Risk-Aware Multi-Horizon Glucose Prediction for Type 2 Diabetes Using Temporal Attention and Conformal Prediction

Giada Lops[1], Vito Andrea Racanelli[1], Luca De Cicco[1], and Saverio Mascolo[1]

***Abstract*— Reliable glucose forecasting is an important tool for proactive and safety-oriented diabetes supervision. However, at long prediction horizons and in heterogeneous type 2 diabetes (T2D) populations, point forecasts alone are insufficient for risk-aware alerting. We propose a multi-horizon glucose prediction framework that employs a lightweight recurrent forecaster along with a temporal attention uncertainty quantification via split conformal prediction. Prediction intervals calibrated separately for each horizon are integrated into a supervisory layer that triggers alerts based on clinically relevant thresholds. Experiments on real-world CGM data show that temporal attention consistently improves forecasting accuracy, reducing the root mean square prediction error (RMSE) at a 120-min horizon from 33.4 to 28.3,mg/dL. When uncertainty is incorporated into supervision, event recall exceeds 0.92 for both hypoglycemia and hyperglycemia, approaching 0.95 at conservative operating points. By tuning the miscoverage level α, daily alarm rates can be reduced from above 90 to approximately 21–31 per day, depending on the event type, while maintaining recall above 0.90.**

***Index Terms*— Continuous glucose monitoring, multi-horizon forecasting, conformal prediction, uncertainty calibration, supervisory decision making, safety-critical monitoring.**

I. Introduction

Continuous glucose monitoring (CGM) allows data-driven decision support in diabetes management by providing glucose measurements under free-living conditions [1]. The increasing availability of longitudinal CGM data has motivated the development of predictive models aimed at early risk detection and proactive intervention, particularly in outpatient settings where delayed responses to adverse glycemic events may compromise patient safety [2], [3].

In type 2 diabetes (T2D), CGM is increasingly used both for retrospective analysis and prospective monitoring as well as for individualized risk assessment across heterogeneous patient populations [1]. Thus, predicting future glucose excursions is particularly important, as delayed responses to hypoglycemic or hyperglycemic events can lead to adverse outcomes and reduced safety margins [3], [4].

From a control standpoint, glucose forecasting can be seen as a prediction module employed in a control loop, where forecast outputs are used to trigger monitoring, supervision, or control actions [5].

This work has been partially funded from the European Union's Horizon Europe research and innovation programme under the project SHIELD, grant agreement No. 101156751.

[1]Department of Electrical and Information Engineering, Politecnico di Bari, Italy. g.lops@phd.poliba.it, {vitoandrea.racanelli, luca.decicco, saverio.mascolo}@poliba.it

Accordingly, predictive performance should be assessed both in terms of point-wise accuracy and its ability to support conservative and safe decisions under uncertainty.

In this context, it is particularly important to consider two important aspects. First, actionable interventions require anticipation on the order of one to two hours, thus motivating multi-horizon forecasting over prediction horizons ranging from 30 to 120 minutes [3], [4], [6]. Second, reliable uncertainty estimates are required, as point predictions alone are insufficient to characterize risk in the presence of physiological variability, sensor noise, and unmodeled disturbances [7], [8].

In this work, we propose a decision-oriented, multi-horizon glucose prediction framework that employs a temporal attention-based sequence modeling together with distribution-free uncertainty quantification and a conservative supervisory layer.

Prediction intervals are explicitly used to define supervisory alert rules for hypoglycemia and hyperglycemia based on clinically established thresholds. This allows to assess the trade-off between missed events and false alarms under strict patient-wise generalization constraints.

The remainder of the paper is organized as follows. Section II reviews related work. Section III presents the proposed framework. Section IV reports experimental results. Section V concludes the paper.

II. Related Work

This section reviews the literature on CGM-based glucose forecasting, focusing on modeling approaches, evaluation protocols, and uncertainty quantification for supervisory and decision-support systems.

A. Glucose forecasting from CGM data

Glucose forecasting from CGM data has long been investigated for anticipatory monitoring and prevention of adverse glycemic events, with early studies showing that even simple predictors can provide clinically useful anticipation [4].

Subsequent work has focused on improving robustness to inter-subject variability and mitigating performance degradation at longer horizons, with heterogeneity-aware models reporting consistent gains across prediction horizons [9].

