Quantitative Predictive Monitoring and Control for Safe Human-Machine Interaction

Shuyang Dong[1], Meiyi Ma[2], Josephine Lamp[3], Sebastian Elbaum[1], Matthew B. Dwyer[1], Lu Feng[1]

[1]University of Virginia
[2]Vanderbilt University
[3]DexCom

{sd3mn, se4ja, md3cn, lf9u}@virginia.edu, meiyi.ma@vanderbilt.edu, josephine.lamp@dexcom.com

Abstract

There is a growing trend toward AI systems interacting with humans to revolutionize a range of application domains such as healthcare and transportation. However, unsafe human-machine interaction can lead to catastrophic failures. We propose a novel approach that predicts future states by accounting for the uncertainty of human interaction, monitors whether predictions satisfy or violate safety requirements, and adapts control actions based on the predictive monitoring results. Specifically, we develop a new quantitative predictive monitor based on *Signal Temporal Logic with Uncertainty* (STL-U) to compute a *robustness degree interval*, which indicates the extent to which a sequence of uncertain predictions satisfies or violates an STL-U requirement. We also develop a new loss function to guide the uncertainty calibration of Bayesian deep learning and a new adaptive control method, both of which leverage STL-U quantitative predictive monitoring results. We apply the proposed approach to two case studies: Type 1 Diabetes management and semi-autonomous driving. Experiments show that the proposed approach improves safety and effectiveness in both case studies.

1 Introduction

There is a growing trend toward AI systems interacting with humans to revolutionize a range of application domains such as healthcare and transportation. However, unsafe human-machine interaction can lead to catastrophic failures (e.g., crashes of automated vehicles (Banks, Plant, and Stanton 2018) and robot-caused fatalities (Yang et al. 2022). Ensuring the safety of human-machine interaction poses significant challenges. First, safety is an emergent property that requires holistic reasoning about AI systems and human operators (Ma, Stankovic, and Feng 2021). Second, modeling human-machine interaction should account for the inherent uncertainty of human behavior. Moreover, safe and prompt decision-making under uncertainty entails predictive monitoring, i.e., making predictions about future states and monitoring if safety requirements would be violated. We concretize these challenges using a motivating example below.

Type 1 Diabetes (T1D) is a chronic disease that affects millions of patients whose pancreas produces little to no insulin to regulate blood glucose (BG). Uncontrolled diabetes may cause hypoglycemia (BG below 70 mg/dL) or hyperglycemia (BG above 180 mg/dL), which over time can lead to serious damage to organs such as the kidneys and heart. Over the past decades, advanced technologies such as continuous glucose monitoring (CGM) sensors and insulin pumps have been developed to reduce the need for patients to check BG via finger-pricking and self-injections of insulin. Recently, there have been promising breakthroughs in the development of Artificial Pancreas Systems (APS), which are automated or semi-automated closed-loop insulin delivery systems to regulate BG levels. A typical APS controller calculates insulin dosages based on CGM sensor readings and user input (e.g., meal carbohydrates). There has been increasing interest in using machine learning techniques to predict future BG levels, which are then fed into an APS controller. To guarantee a safety requirement, such as "BG levels should be regulated within the range of 70-180 mg/dL to avoid hypoglycemia or hyperglycemia", it is not sufficient to only check APS control actions. Instead, it is necessary to reason about the behavior of the entire closed-loop system of APS, including the insulin pump, CGM sensor, as well as patient physiology (e.g., glucose metabolism) and behavior (e.g., eating).

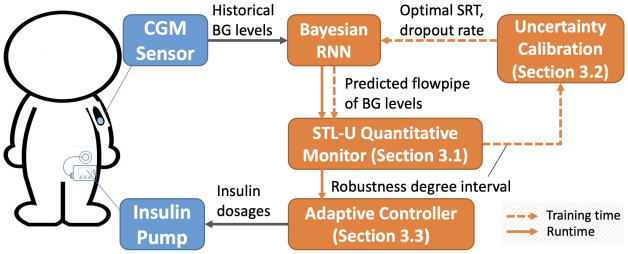

Figure 1: Proposed approach applied to T1D management.

To tackle these challenges, we propose a novel logic-based quantitative predictive monitoring and control approach, as illustrated in Figure 1. We adopt Recurrent Neural Networks (RNNs), which are well-suited to time series data (Bengio, Goodfellow, and Courville 2017), for making sequential predictions about future BG levels. Commonly used RNNs, such as long short-term memory (LSTM) networks, are deterministic models that generate the same predictions with the same input (Hochreiter and Schmidhuber

1997; Ma et al. 2020). To account for the uncertainty of human physiology and behavior, we cast deterministic RNNs into Bayesian RNNs using stochastic regularization techniques (SRTs) (Gal 2016). Bayesian RNNs yield uncertain sequential predictions with the uncertainty estimated by a sequence of posterior probability distributions.

(Ma et al. 2021) characterize Bayesian RNN predictions as a *flowpipe* signal containing an infinite set of predicted traces, and present *Signal Temporal Logic with Uncertainty* (STL-U) to check if a flowpipe strongly or weakly satisfies a requirement (i.e., whether the requirement satisfaction holds for all or some traces contained in a flowpipe). Nevertheless, existing STL-U monitors do not provide quantitative information about the degree to which a flowpipe satisfies or violates a requirement, which is imperative for fine-grained decision-making in safe human-machine interactions such as diabetes management.

