IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 11, 2021, accepted July 22, 2021, date of publication July 26, 2021, date of current version August 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100007

A Blood Glucose Control Framework Based on Reinforcement Learning With Safety and Interpretability: In Silico Validation

MIN HYUK LIM[1], (Graduate Student Member, IEEE), WOO HYUNG LEE[2], BYOUNGJUN JEON[3], AND SUNGWAN KIM[1,4], (Senior Member, IEEE)

[1]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul 03080, South Korea
[2]Department of Rehabilitation Medicine, Seoul National University Hospital, Seoul National University College of Medicine, Jongno-gu, Seoul 03080, South Korea
[3]Interdisciplinary Program in Bioengineering, Graduate School, Seoul National University, Seoul 03080, South Korea
[4]Institute of Bioengineering, Seoul National University, Seoul 03080, South Korea

Corresponding author: Sungwan Kim (sungwan@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) by the Korean Government under Grant 2018M1A3A3A02065779.

ABSTRACT Controlling blood glucose levels in diabetic patients is important for managing their health and quality of life. Several algorithms based on model predictive control and reinforcement learning (RL) have been proposed so far, most of which use prior knowledge of physiological systems, the mathematical structure of blood glucose dynamics, and many episodes including failures for training the policy network in RL. To be smoothly adopted in clinical settings, we propose a fast online learning method underlining safety and interpretability. A random forest regressor and a dual attention network were exploited for glucose prediction and extension of state variables. The soft actor-critic network to determine insulin dosing was guided by proportional-integral-derivative (PID) control in the early phase, and an adaptive safe actor with suspension and additional insulin dosing was incorporated. The performance of the models was validated using an FDA-approved type 1 diabetes simulator. The results showed comparable outcomes with PID control. Using this system, glucose dynamics could be captured despite minimal prior knowledge. The extended state variables were correlated with basic states such as glucose, insulin, and meal intake, their derivatives, and their integrals, which can be fundamental elements of mathematical modeling of physiological responses. Attention scores and attribution scores in the prediction and control models represented the focused features and the internal operation of the models with interpretability. We expect this study to provide some insights on how RL can be practically adopted in clinical environments and how interpretability can provide hints of machines' thoughts for clinical applications.

INDEX TERMS Blood glucose control, reinforcement learning, safe and interpretable control, in silico validation, simulation for clinical application.

I. INTRODUCTION

Regulation of blood glucose levels in diabetic patients is critical for managing their health. Diabetes mellitus is a chronic disorder of improper glucose metabolism that induces complications such as cardiovascular disease [1], and neuropathy [2].

It is estimated that 463 million people worldwide suffer from diabetes [3] and the number of patients is going to increase. Therefore, prevention [4] and care are important. Technology for uncovering the characteristics of individuals with diabetes and suggesting interventions to prevent exacerbation of clinical stages would improve the quality of life and health of patients.

The interaction between glucose and insulin is complex. Several physiological and mathematical models [5]– [8]. have been proposed. The operations of numerous organs related

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

to glucose homeostasis have been depicted by compartment modeling [9] and estimated by measurements through blood sampling, oral tracers [10], [11], and glucose clamp techniques [12], [13]. Clinical indexes including insulin sensitivity [14], [15] and glucose responsivity [16], [17] differentiate groups of diabetic patients to consider treatment plans. The model parameters are related to clinical indexes; therefore, incorrect assumptions regarding the structure of models may lead to biases and inaccurate clinical descriptions.

Patients suffering from type 1 diabetes cannot produce insulin; thus, external injections are required to regulate blood glucose levels. The calculation of insulin dosing [18] should be carefully considered because hypoglycemia and hyperglycemia are common complications in the treatment. In particular, insulin on board (IOB), that is the previous bolus delivered and still active [19], may induce excessive action of insulin to lower blood glucose levels toward hypoglycemia. Continuous glucose monitoring (CGM) [20]–[23] and artificial pancreas [24], [25] have been widely tested in clinical trials and have been adopted in daily life. In comparison to manual calculation and delivery of insulin doses, automated insulin pumps with sensors for measuring blood glucose levels can provide convenience and efficiency upon proper responses to time-varying human conditions. Safety is the most important value [24], [26], [27] for adopting a technology in clinical applications, and all algorithms should be carefully considered and tested.

Simulator-based approaches [28]–[30] cannot provide a perfect validation for clinical studies, but the characteristics, extent of applications, and potential weaknesses of algorithms can be evaluated as preparation for experiments in vivo. Physiological models have been adopted to simulate glucose dynamics, and several control algorithms based on proportional–integral–derivative (PID) control [31], [32], model predictive control (MPC) [33]– [35], and reinforcement learning (RL) [36]–[39] have been tested for validation before clinical trials. Insulin is the main control action in type 1 diabetes, whereas glucagon has been recently used to compensate for the action of insulin in the artificial pancreas, and oral anti-diabetic medications can alternatively be used in type 2 diabetes.

Many simulators use mathematical models to capture dynamics. If the prediction and control models adopt this knowledge to design the internal structure of the models, information on a series of simplified equations is shared, both in the controller and testing platforms. This may induce an over-fitting problem with seemingly high performance in simulation compared with the actual trials of hidden and unobserved dynamics in the body. For example, if a controller and a simulator use the same mathematical equations for physiological descriptions, then the controller will show a good performance in simulations, even though the actual natural functionality may be different from the simplified equations in the controller and the simulator. In this study, we considered clinically applicable approaches without distinguishing between simulation and reality. We attempted to construct prediction and control models with as little prior knowledge of physiology as possible to reduce bias in simulation testing due to shared knowledge of the information of prior models.

RL has successfully achieved goals in numerous problems with complex tasks, from playing games [40] to manipulating robots [41]. The RL framework based on deep learning is flexible for estimating parameters and structures in policy, value function, and Q-function neural networks, model-free and model-based approaches, and application from simulation to the real world. It does not require full knowledge of the model and the dynamics in advance, and trial-and-error testing should be conducted during exploration and exploitation [42]. Moreover, if the RL agent can learn time-varying models online, then RL can respond to the changes and trends of dynamics in glucose regulation in the short and long term. Thus, RL approaches can be expanded to numerous clinical problems with unclear or partial knowledge of internal dynamics and interactions in vivo to generate actions for regulating some biological variables or conditions.

However, some limitations exist for applying RL based on neural networks to clinical trials for diabetes management. First, RL usually requires many episodes for training models. RL agents have to experience successes and failures to shape the information from trial-and-error episodes to determine a good policy for actions. However, patients should not be subjected to health failures such as hypoglycemia or hyperglycemia. Second, clinicians should be able to intervene with controllers and patients when the situation is predicted incorrectly and be provided with some clues about the internal states of patients and controllers. Third, many RL studies also focus on the reduction of dimensions to construct latent states dealing with raw image pixels and/or spatio-temporal data. In contrast, glucose dynamics have unobserved hidden states and interactions with relatively simple observed variables such as glucose levels, insulin levels, and amount of meal intake.

To resolve these difficulties, the present study has the following three major contributions:

1) First, the adoption of PID control as a guiding policy in the early phase and the introduction of a safe actor make training periods shorter than those of other RL methods for glucose regulation. Similar to clinical environments, an adaptive actor for safety to allow patients to escape harmful conditions adjusts the insulin dosing based on soft actor-critic (SAC) control in one continuous episode.
2) Second, the dual attention network (DAN) for forecasting glucose levels can extend the basic states reflecting temporal contexts in dynamics through encoding and decoding processes. The extended states contain interpretable physiological information that can be derived from the transformation of the basic states. It does not require prior knowledge of mathematical equations to describe glucose dynamics.

3) Third, the prediction and control models were investigated from the perspective of interpretability. The internal operations can be explained based on the inherent structure of the models and the ad hoc methods used to analyze neural network-based models. Patients, clinicians, and control designers can understand the behavior of the models by observing how the models act under various conditions and prevent malfunctioning events.

Validation from simulations is not a perfect solution for testing algorithms. However, it can provide insights for further applications in clinical environments. With fewer assumptions, the framework can be flexibly adapted for hidden complex dynamics and is not limited to glucose regulation but can be adapted for several dynamics with biological, physiological, and clinical variables in a time-varying fashion.

II. RELATED WORKS

A. DESCRIPTION OF GLUCOSE DYNAMICS

To describe glucose dynamics in the human body, a series of ordinary or partial equations based on compartment modeling [6]– [9] has been proposed. Several parameter estimation techniques, such as deconvolution [43], [44] and Bayesian inference [45], have been exploited for acquiring accurate coefficients for models. Measurements of various biological variables, including glucose, insulin, C-peptide, and glucagon [46], have been used to expand the structure of the models. In addition to blood sampling, radioactive agents [10], [11] were used to validate the assumption of internal interactions. Glucose clamp techniques [12], [13] have been used to observe the dynamics based on the equilibrium of the phenomena.

The major variables used for modeling are the glucose level of blood or subcutaneous tissues [47], insulin injected or concentrated in blood, and information on meal intake. In addition, glucagon, which has the opposite actions toward insulin, C-peptide, which is the precursor of insulin, and incretin [48], which stimulates insulin secretion, can be included in the modeling of glucose dynamics in the body.

Model identification [49], [50] is a common issue for properly capturing patient characteristics. Infeasible ranges in states may lead to unstable or unreal conditions; thus, accurate measurements and modeling based on physiological knowledge are important. Individuals have the ability to stabilize blood glucose levels, and clinical indexes for differentiating status as normal, pre-diabetes, and diabetes can be derived.

According to a population-based study, mathematical models can support clustering of subjects' groups and organization of virtual clinical trials from the distributions of parameters calculated from actual clinical trials. The UVA/Padova type 1 diabetes simulator [29], [30] is such an effort in which many control algorithms have been tested. Insulin sensitivity, the extent to which the body reacts to given insulin to regulate blood glucose levels, and glucose responsivity, the proxy for how glucose affects the production levels of insulin secretion, can be derived from the observation of the body's responses and mathematical models. For instance, the HOMA [51] and Matsuda [52] indices are widely used in the clinical field and can be obtained using simple calculations. A more complex model cannot ensure accurate evaluation of disease status. However, it can provide more minute differentiation of patient characteristics based on parameters of the model and an opportunity to predict time-varying sequences of related variables in glucose dynamics.

B. FORECASTING OF GLUCOSE LEVELS

Forecasting the state is useful for clinicians and diabetic patients because it can provide opportunities to prepare interventions to cope with undesirable events such as hypoglycemia, which may induce unconsciousness in patients. If the responses to the oral intake of carbohydrates are predicted, dietary plans can be made to reduce hyperglycemia.