More recently, attention-based architectures have been adopted to better capture informative temporal context in CGM time series. Hybrid Transformer–LSTM models and comparative studies across Transformer families report improved point forecasting accuracy, with strong horizon dependence and different architectures performing best at different forecast lengths [6], [10].

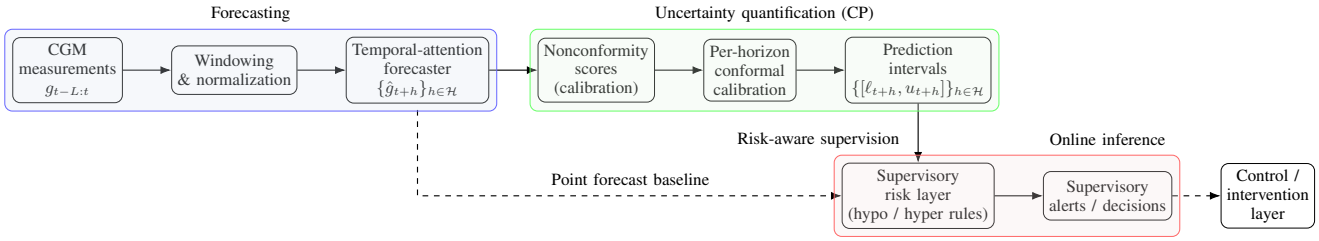

Fig. 1. Overview of the proposed pipeline. CGM measurements are first preprocessed and fed to a temporal-attention forecaster that produces multi-horizon glucose predictions. Prediction intervals for each horizon are then obtained through split conformal calibration applied to calibration residuals. These intervals are used in a supervisory risk layer that triggers hypoglycemia or hyperglycemia alerts according to clinically defined thresholds. The dashed path illustrates a baseline supervision strategy based on point forecasts only, without uncertainty information.

In parallel, pretrained and foundation-style models have been proposed to learn transferable representations from large CGM corpora, alongside personalization strategies tailored to therapy type. These approaches report MAE values in the range of 15–20 mg/dL at horizons between 60 and 120 minutes, depending on the evaluation protocol [11], [12].

B. Decision-oriented uncertainty quantification

In safety-oriented monitoring and decision support, point forecasts are insufficient because they do not expose the reliability margins required to balance *sensitivity* and *nuisance alarms*. Distribution-free conformal prediction is therefore attractive, as it can be applied to arbitrary forecasters and provides finite-sample marginal coverage under exchangeability assumptions [7].

A representative efficiency baseline is conformalized quantile regression (CQR), which combines quantile regression with conformal calibration to adapt interval widths to heteroscedastic prediction errors and achieve empirical coverage close to the nominal level with improved efficiency [8]. Extending conformal prediction to multi-horizon forecasting is non-trivial due to temporal dependence and error accumulation. Recent approaches address this challenge through joint calibration, copula-based modeling, and online or regime-aware updates [13], [14].

Beyond forecasting, recent deep learning approaches have explored decision-oriented and control-aware formulations, particularly in the context of automated insulin delivery and safety-aware reinforcement learning. These works demonstrate the potential of integrating predictive models with downstream decision and control logic, yielding measurable safety improvements in closed-loop settings [5], [15]–[17].

C. Positioning of this work

The existing literature highlights two main limitations.

First, glucose forecasting models are commonly evaluated using point error metrics at fixed horizons, providing limited insight into robustness under patient-wise generalization and into the suitability of long-horizon forecasts for safety-oriented supervision.

Second, although conformal prediction provides finite-sample coverage guarantees, most studies focus on marginal coverage and interval efficiency on generic benchmarks. The use of calibrated uncertainty to define explicit supervisory rules linked to clinically relevant events, and to quantify the trade-off between missed events and false alarms, remains limited.

This work proposes a multi-horizon glucose prediction framework for type 2 diabetes that integrates temporal attention, per-horizon conformal calibration, and risk-aware supervision (Fig. 1). Prediction intervals are directly used to trigger hypoglycemia and hyperglycemia alerts, enabling a quantitative assessment of the trade-off between missed events and false alarms under strict patient-wise evaluation.

The framework is evaluated on real-world CGM data using a split-by-patient protocol.

III. Methodology

This section describes the proposed framework for risk-aware multi-horizon glucose prediction and supervision, following a decision-oriented structure commonly adopted in monitoring and supervisory control systems. We introduce the problem setting and notation, the temporal forecasting models, the uncertainty quantification method based on split conformal prediction, and the supervisory decision logic.