In this work, we develop a new STL-U quantitative monitor that computes a *robustness degree interval*, indicating the degree to which a requirement is satisfied or violated. The lower and upper bounds of a robustness degree interval correspond to the worst-case and best-case estimates of the degree to which a flowpipe satisfies a requirement. Additionally, we define a loss function that leverages STL-U quantitative monitoring results to calibrate the uncertainty estimation of Bayesian RNNs during training. We select the optimal combination of SRT and dropout rate that yields the smallest loss. Predictions made by Bayesian RNNs using the calibrated SRT and dropout rate improve the quality of uncertainty estimates with respect to requirement satisfaction. Furthermore, we adapt control actions based on STL-U robustness degree intervals. For instance, we present a proof-of-concept adaptive APS controller that increases or decreases insulin dosages depending on the robustness degree for predicted BG levels that violate the safety requirement. Finally, we evaluate the proposed approach through experiments using the state-of-the-art UVA/PADOVA T1D patient simulator (Man et al. 2014). To demonstrate the generalizability of the proposed approach, we also apply it to a second case study of semi-autonomous driving using the CARLA simulator (CARLATeam 2023), which is included in the appendix of (Dong et al. 2024) due to page limits.

2 Background

2.1 Uncertain Sequential Prediction

Stochastic regularization techniques (SRTs) are commonly employed to transform deterministic deep learning models into Bayesian models, enabling uncertainty estimation. In this work, we transform a deterministic RNN model into a Bayesian RNN model via SRTs. We consider four commonly used SRTs: Bernoulli dropout, Bernoulli dropConnect, Gaussian dropout, and Gaussian dropConnect (Gal 2016). These SRTs have various ways to determine which neuron connections to drop based on sampling from a probability distribution with certain dropout rate p. The larger the value of p, the more neuron connections are retained.

A Bayesian RNN model yields a set of sequential predictions by applying Monte Carlo sampling N times. At each step t along the sequence, a Gaussian distribution $\Phi_t \sim N(\theta_t, \sigma_t^2)$ can be estimated, whose mean θ_t and variance σ_t are calculated based on the Monte Carlo samples $\{x_t^{(1)}, \cdots, x_t^{(N)}\}$. The uncertainty estimates of Bayesian RNN predictions are bounded by the Gaussian distribution's confidence interval $[\Phi_t^-(\varepsilon), \Phi_t^+(\varepsilon)]$ under a confidence level $\varepsilon \in (0, 1)$. The larger the confidence interval range, the higher the estimated uncertainty.

2.2 Signal Temporal Logic with Uncertainty

(Ma et al. 2021) characterize Bayesian RNN predictions as a *flowpipe* signal ω over a discrete time domain $\mathbb{T}$. At each time t, the flowpipe contains all values bounded within a confidence interval $[\Phi_t^-(\varepsilon), \Phi_t^+(\varepsilon)]$. Figure 2 shows an example flowpipe of predicted BG levels.

Signal Temporal Logic with Uncertainty (STL-U) with the following syntax is proposed in (Ma et al. 2021).

$$\varphi := \mu(\varepsilon) \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \Box_I \varphi \mid \Diamond_I \varphi \mid \varphi_1 \, U_I \, \varphi_2$$

where $\Box_I$, $\Diamond_I$ and U_I are temporal operators "always", "eventually", and "until" with a time interval $I \subseteq \mathbb{T}$, respectively. $\mu(\varepsilon)$ is an atomic predicate whose value is determined by a function $f(x) > 0$ for $x \in [\Phi_t^-(\varepsilon), \Phi_t^+(\varepsilon)]$ under a confidence level ε. For example, $\Box_{[0,3]}(BG_{\varepsilon=90\%} > 70)$ is an STL-U formula expressing the requirement that "the predicted BG level under a 90% confidence level should always be above 70 mg/dL in the next three hours".

STL-U semantics defined in (Ma et al. 2021) include two indices: *strong satisfaction* (i.e., all values bounded within the flowpipe's confidence interval satisfy φ), and *weak satisfaction* (i.e., there exists some value within the flowpipe's confidence interval satisfying φ). For example, the flowpipe shown in Figure 2 strongly satisfies $\Box_{[t,t_1]}(BG_\varepsilon > 70)$, weakly satisfies $\Box_{[t,t_2]}(BG_\varepsilon > 70)$, and strongly violates $\Box_{[t,t_3]}(BG_\varepsilon > 70)$.

Additionally, a loss function, denoted by L_{sat}, is proposed in (Ma et al. 2021) based on STL-U strong/weak satisfaction relations to calibrate the uncertainty estimation of Bayesian RNNs by guiding the choice of SRTs and dropout rates.

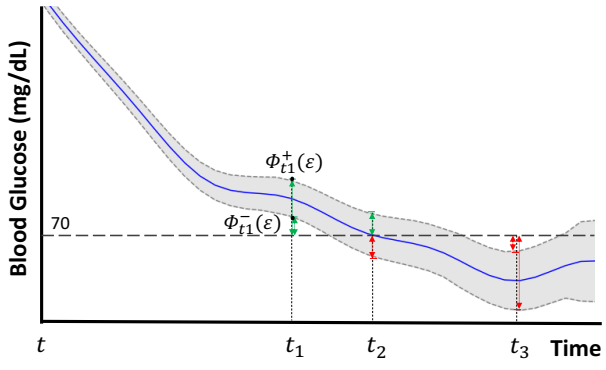

Figure 2: An example flowpipe of predicted BG levels under a confidence level ε.

3 Approach

The goal of our approach is to provide quantitative information about the degree to which an STL-U formula is satisfied or violated, which is imperative for fine-grained decision making in safe human-machine interaction. To achieve this goal, we develop a new STL-U quantitative monitor in Section 3.1. The monitoring results are leveraged to improve the uncertainty calibration through a new loss function in Section 3.2 and an adaptive control method in Section 3.3.

3.1 STL-U Quantitative Monitor

We propose a new quantitative semantics of STL-U, which computes a *robustness degree* function $\rho(\varphi, \omega, t)$ indicating how much an STL-U formula φ is satisfied or violated by a flowpipe signal ω at time t.