Many prediction models use several variables to construct states based on the sequences of data. This process depends on how the states are defined or input into the models. If the model is established based on state-space models, Kalman filtering [53], [54] or other correction methods are usually incorporated to estimate the precise states repeatedly. Machine learning-based regression algorithms [55] using random forests, support vector machines, and gradient boosting can also be adopted for forecasting. Nowadays, neural network-based approaches [56], [57] have received attention based on recurrent neural networks and auto-encoders to represent sequential data. Hybridization methods [58] connect mathematical models of differential equations and variational autoencoders to describe the glucose dynamics of diabetic patients. Some studies [54], [57], [58] have used physiological knowledge such as the glucose appearance rate for data efficiency.

Deep learning-based forecasting methods are more flexible for time-series data. Neural networks can represent non-linear functions according to the universal approximation theorem. Causal and dilated convolution layers to capture temporal changes in glucose levels were adopted in [57]. Recurrent neural networks take sequences as inputs with relatively fewer assumptions than other machine learning-based methods. A model consisting of encoders and decoders with multiple outputs was proposed in [56]. Recently, the attention mechanism [59] has been broadly adopted in several algorithms for processing sequential data to improve performance and to provide explainability. It can also highlight the rationale of the machine's internal calculations. In this study, a dual-attention network [60] was adopted. It uses two attention modules to focus on the importance of variables and temporal sequences, which can provide attention scores for variables and times.

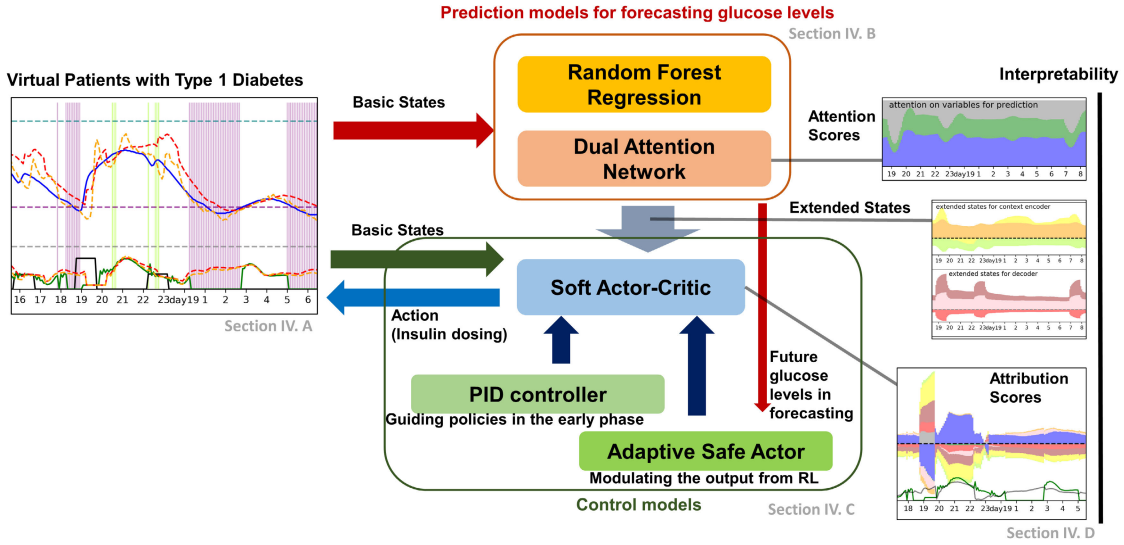

FIGURE 1. The pipeline of our framework for regulating glucose levels is presented in this figure. Prediction models to forecast blood glucose levels and control models to regulate blood glucose levels are connected to interact with a type 1 diabetes mellitus simulator in feedback cycles. The dual attention network generates extended states from basic states. Attention scores for the dual attention network and attribution scores for the soft actor-critic provide interpretable metrics.

C. CLASSICAL CONTROL FOR GLUCOSE REGULATION

State-space models can be used to establish controls using classical control theories. Insulin is a variable that is inherent in many mathematical models for describing glucose dynamics and can be used as an external control.

Several approaches to maintain proper glucose levels in the body, including the linear quadratic regulator [61], dynamic programming, and closed-loop control [62], [63], have been proposed in many studies. These control methods use mathematical models of glucose dynamics, and a small number of samples is required for data efficiency. PID control is also comparable [64] to model-based control in terms of performance, and PID with adaptive weights [65] has been proposed recently. In [66], linear parameter-varying control was incorporated with insulin, food intake, and metabolism subsystems. In addition, the controller's learning was emphasized in control systems [67] to improve control performance by considering the variation in glucose dynamics.

The strength of classical control approaches is data efficiency and ease of deriving closed-form solutions to enhance stability analysis. Control actions can be designed and implemented explicitly with low computation time. Traditional tools for control systems and characteristics, including observability and controllability, can be considered. However, because of the rigid structure of mathematical models, accurate modeling to differentiate hidden dynamics and noise is required.

D. REINFORCEMENT LEARNING FOR DIVERSE FIELDS

Reinforcement learning (RL) has been widely applied to diverse tasks, from solving video games [40] to robotic actions, in constrained environments, and for the optimization of the structure of the neural network. Most RL studies have models for representation learning [68] to extract state representations from raw inputs and pursue end-to-end learning. Fully connected layers, convolutional layers, hidden Markov models, and encoders from autoencoders are examples of feature construction to establish states for RL.

For continuous actions in RL, deep deterministic policy gradient (DDPG) [69] and SAC [70] are efficient solutions. Experience replay [71] and the world model [68] can reduce bias in training policy functions and the number of episodes to be experienced.

RL with safe constraints [72] has emerged to consider applications in the real world. Some studies have refined the constraints or rewards for safe ranges in states, which can help in avoidance of unstable areas. Otherwise, the methods in which policies are trained separately can be applied under normal conditions for conducting tasks and escaping conditions for safety. If the states are classified as unsafe, the policy is switched to ensure safety.

In this study, an adaptive actor for safety was adopted to modulate the insulin dose calculated from the SAC controller. It provides an alternative policy when the glucose level is predicted to be in an unsafe range.

E. REINFORCEMENT LEARNING FOR REGULATION OF GLUCOSE

Numerous studies [73] have adopted RL for the determination of insulin dosing for continuous and discrete values. Physiological models with prior knowledge of glucose dynamics are typically used to design rewards and responses.

Features of hypoglycemia and hyperglycemia events [36], [74] have been used to adapt a bolus based on actor-critic methods. Some rule-based conditions need to be combined, and even explainability can be imposed on the trained

controller. RL-based controllers have the advantage of flexibility if model-free assumptions are pursued; thus, the forms of functions on policy and value are important, which can be from linear combination [75] to deep learning. Some studies have implemented neural networks for deep reinforcement learning to reduce the number of episodes to be trained [37], [39]. For example, transfer learning from a generalized controller with fine-tuning has been proposed to accelerate adaptation to the characteristics of individuals glucose dynamics.

In this study, actor-critic methods consisting of deep learning with a safe actor using features of hypo and hyperglycemia were adopted from the perspectives of reinforcement learning and switching controllers for safety. Actor-critic methods have been widely implemented [39], [70], [72]–[74] in many RL applications.

F. INTERPRETABILITY OF NEURAL NETWORKS

Interpretable models and methods can provide some hints to explain the models. Models can inherently have modules and structures for interpretation or can be analyzed via ad hoc approaches. A model itself can be explained by internal parameters and weights of neural networks, or the features that can affect the outcome of the models can be explored by gradient-based or perturbation-based methods. To date, there is no gold standard for calculating the contribution scores of input features to outputs in a model. It depends on the characteristics of learning and inference, such as back and forward propagation, gradient and integral, perturbation, and boosting. However, although metrics for interpretability sometimes require additional interpretation by humans, attention and attribution scores can provide some hints for the internal operation of the model.

Interpretability has recently been emphasized in the clinical field. Both improvements in accuracy and explanations of why false-positive and false-negative outputs occur are invaluable for improving models for classification problems common in disease detection.

Deep learning important features (DeepLIFT) [76] is a gradient method for calculating attributions from the inputs. Layer-wise relevance propagation (LRP) [77], [78] and the integrated gradients (IG) [79] are alternative approaches for understanding contribution features of neural networks, which can be unified by means of composition and setting reference values.

III. METHODS OF PREDICTION, CONTROL, AND INTERPRETABILITY

This study includes the following three major parts:

1) First, prediction models forecast future glucose levels and generate extended states from basic states to reflect context information. Prediction models are based on random forest regression (RFR) and a dual attention network (DAN). Attention scores show which variables and sequences the DAN is focusing on.
2) Second, insulin dosing is obtained from the soft actor-critic (SAC) policy network. The policy of SAC is guided by a PID controller in the early phase of the simulations and modulated by an adaptive safe actor.
3) Third, extended states are investigated based on correlations with simple forms of basic states. In this study, simple forms of a variable are defined as the variable itself, the first and second derivatives of the variable, and the integrated value of the variable, which can be interpretable and meaningful from the perspective of physiological responses described by mathematical equations. In addition, the attribution scores of extended states in SAC provide a partial explanation of the behavior of SAC.

Prior knowledge and static information regarding the complex physiological structure of virtual patients are minimally assumed for model construction. Learning and actions are concurrently conducted online to reflect the variability of the patient's status by updating the parameters repeatedly. The conditions and responses (e.g., insulin sensitivity and glucose responsivity) of dynamics in patients vary from moment to moment; thus, models should be updated frequently.

PID control was fully introduced to the action in the beginning, and the portion of the action from PID control was gradually replaced by an SAC policy over time. PID control is a type of model agnostic control method; thus, it coincides with the assumption of minimal knowledge of physiology. Fig. 1. and Fig. 2. show the overall scheme of the proposed framework.