A. Problem Setting and Notation

Let $\mathbf{x}_t \in \mathbb{R}^F$ denote the feature vector available at time t, and $\mathbf{X}_t := \mathbf{x}_{t-T+1:t} \in \mathbb{R}^{T\times F}$ the causal input window of length T ending at decision time t, where $\mathbf{x}_{a:b}$ denotes the ordered sequence $\{\mathbf{x}_a, \mathbf{x}_{a+1}, \ldots, \mathbf{x}_b\}$. The feature vector includes CGM measurements, contextual signals related to insulin delivery and dietary intake, and time-of-day encodings. All features are strictly causal.

At each decision time t, the goal is to predict future glucose values at K clinically relevant prediction horizons,

$$\hat{\mathbf{g}}_t = \left[\hat{g}_{t+h_1}, \hat{g}_{t+h_2}, \ldots, \hat{g}_{t+h_K}\right], \tag{1}$$

with corresponding true values $\mathbf{g}_t = [g_{t+h_1}, \ldots, g_{t+h_K}]$. The horizons $h_k > 0$ denote future prediction steps expressed in multiples of the CGM sampling period. In addition to point forecasts, the framework computes a prediction interval $\mathcal{I}_{h_k}(t)$ for each horizon h_k, later used by a supervisory layer to support risk-aware decisions.

B. Temporal Multi-Horizon Forecasting

We adopt a recurrent forecasting architecture based on Long Short-Term Memory (LSTM) units [18]. Given an

input window $\mathbf{X}_t$, the recurrent encoder produces a sequence of hidden states

$$\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T, \qquad \mathbf{h}_\tau \in \mathbb{R}^{d_h}, \tag{2}$$

where d_h denotes the hidden state dimension. These hidden states summarize the temporal information in the past window. The baseline forecaster derives the multi-horizon prediction from the final hidden state,

$$\hat{\mathbf{g}}_t = f_{\text{out}}(\mathbf{h}_T), \tag{3}$$

where $f_{\text{out}}(\cdot)$ is a shared prediction head producing K outputs. We denote the resulting point forecaster by $\hat{f}_\theta$, parameterized by θ, such that $\hat{\mathbf{g}}_t = \hat{f}_\theta(\mathbf{X}_t)$.

To exploit information from different parts of the input window, we extend the recurrent encoder with a temporal attention mechanism [19]. Attention scores are computed as

$$s_\tau = \mathbf{v}^\top \tanh(\mathbf{W}\mathbf{h}_\tau + \mathbf{b}), \tag{4}$$

where $\mathbf{W}$, $\mathbf{v}$, and $\mathbf{b}$ are trainable parameters of the attention mechanism. They are then normalized via the softmax function as $\alpha_\tau = \frac{\exp(s_\tau)}{\sum_{j=1}^{T}\exp(s_j)}$. A context vector is then formed as $\mathbf{c}_t = \sum_{\tau=1}^{T}\alpha_\tau \mathbf{h}_\tau$ and used to generate the multi-horizon forecast, $\hat{\mathbf{g}}_t = f_{\text{out}}(\mathbf{c}_t)$. The baseline and attention-based models share the same recurrent backbone and prediction head, differing only in the aggregation of past information.

C. Model Training

Let $\mathcal{S}_{\text{tr}} := \{(\mathbf{X}_{t^{(i)}}, \mathbf{g}_{t^{(i)}})\}_{i=1}^{n_{\text{tr}}}$ denote the training set. Here $t^{(i)}$ denotes the i-th decision time index extracted from the underlying glucose time series. Model parameters are learned by minimizing a robust regression loss,

$$\min_\theta \sum_{i=1}^{n_{\text{tr}}}\sum_{k=1}^{K} \rho_\delta\left(g_{t^{(i)}+h_k} - \hat{g}_{t^{(i)}+h_k}\right), \tag{5}$$

where $\rho_\delta(\cdot)$ is the Huber loss [20] and θ collects all trainable parameters. Early stopping is performed using a validation subset.