Let $\upsilon = [\underline{\upsilon}, \overline{\upsilon}]$ denote a real-valued interval. We define the following three interval operations:

$-^*\upsilon \stackrel{\text{def}}{=} [-\overline{\upsilon}, -\underline{\upsilon}]$

$min^*(\upsilon_1, \ldots, \upsilon_n) \stackrel{\text{def}}{=} [\min(\underline{\upsilon}_1, \ldots, \underline{\upsilon}_n), \min(\overline{\upsilon}_1, \ldots, \overline{\upsilon}_n)]$

$max^*(\upsilon_1, \ldots, \upsilon_n) \stackrel{\text{def}}{=} [\max(\underline{\upsilon}_1, \ldots, \underline{\upsilon}_n), \max(\overline{\upsilon}_1, \ldots, \overline{\upsilon}_n)]$

Definition 1 (STL-U quantitative semantics)

$$
\begin{aligned}
\rho(\mu(\varepsilon), \omega, t) &= [\min\big(f(x)\big), \max\big(f(x)\big)], \\
&\quad \forall x \in [\Phi_t^-(\varepsilon), \Phi_t^+(\varepsilon)] \\
\rho(\neg\varphi, \omega, t) &= -^*\rho(\varphi, \omega, t) \\
\rho(\varphi_1 \wedge \varphi_2, \omega, t) &= min^*\big(\rho(\varphi_1, \omega, t),\ \rho(\varphi_2, \omega, t)\big) \\
\rho(\Box_I \varphi, \omega, t) &= \min_{t' \in (t+I)}^* \rho(\varphi, \omega, t') \\
\rho(\Diamond_I \varphi, \omega, t) &= \max_{t' \in (t+I)}^* \rho(\varphi, \omega, t') \\
\rho(\varphi_1 U_I \varphi_2, \omega, t) &= \max_{t' \in (t+I)}^* \Big(min^*\big(\rho(\varphi_2, \omega, t'), \\
&\quad \min_{t'' \in [t, t']}^* \rho(\varphi_1, \omega, t'')\big)\Big)
\end{aligned}
$$

Intuitively, a robustness degree function yields an interval whose lower/upper bounds corresponding to the worst/best cases of a flowpipe satisfying or violating an STL-U formula. A positive (*resp.* negative) robustness value indicates the degree of satisfaction (*resp.* violation).

Algorithm 1 illustrates STL-U quantitative monitoring algorithm based on Definition 1. We can apply this algorithm recursively to monitor complex STL-U formulas with multiple levels of nesting temporal operators. The algorithm has a linear time complexity with respect to the length of the flowpipe, $|\omega|$.

Here is an example of checking the flowpipe in Figure 2 against an STL-U formula $\Box_{[t,t_3]}(BG_\varepsilon > 70)$. First, we check the atomic predicate at each time $\tau \in [t, t_3]$, and compute the robustness degree interval $[\min\big(f(x)\big), \max\big(f(x)\big)]$ for all $x \in [\Phi_\tau^-(\varepsilon), \Phi_\tau^+(\varepsilon)]$, where $f(x) = x - 70$. The flowpipe at t_2 is bounded by $[60, 80]$ and yields a robustness degree interval $[-10, +10]$. The flowpipe at t_3 is bounded by $[40, 65]$ and its robustness degree interval is $[-30, -5]$. Finally, we obtain a robustness degree interval for the always operator $\Box_{[t,t_3]}$ by taking the minimal of the lower/upper bounds of the atomic predicate's robustness degree intervals over all time steps $\tau \in [t, t_3]$. The resulting robustness degree interval, $[-30, -5]$, indicates that the predicted flowpipe would violate the requirement by -30 and -5 in the worst and best-case scenarios, respectively.

Algorithm 1 STL-U quantitative monitoring algorithm

Function $Monitor(\varphi, \omega, t)$**:**
- **switch** φ **do**
  - **Case** $\mu(\varepsilon)$
    - $\rho \leftarrow [\min\big(f(x)\big), \max\big(f(x)\big)]$, for all $x \in [\Phi_t^-(\varepsilon), \Phi_t^+(\varepsilon)]$
    - **return** ρ
  - **Case** $\neg\varphi$
    - **return** $-^*\big(Monitor(\varphi, \omega, t)\big)$
  - **Case** $\varphi_1 \wedge \varphi_2$
    - **return** $min^*\big(Monitor(\varphi_1, \omega, t), Monitor(\varphi_2, \omega, t)\big)$
  - **Case** $\Box_I \varphi$
    - $\rho \leftarrow Monitor(\varphi, \omega, t)$
    - **for** $t' \in (t + I)$ **do**
      - $\rho \leftarrow min^*\big(\rho, Monitor(\varphi, \omega, t')\big)$
    - **return** ρ
  - **Case** $\Diamond_I \varphi$
    - $\rho \leftarrow Monitor(\varphi, \omega, t)$
    - **for** $t' \in (t + I)$ **do**
      - $\rho \leftarrow max^*\big(\rho, Monitor(\varphi, \omega, t')\big)$
    - **return** ρ
  - **Case** $\varphi_1\ U_I\ \varphi_2$
    - $\rho \leftarrow (-\infty, -\infty)$; $\rho_1 \leftarrow Monitor(\varphi_1, \omega, t)$
    - **for** $t' \in (t + I)$ **do**
      - $\rho_2 \leftarrow Monitor(\varphi_2, \omega, t')$
      - **for** $t'' \in [t, t']$ **do**
        - $\rho_1 \leftarrow min^*\big(\rho_1, Monitor(\varphi_1, \omega, t'')\big)$
      - $\rho \leftarrow max^*\big(\rho, min^*(\rho_1, \rho_2)\big)$
    - **return** ρ

The soundness of the proposed STL-U quantitative monitor is stated below and the proof is given in the appendix of (Dong et al. 2024).