Algorithm 1 Simultaneous Prediction and Control

```
initialize parameters of all models
while (scenario is valid) do
    receive basic states
    update parameters of all models
    predict the next states
    generate extended states from basic states
    calculate actions from PID, SAC, or a mixed policy of
    PID and SAC
    if (hypoglycemia is predicted to occur) then
        conduct suspension by the safe actor
    else
        conduct actions from PID, SAC, or a mixed policy of
        PID and SAC
        if (hyperglycemia is predicted to occur) then
            conduct additional dosing by the safe actor
        end if
    end if
end while
```

A. PREDICTION MODELS

For glucose and insulin prediction in the future states, two predictive models were adopted, which are based on random forest regression (RFR) and DAN. Blood glucose level,

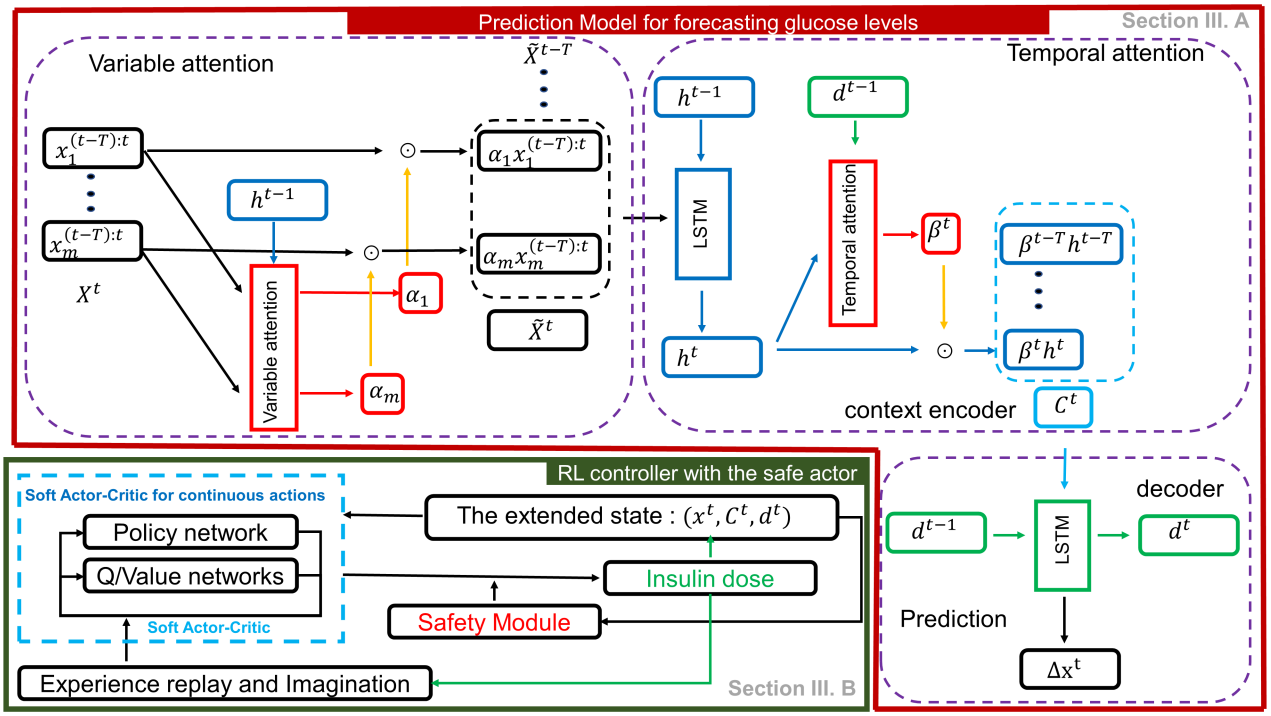

FIGURE 2. A detailed structure of prediction and control models for blood glucose forecasting and regulation were proposed. Attention scores for variables and temporal relationships were obtained through the encoder and decoder. States were extended with hidden variables and context vectors from the dual-attention network to predict changes in the states. The SAC with adaptive safe actors in calculating insulin doses uses the extended states for learning and actions.

insulin dosing for the virtual pump, and the amount of meal intake are the basic states for prediction. Notably, insulin calculated by the control models, which corresponds to the insulin dose from the pump, is used, whereas the IOB inside the patient's body is not explicitly structured in prediction models to intentionally reduce the bias on internal complex dynamics of physiologic variables. In addition, the DAN is used not only for prediction but also for the encoder and decoder to generate extended states from basic states (glucose, insulin, and meal intake). The extended states are inputs for the RL algorithms in this study.

We use a DAN for forecasting and extending the states. Basic states $x^t = (x_1^t, \ldots, x_m^t) \in \mathbb{R}^m$ at the current time t are supposed to partially reflect the current state because of the lack of context. A sequence of basic states $X^t = (x^{t-T_w}, \ldots, x^t)$ with time window T_w is input to the prediction models. $\tilde{X}^t$ is the modified sequence weighted by the attention coefficient, α^t.

$$\tilde{X}^t = \sum_{i=1}^{m} \alpha_i^t \mathbf{x_i^t} \tag{1}$$

$$\begin{aligned} output_{enc}, h^t &= f_{enc}(h^{t-1}, \tilde{X}^t) \\ output_{dec}, d^t &= f_{dec}(d^{t-1}, \tilde{X}^t) \end{aligned} \tag{2}$$

Encoder f_{enc} and decoder f_{dec} have the normal forms of long short-term memory (LSTM), and the hidden states of the encoder and decoder, respectively, and h^t and d^t are concatenated to construct attention coefficients α_i^t, β_i^t in attention modules. In this encoder-decoder connection based on LSTM, $output_{dec}$ is expected to be $\Delta x^t = x^{t+1} - x^t$ in training and prediction.

$$\begin{aligned} e^t &= \mathbf{v}_e^T tanh(\mathbf{W}_e[\mathbf{h}^{t-1}; x^{t-1}]) \\ g^t &= \mathbf{v}_d^T tanh(\mathbf{W}_d[\mathbf{d}^{t-1}; h^{t-1}]) \\ \alpha_k^t &= \frac{exp(e_k^t)}{\sum_{i=1}^{m} exp(e_i^t)}, \beta_k^t = \frac{exp(g_k^t)}{\sum_{j=1}^{T_w} exp(g_j^t)} \end{aligned} \tag{3}$$

With the attention coefficients obtained from attention modules for variables and temporal sequences, the context vector C^t is the weighted sum of the encoder hidden states h^t in the time window T_w.

$$C^t = \sum_{j=1}^{T_w} \beta_j^t \mathbf{h^j} \tag{4}$$

The objective of forecasting is the change in the basic states $\Delta x^t = x^{t+1} - x^t$, and the mean squared error between the predicted and actual changes in basic states is used as the loss function to be minimized during training.

$$loss_{pred} = \frac{1}{N} \sum_{n=1}^{N} (output_{dec} - \Delta x^t)^2 \tag{5}$$

Extended states for control models are made up at the current time t by concatenating the basic state, context vector, and decoder hidden state, $X_{ext}^t = (X^t, C^t, d^t)$.

Two forecasting models, RFR and DAN, conduct predictions in the time window T_{safe}. If the glucose level after T_{safe} is predicted to be lower than G_{susp}, then a suspension of insulin dosing and/or additional oral carbohydrate is considered. If the glucose level after T_{safe} is higher than G_{int}, then an additional insulin dose from the action of the SAC and/or PID controller is administered to the patients.

B. CONTROL MODELS

SAC networks and adaptive actions for safety are combined for blood glucose regulation. As explained above, insulin in the basic states at the current time is also an action from the control models including PID and/or RL with an adaptive safe actuator at the previous time. Thus, three distinct points exist in this study.

First, some states and actions are explicitly connected via insulin, in contrast to many RL problems. The history of actions directly defines the portion of states, which means that the prediction for states should directly consider the current policy of control because states are affected by how actions would be delivered to the patient. Second, actions including insulin and oral carbohydrates cannot have negative values, so exploration in RL should be effectively conducted because the range of actions is limited. Owing to delayed responses in the dynamics, prediction models and controllers should take time to observe the delayed effects of the action at the current time and the equilibrium states to be restored. Third, the RL control and actuator for safety compensate for each other. Predefined actions and rules to rescue the patient from hypo and hyperglycemia are adopted, and the thresholds of glucose levels to invoke the actor for safety are tuned according to the portion of hyper/hypoglycemic events that occurred.

In the normal range of glucose levels, the SAC determines the dosage of insulin by the virtual pump. For policy learning, $s_t = X^t_{ext} \in S$ is the state (i.e., extended state), $a_t \in A$ is the action (i.e., insulin dosing), the state transition probability from the current state to the next state with action a is p (i.e., transition probability from prediction models), and the reward is r. Following the general presentation of RL, the state value function $V_\psi(s_t)$, soft Q-function $Q_\theta(s_t, a_t)$ and policy $\pi_\phi(a_t|s_t)$ are parameterized by neural networks with parameters ψ, θ, and ϕ. In the training process, $\bar{\psi}$, the exponentially moving average of ψ is used for stabilization.

To update ψ and θ by using gradients, the squared residual errors J_Q, J_V to be minimized are introduced as in the original research of SAC. In replay buffer D, the tuple of (s_t, a_t, s_{t+1}) is extracted for sampling. The squared residual error of the soft value function, $J_V(\psi)$, is expressed as follows:

$$J_V(\psi) = \mathbb{E}_{s_t\sim D}[(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi_\phi}[Q_\pi(s_t, a_t) - log\pi_\phi(a_t|s_t)])^2] \tag{6}$$

and, the soft Bellman residual for Q-function, $J_Q(\theta)$ is,

$$J_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim D}[\frac{1}{2}(Q_\theta(s_t, a_t) - \hat{Q}_\theta(s_t, a_t))^2] \tag{7}$$

To obtain $\hat{Q}_\theta(s_t, a_t)$, we consider not only the next state but also the time window T_Q with decay coefficient γ, because the insulin action has delayed effects on the states including glucose level, and future states of long time horizon should be considered.

$$\hat{Q}_\theta(s_t, a_t) = \sum_{i=0}^{T_Q} \gamma^i r(s_{t+i}, a_{t+i}) + \gamma^{T_Q+1}\mathbb{E}_{s_{t+T_Q}\sim\hat{D}}[V_{\bar{\psi}}(s_{t+T_Q})] \tag{8}$$

In the above equation, instead of the actual buffer D, tuples generated from the connection between the prediction and control models are included in the augmented buffer $\hat{D}$, which is the imagination of the RL agent for virtual experiences in the prediction horizon from t to $t+T_Q$. From the current state s_t and action a_t, $\hat{s}_{t+1}$ is predicted. Then, the next action $\hat{a}_{t+1}$ is given by policy $\pi_\phi(\hat{a}_{t+1}|\hat{s}_{t+1})$. This imagination process is repeated until reaching time $t + T_Q$.

To train the policy network, the expectation of Kullback-Leibler (KL) divergence with $Q_\theta(s_t, \cdot)$ and partition function $Z_\theta(s_t)$ for normalizing the distribution are given as follows:

$$J_\pi(\phi) = \mathbb{E}_{s_t\sim D}[D_{KL}(\pi(\cdot|s_t) \parallel \frac{exp(Q_\theta(s_t, \cdot))}{Z_\theta(s_t)})] \tag{9}$$

Algorithm 2 Soft Actor-Critic With Imagination

```
initialize parameters ψ, ψ̄, θ, φ
for each iteration do
  for each environment step do
    a_t ~ π_φ(a_t|s_t)
    s_{t+1} ~ p_env(s_{t+1}|s_t, a_t)
    D ← D ∪ (s_t, a_t, r(s_t, a_t), s_{t+1})
    D̂ ← D
    ŝ_{t+1} ← s_{t+1}
    for each imagination step, i = 1 ... T_Q do
      â_{t+i} ~ π_φ(â_{t+i}|ŝ_{t+i})
      ŝ_{t+i+1} ~ p_pred(ŝ_{t+i+1}|ŝ_{t+i}, â_{t+i})
      D̂ ← D̂ ∪ (ŝ_{t+i}, â_{t+i}, r(ŝ_{t+i}, â_{t+i}), ŝ_{t+i+1})
    end for
  end for
  for each parameter update step do
    update ψ by gradient of J_V(ψ)
    update θ by gradient of J_Q(θ)
    update φ by gradient of J_π(φ)
    ψ̄ ← τψ + (1 − τ)ψ̄
  end for
end for
```

If the condition satisfies the safe actor being recalled as indicated in the forecasting model section, then the action from the policy from RL and/or PID controls is suspended or modified. To determine the condition, future glucose levels from the two forecasting models are averaged or considered to elevate the sensitivity to detect future events.