D. Uncertainty Quantification via Conformal Prediction

Predictive uncertainty is quantified using split conformal prediction [21], [22]. Given a calibration set $\mathcal{S}_{\text{cal}}$, horizon-specific *nonconformity scores* are defined as

$$\epsilon_{h_k}^{(i)} = \left|g_{t^{(i)}+h_k} - \hat{g}_{t^{(i)}+h_k}\right|. \tag{6}$$

For miscoverage level α, the conformal quantile is

$$q_{h_k}(\alpha) := Q_{\lceil (n_{\text{cal}}+1)(1-\alpha)\rceil/n_{\text{cal}}}\left(\{\epsilon_{h_k}^{(i)}\}\right). \tag{7}$$

Here $n_{\text{cal}} := |\mathcal{S}_{\text{cal}}|$ and $Q_p(\cdot)$ denotes the empirical quantile of order p. Prediction intervals are then given by

$$\mathcal{I}_{h_k}(t) = \left[\hat{g}_{t+h_k} - q_{h_k}(\alpha),\ \hat{g}_{t+h_k} + q_{h_k}(\alpha)\right]. \tag{8}$$

Smaller α yields wider, more conservative intervals.

E. Risk-Aware Supervisory Decision Rules

Let $\text{LB}_{h_k}(t)$ and $\text{UB}_{h_k}(t)$ denote the bounds of $\mathcal{I}_{h_k}(t)$ and g_{hypo} and g_{hyper} denote the clinical hypoglycemia and hyperglycemia thresholds, expressed in mg/dL. Alerts are triggered according to worst-case rules:

- Hypoglycemia if $\min_k \text{LB}_{h_k}(t) < g_{\text{hypo}}$,
- Hyperglycemia if $\max_k \text{UB}_{h_k}(t) > g_{\text{hyper}}$.

This envelope-based strategy aggregates information across horizons and is closely related to worst-case reasoning in robust control [23]. The miscoverage level α directly controls supervisory conservatism.

F. Evaluation Criteria

Performance is evaluated along three dimensions: (i) pointwise forecasting accuracy, (ii) uncertainty calibration across horizons, and (iii) risk-aware supervisory performance.

Point forecasting accuracy is assessed using MAE and RMSE, computed across all prediction horizons and decision windows.

Uncertainty calibration is evaluated via empirical marginal coverage and mean interval width (*sharpness*). Supervisory performance is assessed using event-level detection metrics (*recall* and *precision*), together with window-level false positive rate and alarm burden (*alerts per day*), as a proxy for potential alarm fatigue in continuous monitoring systems [24].

In addition, the evaluation includes comparisons against strong baseline models, ablation studies to isolate the contribution of individual components, and subgroup analyses to assess robustness under heterogeneous glycemic dynamics.

IV. Experimental Results

This section reports experimental results for the proposed glucose forecasting and supervision framework under realistic deployment conditions.

All experiments are conducted on real-world CGM data from individuals with T2D using a strict patient-wise data split. Models are evaluated exclusively on previously unseen subjects to assess generalization and robustness in clinical monitoring settings [25].

A. Experimental Setup

The ShanghaiT2DM dataset [1] includes CGM recordings from 100 individuals with T2D, spanning 109 recording periods. To prevent intra-subject leakage, all periods belonging to the same individual are grouped before data splitting.

CGM signals are resampled at $\Delta t = 15$ minutes. Missing gaps shorter than 30 minutes are linearly interpolated, while longer gaps are excluded from window extraction. Glucose values are clipped to $[40, 400]\,\text{mg/dL}$. Dietary intake and insulin signals are resampled on the same grid, with missing entries set to zero.

All feature engineering is strictly causal [26]. Sliding windows of length $T = 8$ samples (120 minutes) are extracted, with prediction targets corresponding to prediction horizons $h_k \in \{30, 60, 120\}$ minutes ahead.

This yields 110,557 input–output pairs, split by patient into training (65,746), calibration (26,945), and test (17,866) sets using a 60/20/20 partition. Normalization parameters are estimated on the training set only [25].

B. Uncertainty Calibration and Robustness

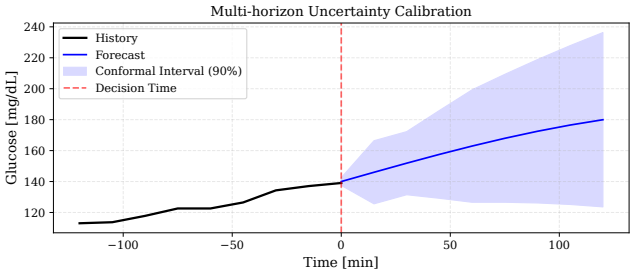

Fig. 2. Illustrative example of multi-horizon conformal prediction intervals. The shaded region represents the calibrated 90% safety envelope. Interval width increases monotonically from 40.8 mg/dL at 30 minutes to 112.7 mg/dL at 120 minutes, reflecting the growth of forecast uncertainty over time.