Theorem 1 *Given an STL-U formula φ and a flowpipe ω, the following properties hold.*

1. $\underline{\rho} > 0 \Rightarrow (\omega, t) \models_s \varphi$
2. $\underline{\rho} \leq 0 \Rightarrow (\omega, t) \not\models_s \varphi$
3. $\overline{\rho} > 0 \Rightarrow (\omega, t) \models_w \varphi$
4. $\overline{\rho} \leq 0 \Rightarrow (\omega, t) \not\models_w \varphi$

*where $\underline{\rho}$ and $\overline{\rho}$ are the lower and upper bounds of the robustness degree interval $\rho(\varphi, \omega, t)$, and $\models_s$ (*resp. $\models_w$*) denotes the strong (*resp. *weak) satisfaction relations.*

3.2 Uncertainty Calibration

A Bayesian RNN model may produce divergent uncertainty estimates for a model trained on identical data, depending on various choices of SRTs and dropout rates. The current

practice often selects an SRT and dropout rate empirically or guided by traditional deep learning metrics such as prediction accuracy, which tend to overestimate uncertainty (i.e., the wider the flowpipe, the higher the accuracy of containing the ground truth trace).

We propose a loss function based on STL-U quantitative monitoring results to guide the choice of SRTs and dropout rate. The proposed loss function, denoted as $L_{qt}(\omega, \hat{\omega}, \varphi)$, is a linear combination of two parts: $\eta_r(\omega, \hat{\omega}, \varphi)$ for comparing the predicted flowpipe ω and the target trace $\hat{\omega}$ in terms of the robustness degrees of satisfying an STL-U formula φ, and $\eta_d(\omega, \hat{\omega})$ for measuring the distance between the predicted flowpipe ω and the target trace $\hat{\omega}$. Formally,

$$\eta_r(\omega, \hat{\omega}, \varphi) = \begin{cases} \underline{\rho}(\varphi, \omega, t) & , \quad \hat{\omega} \models_s \varphi \\ -\overline{\rho}(\varphi, \omega, t) & , \quad \hat{\omega} \not\models_s \varphi \end{cases} \tag{1}$$

where $\underline{\rho}(\varphi, \omega, t)$ and $\overline{\rho}(\varphi, \omega, t)$ are the lower and upper bounds of robustness degree intervals.

$$\eta_d(\omega, \hat{\omega}) = \sum_{\tau=0}^{|\hat{\omega}|} \begin{cases} 0 & , \quad \Phi_\tau^-(\varepsilon) \leq \hat{\omega}_\tau \leq \Phi_\tau^+(\varepsilon) \\ \Phi_\tau^-(\varepsilon) - \hat{\omega}_\tau & , \quad \hat{\omega}_\tau < \Phi_\tau^-(\varepsilon) \\ \hat{\omega}_\tau - \Phi_\tau^+(\varepsilon) & , \quad \Phi_\tau^+(\varepsilon) < \hat{\omega}_\tau \end{cases} \tag{2}$$

where $|\hat{\omega}|$ is the length of the target trace and $\hat{\omega}_\tau$ is the target trace's value at time τ.

The loss function is then given by

$$L_{qt}(\omega, \hat{\omega}, \varphi) = -\beta \cdot \eta_r(\omega, \hat{\omega}, \varphi) + (1 - \beta) \cdot \eta_d(\omega, \hat{\omega}) \tag{3}$$

where $\beta \in [0, 1]$ is a real-valued coefficient indicating the relative importance of the two parts.

The goal is to calibrate the uncertainty estimation of a Bayesian RNN model by choosing an optimal combination of SRT and dropout rate that yield flowpipe predictions with minimal loss $L_{qt}(\omega, \hat{\omega}, \varphi)$. Intuitively, the lower the loss, the higher the quality of predicted flowpipes, which achieve a higher (*resp.* lower) robustness degree of satisfying (*resp.* violating) an STL-U formula and a smaller distance from the target trace.

3.3 Adaptive Controller

A typical (deterministic) controller, denoted by $\pi : S \to A$, takes an input state $s \in S$ from the environment and produces an output action $a \in A$. We develop an adaptive controller $\pi' : S \times \rho \to A$ that adapts control actions based on STL-U quantitative monitoring results $\rho(\varphi, \omega, t)$ as shown in Figure 3. The adaptation schema can be domain-specific.

As a proof of concept, we present an adaptive controller based on the Basal-Bolus Controller (Kovatchev et al. 2009) included in the UVA/PADOVA T1D Simulator. The default Basal-Bolus Controller takes input such as the current BG level and meal carbohydrates, and computes the basal and bolus insulin dosages as control actions to regulate BG levels. A constant amount of basal insulin is delivered at each step, denoted by $defaultBasal$, whose value is calculated by multiplying the patient's body weight with a constant representing the steady state insulin rate per kilogram. Additionally, the controller issues a (non-zero) bolus insulin at a step when the patient takes a meal with carbohydrates. The

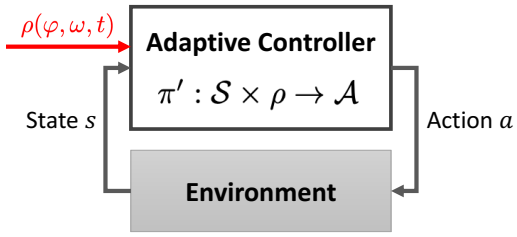

Figure 3: Adapting control actions based on STL-U quantitative monitoring results.

bolus insulin dosage, denoted by $mealBolus$, is calculated based on the amount of carbohydrates, the current and target BG levels, and the patient's carbohydrate ratio and correction factor.