The threshold for suspension, $G_{susp} \in [G^{min}_{susp}, G^{max}_{susp}]$, and the threshold for additional insulin dose, $G_{int} \in [G^{min}_{int}, G^{max}_{int}]$

Algorithm 3 Adaptive Safe Actor for Safety

initialize parameters G_{susp}, G_{int}
for each environment step **do**
  at the current time t
  prediction of the glucose level at the time $t + T_{safe}$
  if $G_{pred}^{t+T_{safe}} < G_{susp}$ **then**
    suspension of insulin dose
    if $G_{pred}^{t+T_{safe}} < G_{susp}^{oral}$ **then**
      oral carbohydrate intake
    end if
  else
    get action I from PID or/and soft actor-critic
    if $G_{pred}^{t+T_{safe}} > G_{int}$ **then**
      action $I \leftarrow I + I_{safety}$
    end if
  end if
  update G_{susp}, G_{int}, I_{safety} at each day
end for

are updated based on the portions of hypoglycemia and hyperglycemia in a day. To cope with very low glucose levels, G_{susp}^{oral} can also be set to elevate glucose levels.

Let ρ_{susp}, ρ_{int} be portions of hypoglycemia and hyperglycemia in a day, which have baselines ρ_{susp}^{thr}, ρ_{int}^{thr} for adjusting G_{susp} and G_{int}. Adjusting rates λ_{susp}, λ_{int} can be constants or can be proportional to ρ_{susp} and ρ_{int}. For additional insulin dose $I_{safety} \in [I_{safety}^{min}, I_{safety}^{max}]$ for intervention to prevent hyperglycemia, ΔI_{safety} is the unit insulin for adjusting I_{safety}. The trade-off between hypoglycemia and hyperglycemia is reflected in a series of adaptations in a safe actor.

$$G_{susp} = \begin{cases} (1-\lambda_{susp})G_{susp}, & \text{if } \rho_{susp} > \rho_{susp}^{thr}. \\ (1+\lambda_{susp})G_{susp}, & \text{if } \rho_{int} > \rho_{int}^{thr}. \\ G_{susp} & \text{otherwise.} \end{cases} \tag{10}$$

$$G_{int} = \begin{cases} (1-\lambda_{int})G_{int}, & \text{if } \rho_{int} > \rho_{int}^{thr}. \\ (1+\lambda_{int})G_{int}, & \text{if } \rho_{susp} > \rho_{susp}^{thr}. \\ G_{int} & \text{otherwise.} \end{cases} \tag{11}$$

$$I_{safety} = \begin{cases} I_{safety} - \Delta I_{safety}, & \text{if } \rho_{supp} > \rho_{supp}^{thr}. \\ I_{safety} + \Delta I_{safety}, & \text{if } \rho_{int} > \rho_{int}^{thr}. \\ I_{safety} & \text{otherwise.} \end{cases} \tag{12}$$

The adaptive actor for safety dominates the action from SAC or PID to maintain glucose levels in proper ranges and to prevent the patient's status from hypoglycemia and hyperglycemia based on the predicted glucose level at the time $t + T_{safe}$ after the current time t.

C. FOR INTERPRETABILITY

Attention scores in the DAN originate from the structure of prediction models, and α^t and β^t can be interpreted as the extent to which the model focuses on the specific state in forecasting. This is because of the inherent structure of the model.

For the policy network from the SAC based on the neural network, ad-hoc techniques for post-training can be applied. DeepLIFT is a gradient-based attribution method for interpretability, and attribution scores are compared with basic states, first and second derivatives of basic states, and integrated values of basic states. Most mathematical models for glucose dynamics are based on a series of first or secondary ordinary equations. Changes, levels, and cumulative values of variables are easily understood from the perspectives of simplicity and approximation

To calculate the attribution scores, let $\Delta y_{target} = y_{target} - y_{target}^{ref}$ and x_{neuron}^i be the difference from the reference of the target value, and let neurons (i = 1…n) be in layers to compute y_{target}. Then, the attribution scores of $C_{\Delta x_{neuron}^i \Delta y}$ satisfy the summation-to-delta property. If z_{neuron}^i exists in intermediate layers, then the chain rule for multipliers $m_{\Delta x_{neuron}^i \Delta y_{target}}$ also holds.

$$\begin{aligned} \Delta y_{target} &= \sum_{i=1}^{n} C_{\Delta x_{neuron}^i \Delta y_{target}} \\ m_{\Delta x_{neuron}^i \Delta y_{target}} &= \frac{C_{\Delta x_{neuron}^i \Delta y_{target}}}{\Delta x_{neuron}^i} \\ m_{\Delta x_{neuron}^i \Delta y_{target}} &= \sum_{j=1} m_{\Delta x_{neuron}^i \Delta z_{neuron}^i} m_{\Delta z_{neuron}^i \Delta y_{target}} \end{aligned} \tag{13}$$

By separating the positive and negative components of Δy_{target} and Δx_{neuron}^i, the attribution scores and target output are also decomposed as $\Delta y_{target} = \Delta y_{target}^+ + \Delta y_{target}^-$

$$C_{\Delta x_{neuron}^i \Delta y_{target}} = C_{\Delta x_{neuron}^{i+} \Delta y_{target}} + C_{\Delta x_{neuron}^{i-} \Delta y_{target}}$$

where $x^{i-} < 0$, $x^{i+} > 0$, $y^- < 0$ and $y^+ > 0$.

In the linear calculation of the neural network, $y = b + \sum_i w_i x_i$, the positive and negative parts are decomposed as follows:

$$\begin{aligned} C_{\Delta x_{neuron}^{i+} \Delta y_{target}^+} &= 1(w_i \Delta x_{neuron}^i > 0) w_i \Delta x_{neuron}^{i+} \\ C_{\Delta x_{neuron}^{i-} \Delta y_{target}^+} &= 1(w_i \Delta x_{neuron}^i > 0) w_i \Delta x_{neuron}^{i-} \\ C_{\Delta x_{neuron}^{i+} \Delta y_{target}^-} &= 1(w_i \Delta x_{neuron}^i < 0) w_i \Delta x_{neuron}^{i+} \\ C_{\Delta x_{neuron}^{i-} \Delta y_{target}^-} &= 1(w_i \Delta x_{neuron}^i < 0) w_i \Delta x_{neuron}^{i-} \end{aligned} \tag{14}$$

$$\begin{aligned} \Delta y_{target}^+ &= \sum_i C_{\Delta x_{neuron}^{i+} \Delta y_{target}^+} + C_{\Delta x_{neuron}^{i-} \Delta y_{target}^+} \\ \Delta y_{target}^- &= \sum_i C_{\Delta x_{neuron}^{i+} \Delta y_{target}^-} + C_{\Delta x_{neuron}^{i-} \Delta y_{target}^-} \end{aligned} \tag{15}$$

Multipliers are considered based on the chain rule and attribution propagation, derived from the above equations.

$$\begin{aligned} m_{\Delta x_{neuron}^{i+} \Delta y_{target}^+} &= m_{\Delta x_{neuron}^{i+} \Delta y_{target}^-} \\ &= 1(w_i \Delta x_{neuron}^i > 0) w_i \\ m_{\Delta x_{neuron}^{i+} \Delta y_{target}^-} &= m_{\Delta x_{neuron}^{i-} \Delta y_{target}^-} \end{aligned}$$

$$= 1(w_i \Delta x^i_{neuron} < 0) w_i \tag{16}$$

For non-linear transformations in perceptions of the neural network, multipliers become the first derivative of y_{target} with regard to x_{neuron} in small ranges.

$$\begin{aligned} \Delta y^+_{target} &= \frac{\Delta y_{target}}{\Delta x_{neuron}} \Delta x^+_{neuron} \\ \Delta y^-_{target} &= \frac{\Delta y_{target}}{\Delta x_{neuron}} \Delta x^-_{neuron} \\ m_{\Delta x^+_{neuron} \Delta y^+_{target}} &= m_{\Delta x^-_{neuron} \Delta y^-_{target}} \end{aligned} \tag{17}$$

Attribution scores are calculated via the forward propagation of multipliers, and end-to-end scores C can be compared with the explainable forms of variables in the basic states. There are global and local attribution scores in a unified view of gradient-based interpretable methods [80], and local attribution considers only the gradient and not the value itself, whereas global attribution considers both. The DeepLIFT adopted in this study belongs to the category of global attribution methods.

To identify information in the attention scores of the prediction model DAN and the attribution scores of SAC, correlations with basic states and their derivative and integral forms can be observed. Basic states themselves, their derivatives and their integrals are relatively simple forms of variables that humans can easily recognize because many controllers and mathematical models contain these forms in structures.

To evaluate attention and attribution scores in groups with correlation with extended states from DAN, the absolute values of correlations should be analyzed. Hidden states in LSTM have no limitation of having positive or negative signs, and they depend on the initialization of the weight in the neural network and local optima during training. In forward propagation in neural networks, (element-wise) production of negative weights and negative values of hidden states lead to the identical output of (element-wise) production of positive weights and positive values of hidden states because different signs are compensated for in production. Therefore, in the case of comparison of correlations for all individuals, extended states can have positive or negative signs, and absolute values of correlations can be the proper measure of how much the variable in extended states affects the prediction and control outputs from the perspective of interpretability with simple forms of basic states.

IV. EXPERIMENTS AND RESULTS

In this section, we describe the experimental setting and the results obtained. First, we describe scenarios for virtual patients in simulations and guiding methods for the fast training of the soft actor-critic with safety. Second, two prediction models, RFR and DAN, were compared. Third, the performance of SAC combined with a safe actor to regulate blood glucose levels was evaluated for each individual and each group. Fourth, the extended states, simple forms of basic states, and attribution scores were investigated from the perspective of interpretability.