Uncertainty calibration is evaluated using empirical marginal coverage and mean interval width. Per-horizon conformal calibration is compared against a global baseline. As illustrated in Fig. 2, the prediction intervals widen progressively as the forecast horizon increases, forming a characteristic fan-shaped envelope typical of multi-step uncertainty propagation.

On the test set, per-horizon calibration maintains empirical coverage close to the nominal level $1-\alpha = 0.9$, ranging from 0.899 to 0.905 across horizons.

Interval width increases monotonically with the prediction horizon, consistent with error accumulation in multi-step forecasting [27].

Compared to global calibration, per-horizon conformal prediction avoids overly conservative intervals at short horizons and under-coverage at long horizons, thereby achieving a more balanced trade-off between coverage and interval sharpness across the full 120-minute prediction window.

C. Risk-Aware Event Detection and Supervisory Behavior

Risk-aware supervision is evaluated by assessing detection of hypo- and hyperglycemic events. Event labels are assigned at the window level if event onset occurs after the decision time t. Multiple alerts associated with the same clinical event are counted once, following [28].

Two alerting strategies are compared: *point-forecast-based alerting* and *interval-based supervision*. In the latter case, an alert is triggered if any horizon-specific prediction interval intersects the corresponding clinical threshold. False positive rates are reported at the window level.

Recall and precision are computed at the event level, while FPR refers to the window-level false positive rate.

As reported in Table I, interval-based supervision substantially increases event recall for both hypoglycemia and hyperglycemia. Hypoglycemia recall improves from 0.276 to 0.925, and hyperglycemia recall from 0.646 to 0.950.

TABLE I

RISK-AWARE EVENT DETECTION PERFORMANCE AND EVENT STATISTICS (TEST SET). CONFORMAL INTERVALS ARE EVALUATED AT $\alpha = 0.1$.

Method	Detection performance			Event statistics	
	Recall	Precision	FPR	# Events	Avg. duration [min]
Hypoglycemia					
Point forecast	0.276	**0.790**	**0.002**	48	41.2
Conformal	**0.925**	0.710	0.045		
Hyperglycemia					
Point forecast	0.646	**0.892**	**0.025**	72	56.7
Conformal	**0.950**	0.845	0.038		

These gains in sensitivity are achieved without a disproportionate loss in precision. Precision remains above 0.70 for hypoglycemia and above 0.84 for hyperglycemia, with only a moderate increase in the window-level false positive rate.

Overall, incorporating predictive intervals leads to markedly improved event detection while preserving a reasonable trade-off between safety and alert burden.

D. Decision Evaluation and Risk–Performance Trade-offs

Supervisory conservatism is analyzed by sweeping the miscoverage level $\alpha \in [0.01, 0.20]$ while keeping the underlying forecaster fixed. Figure 3 summarizes the resulting trade-off between clinical safety and alarm burden.

Lower values of α lead to highly conservative supervision, yielding near-perfect event recall (e.g., 0.98 for hypoglycemia at $\alpha = 0.02$) at the cost of a high alarm rate. As α increases, alarm burden is progressively reduced, from above 90 to approximately 21–31 alarms per day, depending on the event type, while recall remains above 0.90 for both hypo- and hyperglycemic events.

The miscoverage level α therefore acts as an interpretable supervisory gain, enabling explicit tuning of the safety–burden trade-off without requiring retraining or recalibration of the forecasting model. This property supports flexible and personalized risk management under fixed predictive performance.