Algorithm 2 Adapting a Basal-Bolus Controller

Input : STL-U quantitative monitoring results $\rho(\varphi_l, \omega, t)$ and $\rho(\varphi_h, \omega, t)$, current BG level g_t, next planned meal time t_m,time-window length K, and $bolusFlag$ for a meal bolus
Output: Control action a_t at time t

```
// Adapting basal insulin dosages
basal ← defaultBasal
if ρ(φ_l, ω, t) < −20 then
    basal ← 0
else if −20 ≤ ρ(φ_l, ω, t) ≤ 0 then
    basal ← defaultBasal × 0.8
else if −70 ≤ ρ(φ_h, ω, t) ≤ 0 then
    basal ← defaultBasal × 1.2
else if ρ(φ_h, ω, t) < −70 then
    basal ← defaultBasal × 1.5
// Adapting bolus insulin timing
bolus ← 0
if bolusFlag == false then
    if t_m − K ≤ t < t_m then
        if ρ(φ_l, ω, t) ≤ 0 or g_t ≤ 70 then
            bolus ← 0
        else
            bolus ← mealBolus; bolusFlag ← true
    else if t = t_m then
        bolus ← mealBolus; bolusFlag ← true
return a_t = ⟨basal, bolus⟩
```

Algorithm 2 adapts the Basal-Bolus Controller based on quantitative results of monitoring two STL-U formulas: $\varphi_l = \square_{[0,\infty)}(BG_\varepsilon > 70)$ and $\varphi_h = \square_{[0,\infty)}(BG_\varepsilon < 180)$. It decreases or increases basal insulin dosages based on the worst-case robustness degrees of predicted hypoglycemias and hyperglycemias as follows. It reduces the $basal$ dose to zero when $\underline{\rho}(\varphi_l, \omega, t) < -20$, indicating severe hypoglycemia with BG below 50 mg/dL. It decreases the $basal$ dose to 80% of the default when $-20 \leq \underline{\rho}(\varphi_l, \omega, t) \leq 0$, indicating mild hypoglycemia. It increases the $basal$ dose to 120% when $-70 \leq \underline{\rho}(\varphi_h, \omega, t) \leq 0$, indicating mild hyperglycemia. Finally, it increases the $basal$ dose to 150% of

the default when $\underline{\rho}(\varphi_h, \omega, t) < -70$, indicating severe hyperglycemia with BG above 250 mg/dL. We set these BG thresholds and basal change percentages following medical guidelines (American_Diabetes_Association 2022).

Furthermore, the adaptive controller adapts the time to administer a bolus insulin based on the current and predicted BG levels. Studies (Slattery, Amiel, and Choudhary 2018) have shown that taking a bolus 15-30 minutes before a meal helps to improve the glucose control. The optimal bolus timing varies with patient circumstances and insulin effects. Drawing on this insight, we design the adaptive controller to encourage pre-meal boluses. As shown in Algorithm 2, the adaptive controller decides the bolus timing by checking the current and predicted BG levels from K steps before a meal, where K is a constant recommended by medical guidelines (e.g., 45 minutes). The adaptive controller would not issue a pre-meal bolus when the current or predicted BG levels are below 70 mg/dL because of the risk of hypoglycemia.

4 Experiments

We evaluate the proposed approach via experiments using the UVA/PADOVA T1D Patient Simulator (Man et al. 2014), which has been approved by the U.S. Food and Drug Administration (FDA) for pre-clinical experiments with *in silico* populations. We investigate three research questions:

- **RQ1**: How useful is the proposed loss function for the uncertainty calibration of Bayesian RNN predictions?
- **RQ2**: How good is the proposed predictive monitor for the early and accurate detection of safety hazards?
- **RQ3**: How safe and effective is the proposed predictive monitoring and control approach in the closed-loop simulation of T1D management?

Datasets and experimental setup. We use the simulator to generate data based on 30 virtual patient profiles including: 10 adults, 10 adolescents and 10 children. Each patient is simulated for an 85-day period, with each time step in the simulation representing 3 minutes. We use 70-day, 5-day, and 10-day data for training, validation, and testing, respectively. We segment the data into samples of 20 steps (1 hour) by sliding windows, and obtain about 413,326 samples in total for each patient population. Violations are observed in 23% of adolescent data, 48% of adult data, and 52% of child data.

We train three different LSTM models for adults, adolescents, and children, respectively. They seek to predict BG levels in the future 30 minutes using the historical data of the past 30 minutes. The data inputs include CGM sensor readings, meal carbohydrates, insulin dosages, low/high blood glucose indexes, hypo/hyper risk, and time of a day. Each model is trained for 50 epochs. Given a chosen SRT and dropout rate, we repeat Bayesian LSTM predictions for 30 times using the Monte Carlo method. We estimate Gaussian distributions using these predictions and obtain flowpipes under a 95% confidence level. Using a higher confidence level broadens the predicted flowpipe, which can increase requirement violations. While this may appear conservative, it aligns with our focus on safety in STL-U requirements,

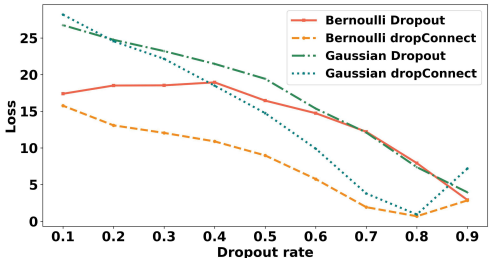

Figure 4: Loss values of using different SRTs and dropout rates in the Bayesian LSTM model for adults.

where conservatism helps mitigate risks proactively. We use the validation datasets to select the optimal SRT and dropout rate under the guidance of the proposed loss function, and use the selected optimal SRT and dropout rate to generate Bayesian LSTM predictions for the testing datasets.

Lastly, we apply the proposed predictive monitoring and control approach in the closed-loop simulation of T1D management over a 7-day period (different from the 85-day data mentioned before) for 30 virtual patients.

The experiments were run on a machine with 2.1GHz CPU, Nvidia Quadro RTX5000 GPU, 128GB memory, and CentOS 7 operating system.

