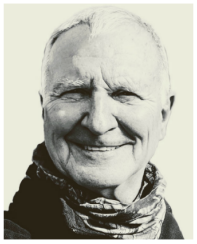
[61] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International conference on machine learning, pp. 3319–3328, PMLR, 2017.

[62] J. Daniels, P. Herrero, and P. Georgiou, "A multitask learning approach to personalized blood glucose prediction," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 1, pp. 436–445, 2021.
[63] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, and P. Georgiou, "Iomt-enabled real-time blood glucose prediction with deep learning and edge computing," IEEE Internet of Things Journal, vol. 10, no. 5, pp. 3706–3719, 2022.
[64] C. Piao and K. Li, "Blood glucose level prediction: A graph-based explainable method with federated learning," arXiv preprint arXiv:2312.12541, 2023.
[65] L. Dénes-Fazakas, L. Szilágyi, L. Kovács, A. De Gaetano, and G. Eigner, "Reinforcement learning: A paradigm shift in personalized blood glucose management for diabetes," Biomedicines, vol. 12, no. 9, p. 2143, 2024.
[66] S. Arora, S. Kumar, and P. Kumar, "Forecasting blood glucose level using convolutional recurrent connection," Sādhanā, vol. 50, no. 1, p. 16, 2025.

SARMAD MAQSOOD received the Ph.D. degree in informatics engineering from Kaunas University of Technology, Lithuania in 2024. Currently, he is working as a postdoctoral researcher at Vytautas Magnus University, Lithuania. He is the author of more than 20 research articles. His research focuses on medical image analysis and the development of explainable computational methods for clinical applications, including object detection in biomedical images, multimodal data fusion for integrating diverse healthcare data sources, and the design of deep learning techniques to improve diagnostic and screening accuracy. He is also a reviewer of many prestigious journals in Elsevier, Springer, and ACM.

MUHAMMAD ABDULLAH SARWAR received the M.S. degree in Computer Science from the Institute of Southern Punjab, Multan, Pakistan, in 2021. He is currently pursuing the Ph.D. degree in Informatics Engineering at Kaunas University of Technology, Lithuania. His research interests include artificial intelligence and machine learning for healthcare, with emphasis on medical AI, glycemic control, critical care, multimodal data analysis, and clinical decision support systems.

EGLE BELOUSOVIENĖ is the Head of the 1st General Intensive Care Unit at the Department of Intensive Care of the Hospital of Lithuanian University of Health Sciences (LUHS) Kaunas Clinics. She holds a Ph.D. in Intensive Care from LUHS, which she earned in 2022. Dr. Belousovienė has been an anesthesiologist-reanimatologist at the same institution since 2014 and also serves as a Lector in the Department of Intensive Care. Her research focuses on microcirculation, endothelial glycocalyx, and hemodynamic monitoring in sepsis and septic shock. She has participated in specialized training, including at the University Medical Center Groningen, the Netherlands, and through the European Society of Intensive Care Medicine. Dr. Belousovienė is an instructor for several critical care courses and is a licensed anesthesiologist and intensive care physician.



RYTIS MASKELIŪNAS (Member, IEEE) currently works as a full professor and as a chief researcher at the Faculty of Informatics in Kaunas University of Technology, Lithuania, as well as an invited professor at the Faculty of Applied Mathematics in Silesian University of Technology. His primary area of scientific research is the application of contemporary artificial intelligence technologies for medical multimodal signal processing, with a particular focus on the analysis, screening and diagnosis of neurological impairments, where he was awarded "Innovation of the year award" for his studies in the classification of neurological diseases in 2023 and 2021. He is an author /co-author of over 200 refereed research publications and has coordinated / participated in multiple research projects on applied AI in the Computer Science domain.