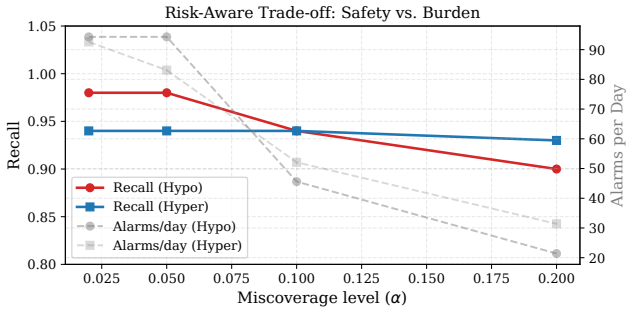

Fig. 3. Safety–burden trade-off as a function of the miscoverage level α. Increasing α reduces the daily alarm rate (burden) with a gradual reduction in event recall (safety), allowing explicit tuning of supervisory conservatism.

E. Comparison with Strong Baselines

We compare the proposed LSTM with temporal attention against two strong baselines representing complementary modeling paradigms: (i) PatchTST-lite, a Transformer-based

forecaster [29], and (ii) quantile regression with conformalized quantile regression (QR+CQR) [8].

PatchTST-lite represents fully attention-based time-series models, while QR+CQR serves as a non-Bayesian probabilistic baseline with finite-sample calibration guarantees. All models are trained and evaluated using identical data splits, preprocessing, and optimization budgets.

paragraphPoint forecasting

Table II reports MAE and RMSE values on the test set across prediction horizons. Across all horizons, the attention-based LSTM achieves the lowest MAE and RMSE.

QR (median) remains competitive, particularly at shorter horizons, but does not outperform the attention-based LSTM under the same optimization budget. PatchTST-lite does not provide systematic improvements, suggesting that recurrent architectures with explicit temporal inductive bias remain effective for short and noisy CGM windows.

TABLE II

POINT FORECASTING ACCURACY ON THE TEST SET. BEST VALUES PER HORIZON ARE HIGHLIGHTED IN BOLD.

	MAE [mg/dL]			RMSE [mg/dL]		
Horizon	LSTM+Attn	PatchTST	QR (median)	LSTM+Attn	PatchTST	QR (median)
30 min	**7.85**	10.95	9.01	**11.20**	18.29	13.03
60 min	**12.40**	17.00	15.59	**18.50**	25.95	22.87
120 min	**19.15**	24.86	24.23	**28.30**	36.16	34.80

a) Uncertainty calibration: Uncertainty estimates are compared using split conformal prediction (CP) for LSTM+Attn and PatchTST-lite, and CQR for the QR baseline. Table III reports empirical coverage and mean interval width at $\alpha = 0.1$.

All methods achieve coverage close to the nominal level, consistent with their calibration guarantees. Differences primarily emerge in the coverage–sharpness trade-off. While QR produces the narrowest intervals at the shortest horizon, the attention-based LSTM achieves tighter intervals at medium and long horizons while maintaining comparable coverage. PatchTST-lite yields wider intervals without systematic gains in coverage.

Overall, temporal attention combined with per-horizon conformal calibration provides the most balanced performance across horizons.

TABLE III

UNCERTAINTY CALIBRATION ON THE TEST SET ($\alpha = 0.1$). BEST VALUES ARE HIGHLIGHTED IN BOLD (COVERAGE CLOSEST TO 0.9; MINIMUM WIDTH).

Model	30 min		60 min		120 min	
	Coverage	Width	Coverage	Width	Coverage	Width
LSTM+Attn	**0.899**	40.7	**0.905**	**62.1**	**0.905**	**92.4**
PatchTST (CP)	0.897	45.5	0.903	77.3	**0.905**	114.3
QR (CQR)	0.891	**40.2**	0.889	70.7	0.883	103.5

F. Ablation Study

We conduct an ablation study to isolate the individual contributions of temporal attention and conformal calibration. All model variants are trained and evaluated using identical data splits, preprocessing pipelines, and optimization settings, ensuring that observed differences can be attributed to architectural or calibration choices.

Four variants are considered: (i) an LSTM without attention; (ii) an LSTM with temporal attention using point forecasts only; (iii) an attention-based LSTM with global conformal calibration; and (iv) an attention-based LSTM with per-horizon conformal calibration (full model).

Table IV summarizes the results of the ablation study. Removing temporal attention consistently increases point forecasting errors across all horizons, confirming its contribution to predictive accuracy.

Regarding uncertainty calibration, global conformal calibration produces horizon-invariant interval widths and exhibits under-coverage at longer horizons, while per-horizon calibration yields interval widths that increase with forecast uncertainty and maintains coverage close to the nominal level. This behavior is consistent with the theoretical properties of multi-horizon conformal prediction [30]. Bootstrap confidence intervals at the 95% level, obtained by resampling recording periods, are tight across all variants.