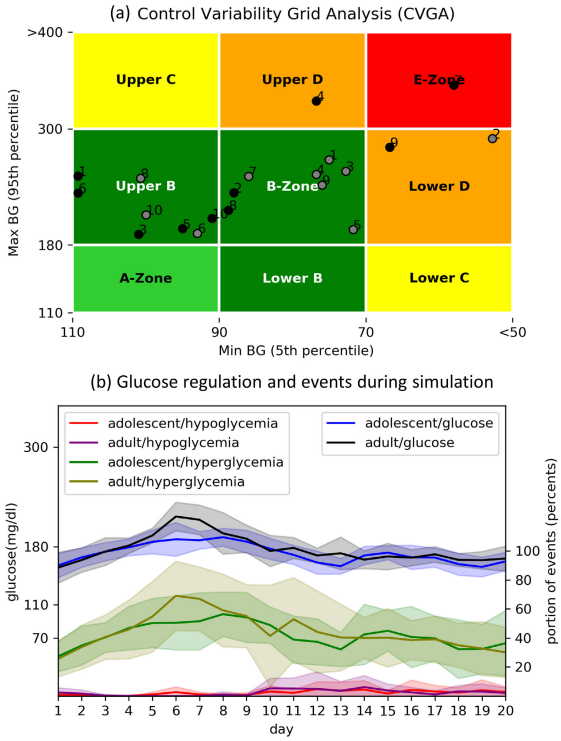

FIGURE 3. (a) Control variability grid analysis for the last two days (RL) is shown. (Gray: adolescents, Black: adults) Most patients are positioned in the safe zone B, whereas some patients are in the relatively dangerous zones D and E. (b) Averaged glucose and insulin curves in 20 days of each group are shown. Hypoglycemia and hyperglycemia were defined as events in which blood glucose levels were lower than 70 mg/dL and higher than 180 mg/dL in this study, respectively. Proportions of hypoglycemia and hyperglycemia over time are shown. (Lines: mean values, shaded regions: standard deviation values).

A. PREPARATION

Simulation platforms for type 1 diabetes have been proposed in several studies. In this study, the FDA-approved type 1 diabetes simulator, T1DMS (version 3.2), known as the UVA/Padova T1D simulator, which is the original version based on MATLAB (version R2019b), interacted with the PYTHON (version 3.6.10) platform containing prediction and control models through a TCP/IP connection. It was assumed that every 5 min in the simulation, sensor information was received and controller action information was sent to the simulator. Thus, the unit time of the simulations was set as 5 min. For the deep learning framework, the Pytorch library (version 1.4.0) was adopted. R (version 4.0.5) was used for the statistical analysis.

Information on glucose levels, insulin levels infused from the pump, and meal intake were measured. These were provided by the simulator and transmitted to the PYTHON environment. Insulin dosing was calculated based on PID and/or SAC with the adaptive safe actor, and this information on the actions of insulin from the calculation from models was delivered to the virtual pump. However, the information of insulin from the pumps in the simulator was also

TABLE 1. Statistics of glucose and parameters for adaptive safe actor for patients in the simulation.

Group	glucose statistics							parameters for safety			MAE of G_{pred}^{30min}	
Patient	Zone	G med.(PID)	G med.(RL)	G 5p.(PID)	G 5p.(RL)	G 95p.(PID)	G 95p.(RL)	G susp.	G int.	I safety	ΔG_{pred}^{DAN}	ΔG_{pred}^{RFR}
adolescent 1	B	150	170	91	75	281	268	102	215	1.5	44.0 ± 25.5	15.6 ± 17.0
adolescent 2	D	139	117	64	53	307	290	101	216	1.5	42.0 ± 28.7	28.2 ± 27.9
adolescent 3	B	147	165	96	73	212	256	82	204	1.25	23.7 ± 20.1	11.8 ± 10.2
adolescent 4	B	178	198	123	77	255	253	153	240	0.125	13.9 ± 14.7	10.9 ± 12.3
adolescent 5	B	142	120	102	72	205	196	131	254	0.5	36.3 ± 23.3	10.8 ± 12.1
adolescent 6	B	141	127	100	93	196	192	72	185	1.5	14.4 ± 10.0	8.8 ± 8.8
adolescent 7	B	177	139	128	86	272	251	107	184	1.0	41.8 ± 26.7	15.0 ± 16.9
adolescent 8	B	169	160	116	101	227	249	95	221	1.375	24.4 ± 15.5	10.2 ± 11.1
adolescent 9	B	155	198	106	144	206	242	131	195	1.0	21.8 ± 10.2	8.9 ± 9.5
adolescent 10	B	157	148	104	100	231	211	110	238	0.75	18.6 ± 16.2	12.4 ± 13.5
average		155.5 ± 14.6	154.2 ± 29.4	103.0 ± 18.0	87.4 ± 24.7	239.2 ± 37.7	240.8 ± 31.6	108.4 ± 24.3	215.2 ± 23.8	1.05 ± 0.47	28.1 ± 19.1	13.3 ± 13.9
adult 1	B	142	165	67	111	249	251	106	230	1.375	32.3 ± 22.6	13.6 ± 14.1
adult 2	B	149	182	60	132	249	234	124	181	1.0	14.6 ± 16.4	11.4 ± 11.6
adult 3	B	161	137	103	101	258	191	124	204	0.875	9.1 ± 7.6	7.6 ± 7.6
adult 4	D	142	168	91	77	193	329	109	238	0.875	14.3 ± 11.3	9.0 ± 8.3
adult 5	B	149	141	98	95	215	197	129	250	0.625	19.7 ± 13.0	9.5 ± 8.8
adult 6	B	173	153	117	111	290	234	125	248	0.5	16.8 ± 11.3	10.7 ± 12.1
adult 7	E	199	184	82	58	351	345	72	180	1.5	45.9 ± 27.6	23.6 ± 18.4
adult 8	B	143	163	97	89	210	216	79	203	1.5	15.1 ± 10.2	8.8 ± 8.7
adult 9	D	139	156	69	67	226	281	115	236	1.25	29.0 ± 21.3	17.8 ± 13.7
adult 10	B	140	156	100	129	212	208	116	237	1.0	20.2 ± 12.2	8.1 ± 9.5
average		153.7 ± 19.2	160.5 ± 15.4	88.4 ± 18.3	97 ± 24.8	245.3 ± 46.8	248.6 ± 53.7	109.9 ± 19.6	220.7 ± 26.5	1.05 ± 0.35	21.8 ± 15.4	12.4 ± 11.7

The zone is determined by the 5^{th} percentiles and 95^{th} percentiles of glucose levels in the last two days (RL) in the control variability grid analysis. The subscript PID indicates data from the first two days with PID control, and the subscript RL indicates data from the last two days with soft actor-critic control. Subscripts med., 5p., and 95p. denote the median, 5^{th} percentiles, and 95^{th} percentiles, respectively. The subscript G indicates glucose level. G_{susp} and G_{int} are thresholds for glucose levels to determine the suspension and additional insulin dosing by the safe actor. I_{safe} is the insulin dose when the safe actor is activated to alleviate hyperglycemia. The MAE of G_{pred}^{30min} denotes the mean absolute error of the predicted glucose levels in the next 30 min. The parameters for safety and MAE of G_{pred}^{30min} were measured in the last two days (RL) in the simulations.

TABLE 2. Portions of activation events from safe actor in control.

patient	susp.(PID)	susp.(RL)	add.(PID)	add.(RL)
adolescent 1	44.8	29.9	24	50
adolescent 2	54.5	36.1	37.5	50.3
adolescent 3	41.7	20.5	23.3	34.7
adolescent 4	18.4	22.9	9.4	33.7
adolescent 5	39.6	13.9	54.2	12.2
adolescent 6	45.8	6.3	50.7	2.1
adolescent 7	22.2	22.2	8.7	59.4
adolescent 8	24	20.1	1	39.9
adolescent 9	34.7	15.3	0	41.7
adolescent 10	34.7	20.8	40.6	14.6
adult 1	57.6	30.9	14.2	43.4
adult 2	53.5	33	0	20.8
adult 3	38.9	33.7	0.7	6.3
adult 4	44.8	18.8	30.6	54.9
adult 5	42.7	19.8	45.1	1
adult 6	32.6	30.6	0	28.5
adult 7	36.1	57.6	38.5	58.7
adult 8	41.7	11.8	10.1	22.6
adult 9	47.9	21.9	36.1	42.7
adult 10	46.9	7.6	0	19.8
average	40.2 ± 10.4	23.7 ± 11.6	21.2 ± 19.0	31.9 ± 18.7
p-value	0.0014^*		0.0045^*	

The abbreviations susp. and add. indicate the insulin suspension and additional insulin dosing by the safe actor, respectively. PID denotes that the PID policy of actions for control is used, and RL denotes that the SAC policy of actions for control is used. The units are percentages, and values except p-values are portions of activation events of the safe actor. Statistical comparisons of the portions of activation of the safe actor were conducted. The Wilcoxon signed-rank test was performed to compare suspension events, and the t-test was used for the events of additional insulin dosing.

gathered because actual delivery amounts from the virtual pump should be considered in state estimation. In short, glucose levels and amount of meal intake at a certain time were the only variables from the patients, and insulin dosing was measured from the pump, not from the patient.

As provided in the simulator, the glucose dynamics of 10 adults and 10 adolescents were tested over 20 days. In the first two days, only PID control with the adaptive actor for safety determined the entire insulin dose. The portion of insulin dose from the SAC for the virtual pump increased by 20% in all insulin actions every two days until only RL with a safe actor considered the dose of insulin. In short, the insulin dose was determined only by the RL policy after day 10 of the simulation.

Every virtual patient took four or five meals in the predefined scenario, which included 40 g, 50 g, 20 g, 50 g, and 20 g of carbohydrates at 7 am, 12 pm, 16 pm, 20 pm, and 23 pm, respectively. The randomness of uniform distribution for mealtime (-30 min to $+30$ min) and intake amounts (-12.5% to $+12.5\%$) was additionally imposed on meal intakes.

The analysis was conducted on groups of adolescents and adults to investigate the overall outcomes of prediction and control models and individuals to study how the prediction and control models work. Group-based statistics are meaningful, as T1DMS has virtual patients representing the population of real diabetic patients according to FDA approval.