4.1 RQ1: Evaluation of Uncertainty Calibration

Figure 4 plots values of the proposed loss function L_{qt} with respect to an STL-U formula $\varphi = \square_{[0,\infty)}(70 < BG_\varepsilon < 180)$, when varying SRTs and dropout rates of the Bayesian LSTM model for adults with the validation dataset. The results show that Bernoulli dropConnect with a dropout rate of 0.8 is the best choice with the smallest loss. For the LSTM models of adolescents and children (plots omitted due to page limits), we found Bernoulli dropConnect with a rate of 0.9 and Gaussian Dropout with a rate of 0.9 to be the best, respectively.

We compare the performance of the proposed loss function L_{qt} with two baselines taken from (Ma et al. 2021): L_{acc} which captures prediction accuracy (i.e., whether the target trace is entirely contained in a predicted flowpipe), and L_{sat} which captures STL-U strong/weak satisfaction. We evaluate the *F1 scores of requirement satisfaction* metric defined as $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$, where TP denotes the number of true positives (i.e., when the target trace satisfies φ and the predicted flowpipe ω yields $\underline{\rho}(\varphi, \omega, t) > 0$), FP denotes the number of false positives (i.e., when the target trace violates φ and the predicted flowpipe ω yields $\underline{\rho}(\varphi, \omega, t) > 0$), and FN denotes the number of false negatives (i.e., when the target trace satisfies φ and the predicted flowpipe ω yields $\underline{\rho}(\varphi, \omega, t) < 0$).

Table 1 shows F1 scores achieved by Bayesian LSTM predictions generated for the testing datasets using the optimal SRTs and dropout rates selected via different loss functions. L_{acc} has the worst results because it tends to over-estimate the uncertainty (i.e., making the flowpipe wider) in order to improve the prediction accuracy. Both L_{sat} and L_{qt} check

Model	L_{acc}	L_{sat}	L_{qt} (proposed)
Adults($\beta = 0.5$)	0.66	0.88	**0.93**
Adolescents($\beta = 0.5$)	0.38	0.60	**0.71**
Children($\beta = 0.5$)	0.68	0.81	**0.90**

Table 1: F1 scores of requirement satisfaction for comparing the proposed loss function with two baselines.

if the target trace and the predicated flowpipe yield consistent STL-U monitoring results. L_{qt} achieves higher F1 scores than L_{sat} in all three models, because L_{qt} accounts for quantitative information in the form of robustness degree of requirement satisfaction.

In summary, experiments demonstrate that the proposed loss function can be used to calibrate the uncertainty estimation of Bayesian RNNs by guiding the selection of optimal SRTs and dropout rates, which improve the quality of uncertainty estimates and predictions.

4.2 RQ2: Evaluation of Predictive Monitors

We compare the performance of the proposed STL-U quantitative predictive monitor with a baseline predictive monitor that does not account for the uncertainty. Specifically, we use the STL monitor (Maler and Nickovic 2004) to check the predicted flowpipes' mean traces as the baseline. For a fair comparison, both predictive monitors use the same Bayesian LSTM models equipped with the optimal SRTs and dropout rates. We consider two performance metrics: the average pre-alert time (i.e., the time interval between the earliest point when a predictive monitor detects an impending hazard and its actual occurrence time) over all possible safety hazards, and F1 score of requirement satisfaction. Following the convention in hazard calculations for artificial pancreas systems (Lum et al. 2021), we make the assumption that if two hazards of the same type happen closely within 30 minutes, then they are counted as one hazard.

		Baseline		Proposed	
Model	**Hazard**	**Time**	**F1**	**Time**	**F1**
Adults	Hypo	0.6	0.54	**23.9**	**0.96**
	Hyper	1.9	0.56	**22.2**	**0.63**
	Overall	1.2	0.54	**23.0**	**0.93**
Adolescents	Hypo	4.2	**0.57**	**24.9**	0.48
	Hyper	10.6	**0.82**	**22.6**	0.78
	Overall	9.2	**0.78**	**23.1**	0.71
Children	Hypo	3.9	0.89	**13.1**	**0.91**
	Hyper	21.2	0.71	**27.7**	**0.75**
	Overall	10.6	0.88	**18.7**	**0.90**

Table 2: Comparing the baseline and the proposed predictive monitors in terms of average pre-alert time (minutes) and F1 scores of requirement satisfaction.

Table 2 shows that, compared to the baseline, the proposed STL-U quantitative predictive monitor leads to earlier detection (i.e., larger pre-alert time) of impending hazards across all three patient populations with statistical significance. The results of paired t-test are (t(127)=28.2, p<0.01, d=3.2, significant) for adults, (t(316)=22.6, p<0.01, d=1.4, significant) for adolescents, and (t(354)=17.0, p<0.01, d=0.7, significant) for children. Table 2 also shows that the proposed predictive monitor achieves higher F1 scores than the baseline for adults and children models, indicating more accurate detection of impending hazards. But the proposed predictive monitor has slightly lower F1 scores than the baseline for the adolescents model. One possible explanation is that the simulated adolescent patients have higher glycemic variability index (1.96) than adults (1.69) and children (1.80), making the prediction more challenging.

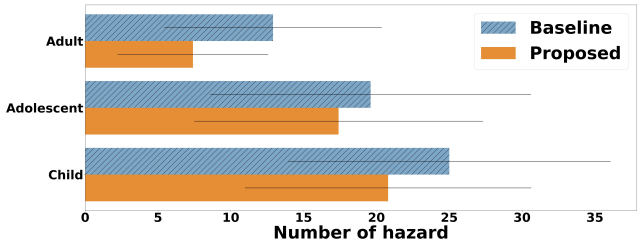

Figure 5: Comparing the number of hazards that occurred during the simulated 7-day period (mean and standard deviation over each patient population).

In summary, experiments demonstrate that the proposed STL-U quantitative monitor can provide early and accurate detection of impending safety hazards.