Consistent with the calibration results, downstream event detection shows that point-only supervision leads to reduced event recall, as it does not account for predictive uncertainty, whereas global calibration results in mismatched conservatism across horizons.

Overall, the attention-based model with per-horizon conformal calibration provides the most consistent balance between safety and alarm burden across prediction horizons.

TABLE IV

ABLATION STUDY ON THE TEST SET ($\alpha = 0.1$). BEST FORECASTING VALUES ARE HIGHLIGHTED IN BOLD.

	MAE [mg/dL]			RMSE [mg/dL]		
Model	30	60	120	30	60	120
LSTM (no Attn)	9.10	15.85	23.95	13.20	22.70	33.40
LSTM + Attn (point)	**7.85**	**12.40**	**19.15**	**11.20**	**18.50**	**28.30**
	Coverage			Width [mg/dL]		
Model	30	60	120	30	60	120
Attn + global CP	0.912	0.898	0.865	118.5	118.5	118.5
Attn + per-horizon CP	**0.899**	**0.905**	**0.905**	**40.7**	**62.1**	**92.4**

G. Subgroup Analysis

Robustness under heterogeneous operating conditions is evaluated by stratifying test recording periods according to CGM-derived glycemic variability, measured as the per-period CGM standard deviation.

Three disjoint subgroups are defined using quartiles of the empirical distribution: low variability ($\sigma_{\mathrm{CGM}} \leq 31.4\,\mathrm{mg/dL}$), medium variability (interquartile range), and high variability ($\sigma_{\mathrm{CGM}} \geq 45.7\,\mathrm{mg/dL}$), with seven recording periods per subgroup.

Table V reports point forecasting accuracy and risk-aware event detection metrics for the LSTM with temporal attention and conformal supervision. Results are reported as subgroup means with 95% bootstrap confidence intervals obtained by resampling recording periods within each subgroup.

Point forecasting errors increase monotonically with glycemic variability, reflecting the greater difficulty of pre-

TABLE V

SUBGROUP ANALYSIS BY CGM VARIABILITY (TEST SET). MAE AND EVENT-LEVEL METRICS WITH 95% BOOTSTRAP CONFIDENCE INTERVALS.

Group	MAE_{30}	MAE_{60}	MAE_{120}	Hypo R	Hyper R	Alarms/d
Low	7.51 [6.80,8.35]	12.15 [11.36,13.28]	17.72 [16.91,18.68]	0.96 [0.75,1.00]	0.92 [0.88,1.00]	31–67
Medium	9.84 [9.30,11.03]	16.41 [15.40,17.99]	25.73 [23.39,28.52]	**1.00**	0.93 [0.91,0.95]	36–61
High	13.63 [10.90,15.09]	21.43 [18.04,23.33]	33.59 [29.14,36.19]	**1.00**	**0.95** [0.93,0.97]	27–70

dicting unstable glucose dynamics [31]. In contrast, event-level recall remains high across all subgroups, indicating that the interval-based supervisory layer preserves detection sensitivity under heterogeneous dynamics, at the cost of increased alerting in high-variability regimes.

V. CONCLUSION

This paper presented a risk-aware, multi-horizon glucose prediction framework for type 2 diabetes that combines temporal attention-based forecasting with distribution-free uncertainty quantification.

Compared to a capacity-matched LSTM baseline, temporal attention consistently improved forecasting accuracy across horizons, reducing MAE by approximately 20% at medium and long horizons and achieving $12.4\,\mathrm{mg/dL}$ at 60 minutes and $19.2\,\mathrm{mg/dL}$ at 120 minutes on the test set.

Per-horizon conformal calibration provided empirical coverage close to the nominal 90% level across all horizons. When prediction intervals were integrated into a supervisory decision layer, event recall for both hypoglycemia and hyperglycemia exceeded 0.92, approaching 0.95. This reduced missed adverse events by nearly 90% relative to point-forecast-based alerting, at the expense of a moderate increase in false positives.

Overall, the results show that explicitly incorporating predictive uncertainty enables a transparent and tunable trade-off between safety and specificity in glucose monitoring systems.

Future work will focus on adaptive conformal calibration under non-stationary conditions and on tighter integration with downstream decision-making or control architectures.