B. FORECASTING

The experiments in this study were designed to mimic the environments of the clinical setting, which do not allow severe dangerous conditions in patients. The parameters

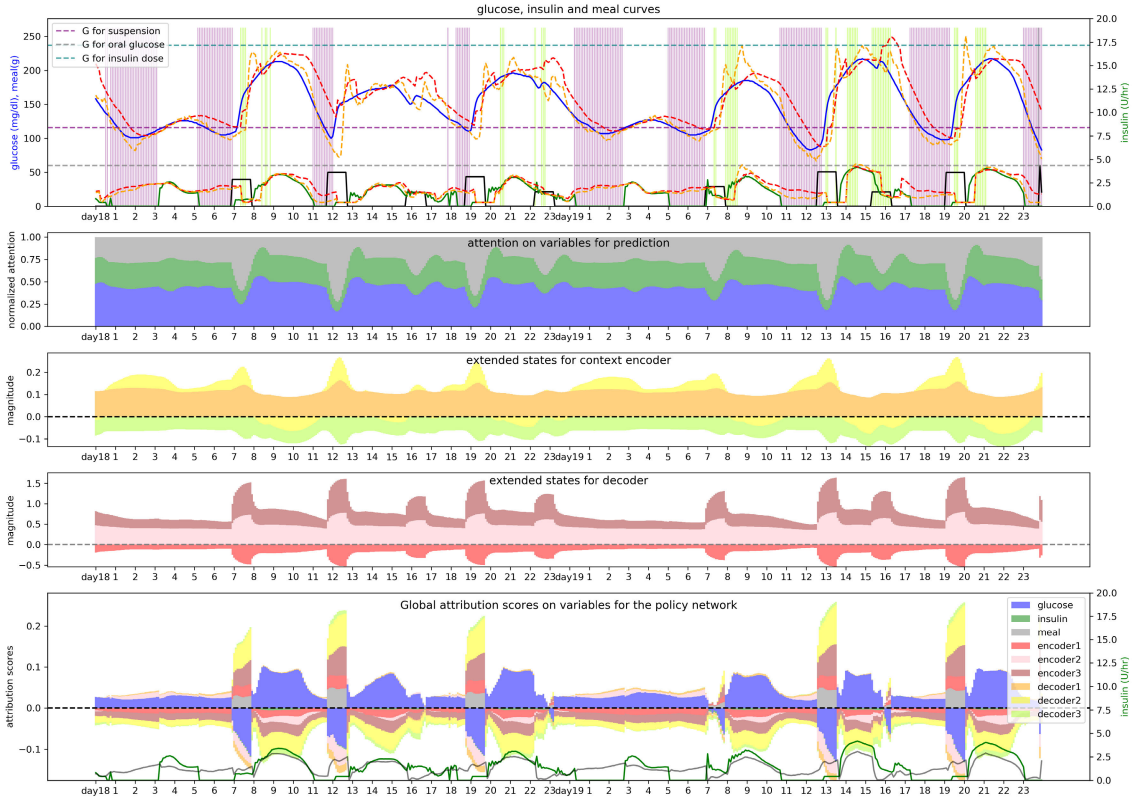

FIGURE 4. Trajectories, states, attention scores, and attribution scores based on prediction models and the controller for adolescent 10 are shown. The simulation started on day 0, and curves for the last two days (i.e. the start of day 18 to the end of day 19) are depicted in the figures. The x-axis is the time axis. In the graph of the first row, the blue, green, and black lines are the trajectories of glucose, insulin administered by the pump, and occurrences of meal intakes, respectively. Dashed lines indicate the curves predicted by the DAN (red) and RFR (orange). Horizontal dashed lines are adaptive threshold values for predictive glucose levels after 45 min in the safe actor to induce actions including additional insulin dosing (sky blue), suspension (purple), and oral carbohydrate rescue (gray). Activation of the safe actor is indicated in shaded regions: suspension (purple) and additional insulin dosing (lime). Variable attention scores with a weighted summation of temporal attention scores on basic states are shown in the graph in the second row. The summation of the attention scores was set to 1. A higher attention score of a variable means that DAN focuses more on that variable to predict future responses. In the third and fourth row graphs, the values of the extended states from the context (encoder) vector C^t (third row) and hidden states d^t (fourth row) are shown. The area of a variable is proportional to the values of that variable, and whether the positions of areas are in positive/negative domains indicates the positive/negative signs of values at a specific time. Attribution scores for SAC based on DeepLIFT are shown at the graph in the last row. The area and the position of a domain for each variable indicate the value and sign of that variable, as in the graphs in the 3rd and 4th rows. In short, a positive area of a variable means that the variable positively affects (i.e., promotes) the action from SAC, whereas a negative area of a variable means that the variable negatively affects (i.e., reduces) the action from SAC. The measured insulin from the pump (green curve) and the output from SAC (gray curve) are shown.

of the models must be updated under online settings to reflect time-varying characteristics. The basic state x^t is the three-dimensional variable of glucose, insulin, and the amount of meal intake. The context (encoder) vector C^t and hidden states d^t were set as three-dimensional variables. Thus, the extended states (x^t, C^t, d^t) belong to the nine-dimensional
state space.

The time window for training T_w, prediction window T_{pred}, and the time window for determination of unsafe states in the future T_{safe} were set as 60, 30, and 45 min, respectively. The time window for the imagination T_Q was set as 150 min. To consider recent changes in glucose dynamics in simulations and computational burdens in the case of training data of whole sequences, data in the last 250 min from the current time t were exploited for the training of DAN and RFR in each batch.

The metric of the mean absolute error for glucose prediction based on RFR and DAN was used. As shown in Table 1, the values of MAE for forecasting glucose levels (mg/dL) in 30 min were $28.1 \pm 19.1(mg/dL)$ (DAN) and $13.3 \pm 13.9(mg/dL)$ (RFR) in the adolescent group, and $21.8 \pm 15.4(mg/dL)$ (DAN) and $12.4 \pm 11.7(mg/dL)$ (RFR) in the adult group, respectively. The MAE of DAN was higher than that of RFR because DAN tended to overestimate the glucose compared with RFR.

As shown in Fig. 4, which depicts the prediction and control of the virtual patient adolescent 10, not only glucose levels but also future insulin actions from the policy network were predicted. The two prediction models demonstrated

different characteristics. Overall, RFR showed fluctuating values, and DAN's prediction overestimated values. Thus, DAN and RFR were concurrently used to complement each other.

When meal intakes occurred, attention scores of meal intakes also increased, as shown in the second-row graph in Fig. 4. This means that DAN focused on the meal intake events to predict future glucose levels. Glucose levels were more steadily focused on insulin dosing.

The extended states produced by the DAN reflected basic states. T-stochastic neighbor embedding (t-SNE) [81] was implemented for visual description in Fig. 5. (a), and showed that the clustered trajectories evolved as basic states. The closer the two basic states were, the closer the two extended states were expected to be. However, these were not completely matched because additional information on the context of the basic states also existed. A detailed explanation is provided in section IV. (D).

C. ACTIONS FOR CONTROL

PID control acted dominantly in the simulation and was gradually replaced by a SAC with every 20% increase every two days. Policy combinations of PID and SAC existed from the third to the tenth day. The SAC only determined the continuous action of insulin dosing after day 10.

$$r = exp(-\epsilon|G - G_{target}|), \quad \epsilon > 0 \tag{18}$$

The reward for reinforcement learning for this study is shown above, and G_{target} was set to 127 (mg/dL) according to the median value of the border between hypoglycemia and hyperglycemia.

The control action was stochastically generated from the output in the case of SAC, which generated values of the mean and logarithm of the standard deviation for the insulin action of the policy network. The insulin dose cannot be negative; thus, the output of the SAC was constrained to be greater than zero.

No significant difference was observed in the median glucose levels between the control algorithms for each group. In the adolescent group, medians of glucose levels were $155.5 \pm 14.6(mg/dL)$ in PID and $154.2 \pm 29.4(mg/dL)$ in RL. In the adult group, medians of glucose levels were $153.7 \pm 19.2(mg/dL)$ in PID and $160.5 \pm 15.4(mg/dL)$ in RL.

Each patient belonged to either one of zones A to E, where A and B were relatively safe, and zone E was dangerous, as shown in Fig. 2. Most virtual patients belonged to zone B, whereas three were in zone D and one in zone E. Table 1 shows a summary of the prediction and control outcomes in individuals of all groups. Median glucose levels were higher than the objective glucose level of 127 mg/dL in most cases. To train the parameters of the SAC network, a phase transition scheme of combination in actions of insulin dose was adopted for 10 days. Hyperglycemia trends were observed during this period, as shown in Fig. 3. (b).

A comparison of portions of safe actor activation was conducted. For the portion of suspension, $40.2 \pm 10.4\%$ in PID duration was significantly higher than $23.7 \pm 11.6\%$ in RL duration with statistical significance. For the portion of additional insulin dosing, $21.2 \pm 19.0\%$ in PID duration was significantly lower than $31.9 \pm 18.7\%$ in RL duration with statistical significance, as indicated in Table 2. In short, PID required more suspensions and SAC required additional insulin dosing events.

D. INTERPRETABILITY FOR PREDICTION AND CONTROL

Extended states constructed from the DAN had characteristics of physiological responses in each patient. As shown in Fig. 5. (b), extended states were analyzed based on correlations between values, first derivatives, second derivatives, and integrals of basic states such as glucose levels, insulin, and meal intake.

An example of adolescent 10 is shown in Fig. 4 and Fig. 5 (b). First, we can investigate the extended states internally from the DAN prediction model. A positive or negative correlation between simple forms of basic states occurred in some cases, where simple forms are variables of basic states and their first and second derivatives, and integrals. In this case, correlations between two variables of the context vector (enc1 and enc2) and the integral of insulin (intI) can be emphasized because they can be interpreted as surrogates of IOB. The first derivative of glucose levels (D1G), integral of glucose levels (intG), and integral of meal intake (intM) were also correlated with the variables in the extended states. In the third and fourth row graphs of Fig. 4, the context vector and hidden states evolved over time and corresponded to the basic states.

Second, we can investigate the control model, SAC, to understand its behavior from the perspective of interpretability. For example, as shown in the fifth-row graph of Fig. 4, attribution scores around 12 pm can be explainable by the SAC policy. At the start of a meal intake, the information on food ingestion (positive gray area) promotes positive action, whereas information on proper glucose levels (negative blue area) was inclined to reduce the action. The Decoder2 (dec2) variable had a positive correlation with the first derivative of the glucose level and the integrals of meal intake. This means that dec2 played a role in considering the change in glucose levels and cumulative meal intake over time. During meal intake, dec2 (positive yellow area) had positive attribute scores, which means that the first derivative of glucose levels contributed positively to the action of SAC.

After meal time, information on high levels of glucose (positive blue area) supports positive action; variables of Encoder1 (enc1) and Encoder2 (enc2), which were correlated to the integral of the insulin dosing (intI), negatively affected the action of SAC. intI can be a proxy for IOB; thus, SAC learned responses to IOB from some extended states without prior knowledge of physiology.