4.3 RQ3: Closed-Loop Simulation

We apply the proposed STL-U quantitative monitor and adaptive controller in the closed-loop simulation of T1D management. We use the simulator's Basal-Bolus Controller (Kovatchev et al. 2009) as a baseline for comparison.

We measure the following commonly used metrics for the safety and effectiveness of T1D management: *number of hazards* (i.e., number of hypoglycemias and hyperglycemias that occur to a patient during the simulated period) and *time in range* (i.e., the percentage of time that a simulated patient's BG levels stay within the range of 70-180 mg/dL).

Figure 5 shows that the proposed approach reduces the average number of hazards for all three types of patients. The results of paired t-test are (t(9)=-3.5, p=0.01, d=-0.9, significant) for adults, (t(9)=-1.9, p=0.09, d=-0.2, insignificant) for adolescents, and (t(9)=-1.0, p=0.33, d=-0.4, insignificant) for children.

Table 3 shows that the proposed approach improves the time in range across all three patient populations with statistical significance. The results of paired t-test are (t(9)=3.4, p=0.01, d=0.6, significant) for adults, (t(9)=2.7, p=0.02, d=0.3, significant) for adolescents, and (t(9)=2.6, p=0.03, d=0.4, significant) for children. We observe that adults have the most effective glucose control, while adolescents (*resp.* children) tend to have more hyperglycemias (*resp.* hypoglycemias) in the simulation.

In summary, results of the closed-loop simulation demonstrate that the proposed predictive monitoring and control approach improves the safety and effectiveness of T1D management by increasing the time in range and decreasing the

Model	Metric	Baseline	Proposed
Adults	Time in range	90.9%	**96.3%**
	Hypo time	4.8%	**2.4%**
	Hyper time	4.3%	**1.3%**
Adolescents	Time in range	80.9%	**85.2%**
	Hypo time	4.2%	**3.3%**
	Hyper time	14.9%	**11.5%**
Children	Time in range	63.6%	**74.8%**
	Hypo time	29.4%	**19.3%**
	Hyper time	7.0%	**5.9%**

Table 3: Comparing the baseline controller and the proposed approach's performance in closed-loop simulation.

number of safety hazards.

5 Related Work

Safe human-machine interaction via formal methods. Traditional methods of human-machine interaction design primarily rely on user studies for safety assurance (Sharp, Preece, and Rogers 2019). To reduce the burden of human testing, model-based design using formal methods has been explored to provide mathematically rigorous safety guarantees for human-machine interaction (Bolton, Bass, and Siminiceanu 2013). For example, a control protocol for an unmanned aerial vehicle (UAV) is synthesized by modeling the human-UAV interaction as a two-player stochastic game (Feng et al. 2016). There are also several works on the formal specification and verification of human-robot interaction (Luckcuck et al. 2019).

However, these existing works mostly focus on modeling and analyzing human-machine interaction at design time. By contrast, in this work, we consider the predictive monitoring and control of safe human-machine interaction at runtime. We use data-driven RNN models as abstract representations of complex human behaviors, capturing their uncertainty through Bayesian deep learning. We then develop logic-based approaches to monitor predictions made by Bayesian RNNs and adapt control actions based on the monitoring results.

Logic-based predictive monitoring and control. Temporal logic specifications such as Signal Temporal Logic (STL) have been used for runtime monitoring (also called runtime verification) and control of cyber-physical systems, such as automobiles and medical devices (Bartocci et al. 2018). Recently, there have been increasing efforts on predictive monitoring, which checks predictions about future states to support prompt decision making. For instance, (Qin and Deshmukh 2020) apply an STL monitor to check glucose levels predicted by an ARIMA statistical model. (Yoon and Sankaranarayanan 2021) presents a logic-based Bayesian intent inference to forecast a robot's future positions and avoid impending collisions. (Ma et al. 2021) presents an STL-U predictive monitor to check predictions of air pollution and traffic in smart cities.

Following this line of work, we contribute to the state-of-the-art by developing a novel STL-U *quantitative* predictive monitor that computes *robustness degrees* indicating how far a signal is from satisfying or violating a logic specification. Our work is inspired by (Fainekos and Pappas 2009), which introduces the concept of robustness degree for checking continuous-time (single-sequence) signals against temporal logic specifications. We adopt this concept and extend it for checking flowpipe signals that represent uncertain sequential predictions made by Bayesian RNNs.

There is some related work on uncertainty-aware STL monitoring. For example, (Baharisangari et al. 2021) define the lower and upper bounds of robustness by selecting a single trace out of an interval trajectory. (Visconti et al. 2021) define lower and upper bounds for specific time points with missing values for a single trace. (Lindemann et al. 2023) construct prediction regions that quantify prediction uncertainty using conformal prediction, a statistical tool for uncertainty quantification. By contrast, we define STL-U quantitative semantics over all possible traces included in a flowpipe, capturing the uncertainty of Bayesian RNNs via confidence intervals at every time point.

Lastly, existing work on utilizing predictive monitoring results for downstream tasks such as control and design modification is limited. This work addresses this gap by developing a new loss function to guide the uncertainty calibration of Bayesian deep learning and a new adaptive control method, both of which leverage STL-U quantitative predictive monitoring results.

6 Conclusion

In this work, we present a logic-based quantitative predictive monitoring and control approach to enhance the safety of human-machine interaction under uncertainty. Using Bayesian RNN models, we represent the uncertainty of human behavior and employ a novel STL-U quantitative predictive monitor to compute robustness degree intervals, which indicate the satisfaction or violation of STL-U requirements. We design a new loss function leveraging STL-U results to optimize uncertainty estimation in Bayesian RNNs by selecting the best combination of SRT and dropout rate. Additionally, adaptive controllers adjust control actions based on robustness intervals.