REFERENCES

[1] Q. Zhao, J. Zhu, C. Wang, and W. Rao, "Diabetes datasets: Shanghait1dm and shanghait2dm," 2022.

[2] B. Cinar, L. van den Boom, and M. Maleshkova, "Review of machine learning models in short- and long-term glucose forecasting and hypoglycemia classification," *Informatics in Medicine Unlocked*, vol. 60, p. 101723, 2026.

[3] A. Facchinetti, "Continuous glucose monitoring and hypoglycaemia prevention," *Diabetes Care*, vol. 42, no. 8, pp. 1401–1412, 2019.

[4] G. Sparacino, F. Zanderigo, S. Corazza, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 931–937, 2007.

[5] F. Liu *et al.*, "Time-series optimization with downstream decision-awareness," *IEEE Transactions on Automatic Control*, 2025.

[6] M. Karagoz, M. Breton, and A. El Fathi, "A comparative study of transformer-based models for multi-horizon blood glucose prediction," *IFAC Diabetes Technology Conference Proceedings*, 2025.

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

[8] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," *Advances in Neural Information Processing Systems*, 2019.

[9] Y. Huang, Z. Ni, Z. Lu, X. He, J. Hu, B. Li, H. Ya, and Y. Shi, "Heterogeneous temporal representation for diabetic blood glucose prediction," *Frontiers in Physiology*, vol. 14, p. 1225638, 2023.

[10] Q. Bian, A. As'arry, X. Cong *et al.*, "A hybrid transformer-lstm model applied to glucose prediction," *PLOS ONE*, vol. 19, p. e0310084, 2024.

[11] Y. Lu, D. Liu, Z. Liang *et al.*, "A pretrained transformer model for decoding individual glucose dynamics from continuous glucose monitoring data," *National Science Review*, vol. 12, no. 5, p. nwaf039, 2025.

[12] F. Lara-Abelenda *et al.*, "Personalized glucose forecasting for people with diabetes using deep learning models," *Journal of Biomedical Informatics*, 2025.

[13] S. Sun and R. Yu, "Copula conformal prediction for multi-step time series forecasting," *OpenReview ICLR 2024*, 2024.

[14] F. Schlembach, E. Smirnov, and I. Koprinska, "Conformal multistep-ahead multivariate time-series forecasting," *Machine Learning*, 2025.

[15] G. Lops, V. A. Racanelli, G. Manfredi, L. De Cicco, and S. Mascolo, "A deep deterministic policy gradient control algorithm for automatic insulin delivery," *IFAC-PapersOnLine*, vol. 59, no. 2, pp. 13–18, 2025.

[16] G. Lops, G. Manfredi, V. A. Racanelli, L. De Cicco, and S. Mascolo, "A safety-aware deep reinforcement learning technique for automated insulin delivery," in *2025 33rd Mediterranean Conference on Control and Automation (MED)*. IEEE, 2025, pp. 120–125.

[17] F. Baldisseri, M. M. H. Atanasious, V. Becchetti, A. Di Paola, G. Lops, D. Menegatti, A. Wrona, S. Mascolo, and F. Delli Priscoli, "A quantitative comparison of deep reinforcement learning algorithms for type 1 diabetes control," in *2025 11th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2025.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *International Conference on Learning Representations*, 2015.

[20] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[21] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[22] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, pp. 486–507, 2021.

[23] J. B. Rawlings, D. Q. Mayne, and M. Diehl, *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2017.

[24] S. Sendelbach and M. Funk, "Alarm fatigue: a patient safety concern," *AACN Advanced Critical Care*, vol. 24, no. 4, pp. 378–388, 2013.

[25] V. M. Young *et al.*, "Data leakage in deep learning for alzheimer's disease diagnosis: A scoping review of methodological rigor and performance inflation," *Diagnostics*, vol. 15, no. 18, p. 2348, 2025.

[26] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, 03 2018.

[27] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. OTexts, 2021.

[28] H. Rezatofighi *et al.*, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

[29] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations*, 2023.

[30] K. Stankeviciute, F. Giannotti, Y. Elgebaly *et al.*, "Conformal time-series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6216–6228, 2021.

[31] L. Monnier and C. Colette, "Glycaemic variability: a strong independent predictor of hypoglycaemia in diabetes," *Diabetes & Metabolism*, vol. 33, no. 4, pp. 295–301, 2007.