To investigate the extended states in SAC, correlations between attribution scores and simple forms of basic states were analyzed, as shown in Fig. 5. (c). Basic states, such as glucose levels, insulin dosing, and meal intake, were strongly

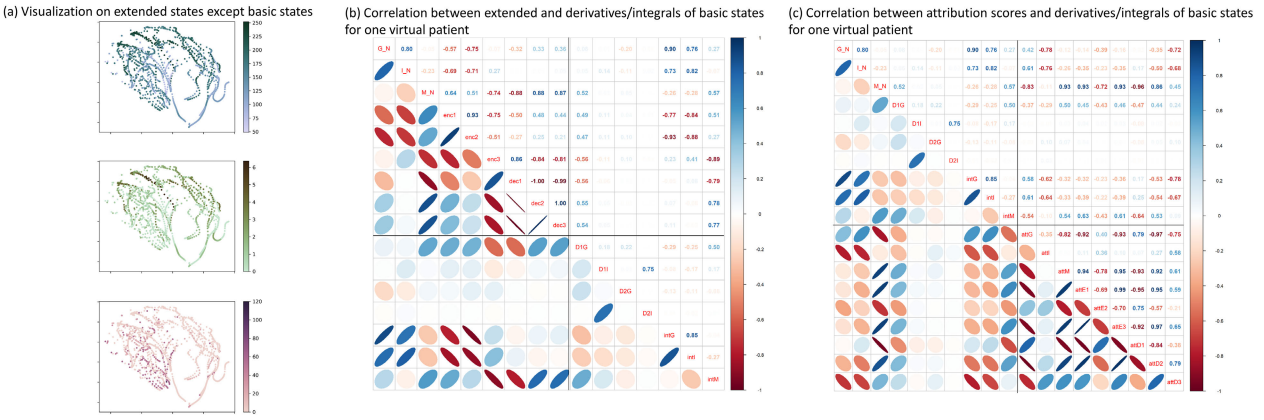

FIGURE 5. (a) T-SNE visualization of extended states without basic states. The concatenated vectors of C^t and d^t are represented in the reduced dimensional space. (b) Correlations between derivatives/integrals of basic states and extended states of adolescent 10 are shown. (c) Correlation between extended states and attribution scores for the SAC network of adolescent 10 are shown. In both figures (b) and (c), enc or E represents the context vector, and dec or D denotes the hidden vector from the decoder in the DAN. For example, dec2 denotes the second hidden state variable. D1 and D2 represent the first and second derivatives, respectively, and int denotes the integral. G, I, and M are glucose levels, insulin, with min-max normalized as N. Attribution scores are denoted as att. In the correlation plots, as the shape becomes a line from a circle, the correlation becomes stronger. For example, enc2 had strong negative correlations with intG and intI, whereas enc3 had a strong negative correlation with intM.

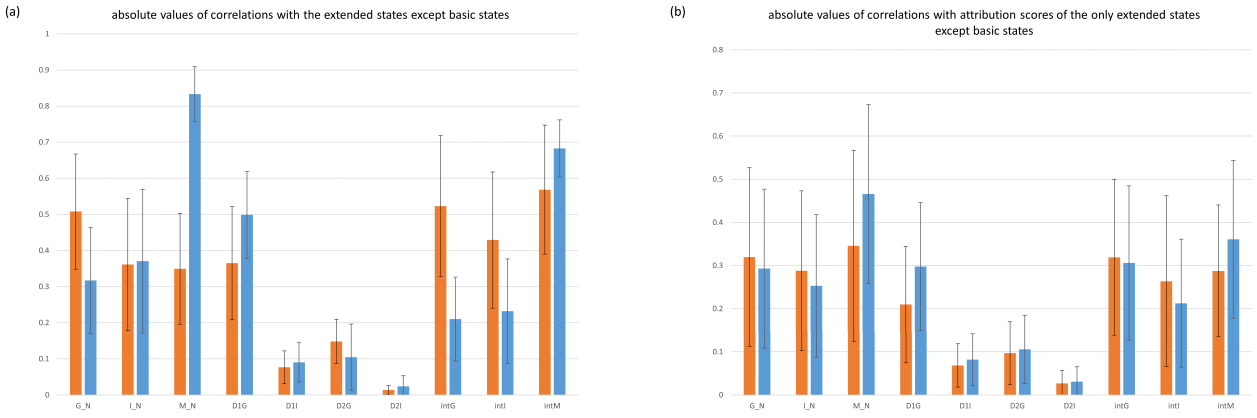

FIGURE 6. (a) Absolute values of correlations between the context vector C^t and hidden states d^t of the decoder, and basic states and their derivatives/integrals (i.e. simple forms) are shown. In short, the relations between the extended states and simple forms of basic states are investigated. The simple forms derived from the basic states provide hints of explainable information on extended states, which are constructed from DAN. (b) Absolute values of correlations of attribution scores from the context vector C^t and hidden states d^t of the decoder, with basic states and their derivatives/integrals shown. In short, the relations between the attribution scores of SAC and simple forms of basic states are investigated. This shows how the policy network considers the explainable components of basic states, including derivatives and integrals of biological variables. In both figures, the two bar plots are combined. The left bar (orange) indicates the absolute value of the correlation with the context encoder variables C^t, and the right bar (blue) indicates the absolute value of correlation with the hidden states d^t from the decoder. The mean (thick bars) and standard deviation (thin lines) are shown.

correlated with attribution scores. The integrals of the basic states and the first derivatives of glucose levels showed moderate correlations with attribution scores.

Notably, the interpretation of models and behaviors of controllers with extended states and attribution scores is subject-specific. The variables of context vectors and hidden states from the decoder can have different roles owing to initialization and training in neural networks for each individual.

To obtain generalized interpretations of all subjects from subject-specific results, we investigated the absolute values of correlations between extended states, simple forms of basic states, and attribution scores. As mentioned in section III. (C), both positive and negative correlations can exist and can be used to describe the phenomena without distinction; the absolute values of correlations were used for comparison.

The context vector C^t had relatively high absolute values of correlations with integrals of glucose (intG), integrals of insulin (intI), integrals of meal intakes (intM), and glucose levels (G), whereas the hidden states d^t of the decoder were more focused on meal intake (M), integral of meal intake (intM), and the first derivative of glucose level (D1G), as shown in Fig. 6. (a). Based on these results, we believe that DAN can generate extended states that reflect highly understandable features. For attribution scores related to SAC

control, basic states, integrals of basic states, and the first derivative of glucose level are relatively more important to determine actions, as shown in Fig. 6. (b). Meal intake was the most important factor affecting the SAC.

V. DISCUSSION

Because of the minimal assumptions on the mathematical structures in physiological responses in this research, this framework can be applied to various clinical conditions with time-series data, including some cases that have no explicit mathematical models for physiological phenomena for patients and diseases. The prediction and control models conduct trial-and-error testing for safety with interpretable components. This approach can reduce the bias from the explicit structure of structured models; however, it requires more data sampling and accurate training of the models. Reduction of participation time and ensuring safety are important in clinical trials; thus, essential guides for training in the early phase and adaptive additional insulin dosing can resolve these issues. In this study, PID control for policy training and a safe actor with suspension and additional insulin dosing were introduced for these objectives.

Recently, methods and applications of interpretability have attracted attention from several studies. However, most studies have focused on classification problems, not neural network-based controllers. In contrast to prediction tasks, the controller affects the external environment, especially in the case of the human body, during the regulation of blood glucose levels. The clues of the machine's thoughts and actions should be provided for manipulation and intervention for controllers and patients in advance, and we hope the approaches in this framework can provide some insights into the behaviors of neural network-based controllers in clinical applications.

We investigated the internal operations of neural network-based models, DAN and SAC, from the perspective of interpretability. The safe actor had explicit rules to modulate the actions of insulin dosing, and it was quite explainable. However, the models' policies were partially separate for safe and unsafe conditions. The policy from SAC could be explained with regard to extended states and simple forms of basic states by means of DeepLIFT, whereas the explanations of the entire policy were not smoothly integrated. To resolve this limitation, we consider the adoption of a neural network-based safe actor connected to the SAC instead of an adaptive rule-based actor in future research. Training and investigating models in an end-to-end fashion can provide integrated perspectives for interpretation.

We can use this framework as an alarm system to deal with hyperglycemia and hypoglycemia. Conditions to activate the safe actor can be regarded as alarms for clinicians and patients to prepare for unexpected events. The activation of early alarms can be determined by the length of time window, extended states as the context of glucose dynamics, future policies of controllers which are elements of prediction and control models. Clinical decisions including interventions and treatment plans by human experts can be supported and improved by interpretable operations of generating alarms. We expect that an alarm system can be significantly helpful in clinical trials and diabetes management, alarm systems with interpretability should be further studied for real patients.

A. PERFORMANCE OF PREDICTION AND CONTROL MODELS

As the horizontal time window increases, the forecasting performance generally decreases. If meal intake suddenly occurs, the prediction becomes inaccurate because meal intake has a strong influence on glucose dynamics. Time can be a potential feature related to meal intake, but the subject should follow a regular meal time schedule to train the model regarding the temporal patterns of food ingestion. If the meal schedule becomes irregular, then the forecasting performance of the prediction model that has a time variable as a feature deteriorates. Taking this into account, we did not consider time itself as a variable for the prediction model, even though the randomness of amounts and schedules for meal intake was considered.

Although the DAN was used for establishing extended states, the performance of prediction of future glucose levels based on MAE was not better than that of RFR. Models were trained in an online fashion with a limited length of sequences, not whole sequences at each training epoch, because of computational burden and to ensure reflection of recent variability in dynamics. In addition, DAN was designed to focus on specific temporal sequences by adopting weights dynamically, whereas RFR had many decision trees, and unexpected disturbances could impact only some trees because of ensembles in RFR. Thus, we suppose that this locality in the prediction processes and the unexpected meal intake in the prediction horizon may have a greater impact on the attention-based methods than the boosting methods.

We adopted online learning processes to mimic the clinical environment, and the time complexity of the models should be considered. Training neural networks requires many calculations in automatic differentiation steps via backpropagation processes. The training step of a DAN is a bottleneck, through which extended states must be constructed to compensate for minimal knowledge of glucose dynamics. Thus, a trade-off exists between the prediction performance and computational burdens. A balanced solution would be practical. RFR and DAN are feasible solutions for prediction because RFR has a light computational burden, and DAN has an encoder and a decoder to construct extended states.

The SAC control with a safe actor showed comparable performance to that of the PID control. PID control is known to be comparable [64] with model-based control; thus, we think that it is a feasible solution with decent performance, even though RL in this study did not exhibit better performance. In addition, many episodes were not required for RL training in this study because of the guided PID control at the beginning of the simulations and the existence of the adaptive soft actor. However, it would be better if the performance

of RL could be improved. Though SAC can be a feasible backbone to regulate blood glucose levels from the perspective of RL, an additional investigation to improve the control performance should be explored in future research.

B. BALANCE BETWEEN REINFORCEMENT LEARNING AND INTERVENTION BASED ON PRIOR KNOWLEDGE

In this study, forecasting models and controllers use as little prior knowledge of physiology as possible because we assume that mathematical models with rigid structures may have biases that stem from the assumptions of structures or information lost by simplifying the dynamics, despite the sample efficiency. However, flexible models usually require more data. Guiding policies such as PID control in the early phase and a safe actor were introduced to resolve this issue. In addition, this framework can be adopted not only in the regulation of glucose dynamics, but also in various clinical applications including administration of medication to the heart in terms of cardiac physiology represented by electrocardiogram signals, and long-term interventions for chronic disease at various time scales.