Experiments with a T1D patient simulator demonstrate that the proposed approach enables early and accurate detection of safety hazards and improves the safety and effectiveness of T1D management. Furthermore, the loss function outperforms state-of-the-art baselines in uncertainty estimation. Results from a semi-autonomous driving case study also show enhanced safety, confirming the approach's generalizability.

Future work includes testing alternative RNN models, developing principled adaptive control for broader domains, incorporating priority-based dynamic enforcement of requirements, validating in real-world settings, and extending to diverse case studies like human-robot interaction.

Acknowledgments

This work was supported in part by the U.S. Air Force Office of Scientific Research under Grant FA9550-21-1-0164 and

the U.S. National Science Foundation under Grants CCF-1942836, CCF-2131511, 2220401, and 2427711. The opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

References

American_Diabetes_Association. 2022. Standards of Medical Care in Diabetes—2022. *Diabetes Care*, 45: S17.

Baharisangari, N.; Gaglione, J.-R.; Neider, D.; Topcu, U.; and Xu, Z. 2021. Uncertainty-Aware Signal Temporal Logic Inference. In *Software Verification*, 61–85. Springer.

Banks, V. A.; Plant, K. L.; and Stanton, N. A. 2018. Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety science*, 108: 278–285.

Bartocci, E.; Deshmukh, J.; Donzé, A.; Fainekos, G.; Maler, O.; Ničković, D.; and Sankaranarayanan, S. 2018. Specification-based monitoring of cyber-physical systems: a survey on theory, tools and applications. In *Lectures on Runtime Verification*, 135–175. Springer.

Bengio, Y.; Goodfellow, I.; and Courville, A. 2017. *Deep learning*. MIT press Cambridge, MA, USA.

Bolton, M. L.; Bass, E. J.; and Siminiceanu, R. I. 2013. Using formal verification to evaluate human-automation interaction: A review. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(3): 488–503.

CARLATeam. 2023. CARLA 0.9.13 Release. https://carla.org/2021/11/16/release-0.9.13/. Accessed: 2025-02-03.

Dong, S.; Ma, M.; Lamp, J.; Elbaum, S.; Dwyer, M. B.; and Feng, L. 2024. Quantitative Predictive Monitoring and Control for Safe Human-Machine Interaction. arXiv:2412.13365.

Fainekos, G. E.; and Pappas, G. J. 2009. Robustness of temporal logic specifications for continuous-time signals. *Theoretical Computer Science*, 410(42): 4262–4291.

Feng, L.; Wiltsche, C.; Humphrey, L.; and Topcu, U. 2016. Synthesis of human-in-the-loop control protocols for autonomous systems. *IEEE Transactions on Automation Science and Engineering*, 13(2): 450–462.

Gal, Y. 2016. *Uncertainty in deep learning*. Ph.D. thesis, University of Cambridge.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Kovatchev, B. P.; Breton, M.; Dalla Man, C.; and Cobelli, C. 2009. Biosimulation Modeling for Diabetes: In Silico Preclinical Trials: A Proof of Concept in Closed-Loop Control of Type 1 Diabetes. *Journal of diabetes science and technology (Online)*, 3(1): 44.

Lindemann, L.; Qin, X.; Deshmukh, J. V.; and Pappas, G. J. 2023. Conformal prediction for STL runtime verification. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, 142–153.

Luckcuck, M.; Farrell, M.; Dennis, L. A.; Dixon, C.; and Fisher, M. 2019. Formal specification and verification of autonomous robotic systems: A survey. *ACM Computing Surveys (CSUR)*, 52(5): 1–41.

Lum, J. W.; Bailey, R. J.; Barnes-Lomen, V.; Naranjo, D.; Hood, K. K.; Lal, R. A.; Arbiter, B.; Brown, A. S.; DeSalvo, D. J.; Pettus, J.; et al. 2021. A real-world prospective study of the safety and effectiveness of the loop open source automated insulin delivery system. *Diabetes technology & therapeutics*, 23(5): 367–375.

Ma, M.; Gao, J.; Feng, L.; and Stankovic, J. 2020. STLnet: Signal temporal logic enforced multivariate recurrent neural networks. *Advances in Neural Information Processing Systems*, 33: 14604–14614.

Ma, M.; Stankovic, J.; Bartocci, E.; and Feng, L. 2021. Predictive monitoring with logic-calibrated uncertainty for cyber-physical systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s): 1–25.

Ma, M.; Stankovic, J. A.; and Feng, L. 2021. Toward formal methods for smart cities. *Computer*, 54(9): 39–48.

Maler, O.; and Nickovic, D. 2004. Monitoring temporal properties of continuous signals. In *Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems*, 152–166. Springer.

Man, C. D.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; and Cobelli, C. 2014. The UVA/PADOVA type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1): 26–34.

Qin, X.; and Deshmukh, J. V. 2020. Clairvoyant Monitoring for Signal Temporal Logic. In *International Conference on Formal Modeling and Analysis of Timed Systems*, 178–195. Springer.

Sharp, H.; Preece, J.; and Rogers, Y. 2019. *Interaction design: beyond human-computer interaction*. NY: Wiley.

Slattery, D.; Amiel, S.; and Choudhary, P. 2018. Optimal prandial timing of bolus insulin in diabetes management: a review. *Diabetic Medicine*, 35(3): 306–316.

Visconti, E.; Bartocci, E.; Loreti, M.; and Nenzi, L. 2021. Online monitoring of spatio-temporal properties for imprecise signals. In *Proceedings of the 19th ACM-IEEE International Conference on Formal Methods and Models for System Design*, 78–88.

Yang, S.; Zhong, Y.; Feng, D.; Li, R. Y. M.; Shao, X.-F.; and Liu, W. 2022. Robot application and occupational injuries: are robots necessarily safer? *Safety science*, 147: 105623.

Yoon, H.; and Sankaranarayanan, S. 2021. Predictive runtime monitoring for mobile robots using logic-based bayesian intent inference. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 8565–8571. IEEE.