We attempted to avoid the common knowledge of mathematical models that affect the performance of forecasting models and controllers in evaluations. For example, if a specific mathematical model is shared by the simulator, the prediction models and/or controllers, then the performance of models tested in simulation may be easily higher than in actual environments because even efficient closed-form solutions without observation may exist. However, the actual physiology may differ from the approximated equations of the models. In this research, we did not use explicit mathematical models based on physiology in the simulator as prior knowledge. We expect that the performance of the models in practical conditions would not be significantly different from that in the simulation. We plan to apply this framework to a real clinical environment for validation in the future.

However, prior knowledge of physiology, including mathematical equations based on physiology, is essential to apply control algorithms and to train models and controllers in real clinical environments. This knowledge can reduce trials and errors during training and delivering interventions. Biological and physical laws in nature described by mathematical equations can be effective guidelines for modeling and training. Thus, we plan to study the hybridization of mathematical model-based approaches and model-free RL controllers in the future.

C. EXTENDING STATES FROM LOW TO HIGH DIMENSIONS

In this study, low-dimensional data had to be extended to construct features containing physiological information in temporal contexts. Extended states contain information on basic forms derived from basic states, and derivatives and integrals are fundamental elements for some mathematical equations. For example, linear state-space equations can be represented as a series of ordinary differential equations with first derivatives. However, the remaining information on more complex relationships should be clarified. Correlation analysis provided meaningful insights into the information contained in the extended states and attribution scores. However, there may be more interpretable variables with hidden information. Extracting knowledge from extended states is required and should be further studied.

Approaches related to mutual information and disentanglement [82], [83] can be alternatives for constructing extended states. It would be better for each variable to have a clear meaning and be orthogonal to the others. To be implemented in clinical environments, it would be helpful to easily understand the features exploited in models without any redundancy to provide clear explanations.

VI. CONCLUSION

We proposed a framework for forecasting and controlling blood glucose, which can be safely adopted in clinical environments, and provide an interpretation of the behaviors of models for intervention in advance. An FDA-approved simulator was used to validate the algorithms, and the performance of SAC algorithms for regulation of blood glucose levels was comparable to that of PID control.

The models exploited prior knowledge of internal physiological dynamics as little as possible because of the flexibility of reflecting time-varying dynamics and minimizing the performance gap between simulations and actual environments during testing. To compensate for minimal prior knowledge, PID control guided the training of SAC, and adaptive safe actors modulated the insulin dosing.

The extension of states is an effective approach for capturing physiological relations between variables based on data, and correlations with simple forms of basic states provide internal dynamics information. Attention and attribution scores for prediction and control models clarify the intentions and behaviors of models from the perspective of interpretability.

We hope that this study can provide novel and practical insights into aspects that must be considered when adopting RL for clinical applications.

APPENDIX A
STRUCTURES OF MODELS AND HYPERPARAMETERS

In the prediction models, the RFR had a maximum depth of 2. LSTM had one layer in the DAN. The encoder of the DAN used one linear layer as the attention module, and the decoder used sequential linear, Tanh, and linear layers for the attention module. Weights and biases were initialized in a uniform distribution in [-1e-6,1e-6].

The basic states were transformed using min-max normalization. Minimal values were set as zero in all basic states such as glucose, insulin, and meal intake, and maximal values were set as 500 (mg/dL), 20 (U/h), and 100 (g), respectively. With these normalized basic states, the extended states were estimated using DAN.

For PID control, the coefficients of the proportional, integral, and derivative components were considered as 2.0, 0,

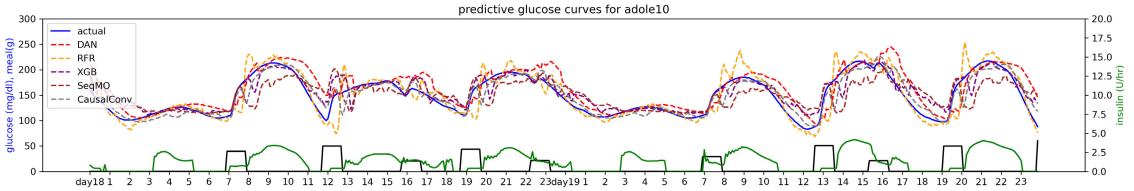

FIGURE 7. A comparison of prediction models to forecast blood glucose levels for the subject of adolescent10 in last two days. Each model was modified to be trained in online learning without prior knowledge of glucose dynamics. The measured values of glucose, insulin, and meal intake are represented in blue, green, and black curves. Dashed lines represent the predictive values of glucose levels after 30 min, which are compared to the measured values.

TABLE 3. Comparison of the predictive performance of prediction models.

Group	MAE of G_{pred}^{30min}					R				
Patient	DAN	RFR	XGB	SeqMO	CausalConv	DAN	RFR	XGB	SeqMO	CausalConv
adolescent 1	44.0 ± 25.5	15.6 ± 17.0	18.3 ± 13.7	35.7 ± 28.5	8.3 ± 9.6	0.82	0.94	0.92	0.64	0.98
adolescent 2	42.0 ± 28.7	28.2 ± 27.9	26.2 ± 24.4	48.3 ± 39.2	12.7 ± 13.1	0.85	0.92	0.88	0.60	0.97
adolescent 3	23.7 ± 20.1	11.8 ± 10.2	15.6 ± 12.5	29.9 ± 22.6	8.0 ± 7.1	0.91	0.97	0.94	0.71	0.99
adolescent 4	13.9 ± 14.7	10.9 ± 12.3	16.7 ± 11.4	28.7 ± 21.8	8.2 ± 6.9	0.95	0.97	0.95	0.75	0.98
adolescent 5	36.3 ± 23.3	10.8 ± 12.1	12.1 ± 11.9	21.9 ± 17.4	6.6 ± 6.4	0.72	0.94	0.91	0.68	0.97
adolescent 6	14.4 ± 10.0	8.8 ± 8.8	10.5 ± 9.4	18.4 ± 15.3	5.8 ± 5.7	0.94	0.95	0.91	0.63	0.97
adolescent 7	41.8 ± 26.7	15.0 ± 16.9	18.2 ± 15.5	35.2 ± 29.4	9.6 ± 9.7	0.93	0.94	0.90	0.53	0.97
adolescent 8	24.4 ± 15.5	10.2 ± 11.1	10.0 ± 8.7	16.4 ± 16.9	6.3 ± 6.1	0.98	0.95	0.95	0.83	0.98
adolescent 9	21.8 ± 10.2	8.9 ± 9.5	8.4 ± 6.8	14.5 ± 12.2	5.1 ± 4.5	0.95	0.94	0.94	0.77	0.98
adolescent 10	18.6 ± 16.2	12.4 ± 13.5	12.9 ± 11.6	22.8 ± 17.4	10.8 ± 9.0	0.85	0.92	0.91	0.64	0.93
average	28.1 ± 19.1	13.3 ± 13.9	14.9 ± 12.6	27.2 ± 22.1	8.1 ± 7.8	0.89 ± 0.08	0.95 ± 0.02	0.92 ± 0.02	0.68 ± 0.09	0.97 ± 0.02
adult 1	32.3 ± 22.6	13.6 ± 14.1	15.4 ± 18.2	27.5 ± 27.3	8.2 ± 8.1	0.93	0.94	0.86	0.55	0.97
adult 2	14.6 ± 16.4	11.4 ± 11.6	11.9 ± 10.5	23.5 ± 18.7	8.5 ± 6.7	0.88	0.92	0.88	0.47	0.96
adult 3	9.1 ± 7.6	7.6 ± 7.6	9.4 ± 7.8	16.2 ± 13.0	4.8 ± 4.3	0.94	0.96	0.92	0.62	0.96
adult 4	14.3 ± 11.3	9.0 ± 8.3	13.0 ± 12.8	23.1 ± 17.8	7.1 ± 5.8	0.98	0.99	0.98	0.93	0.99
adult 5	19.7 ± 13.0	9.5 ± 8.8	12.8 ± 10.1	22.2 ± 15.2	7.0 ± 5.4	0.93	0.95	0.90	0.64	0.97
adult 6	16.8 ± 11.3	10.7 ± 12.1	13.1 ± 10.9	22.7 ± 19.3	7.8 ± 7.9	0.93	0.94	0.90	0.63	0.95
adult 7	45.9 ± 27.6	23.6 ± 18.4	31.2 ± 26.1	68.0 ± 40.7	16.0 ± 11.7	0.83	0.98	0.93	0.62	0.99
adult 8	15.1 ± 10.2	8.8 ± 8.7	11.4 ± 7.7	21.1 ± 14.2	6.2 ± 4.6	0.95	0.97	0.95	0.75	0.98
adult 9	29.0 ± 21.3	17.8 ± 13.7	22.9 ± 20.8	48.5 ± 35.0	12.7 ± 9.8	0.88	0.97	0.90	0.44	0.97
adult 10	20.2 ± 12.2	8.1 ± 9.5	7.2 ± 7.3	12.3 ± 13.3	4.0 ± 4.8	0.94	0.91	0.91	0.68	0.97
average	21.8 ± 15.4	12.4 ± 11.7	14.8 ± 13.2	28.6 ± 21.5	8.2 ± 6.9	0.92 ± 0.04	0.95 ± 0.03	0.91 ± 0.03	0.63 ± 0.14	0.97 ± 0.01

Five prediction models (DAN: Dual attention network, RFR: Random forest regressor, XGB: XGBoost based model, SeqMO: Sequential multi-output model, CausalConv: Causal convolution based model) were compared based on mean absolute errors (MAE) and correlation between measured and predicted values (R).

and 2.0, respectively, with regard to the normalized glucose levels. Derivatives were obtained by the difference of values of unit time (5 min), and integration was conducted through the 60 min horizon.

For the controller models of SAC, policy, value, and two Q-function networks exist. In the policy network, two modules with linear layers, batch normalization, and LeakyReLU with a negative slope of 0.5, a linear layer, and a layer for absolute value $ReLU(x) + ReLU(-x)$ were connected to obtain the outputs of the mean and logarithm of the standard deviation for the action values. The value network consisted of three modules: one linear layer, batch normalization, and one ReLU activation layer. The structures of the Q-function networks were identical to those of the value network. For the reward r, the coefficient $\epsilon = 1$, and the decay coefficient $\gamma = 0.99$.

For the additional insulin dose of the safe actor, I_{int}^{max}, I_{int}^{min}, ΔI_{int} were set as 0, 1.5 and 0.125 (U/hr), respectively. The thresholds for updating parameters ρ_{susp}^{thr} and ρ_{int}^{thr} of the safe actor were determined as 0.05 and 0.1, respectively. For rescue to escape from severe hypoglycemia, G_{susp}^{oral} was set as 50 (mg/dL) to deliver 10 g of oral carbohydrate at the early stage. The initial values of G_{susp} and G_{int} were 110 and 225 (mg/dL), respectively.

APPENDIX B
TIME COMPLEXITY OF MODELS

In clinical settings, online learning is required to predict and control blood glucose levels. This means that not only the accuracy but also the computational burden in training should be considered. Asymptotic sketches for the time complexity can be derived.

Let n_{data} be the number of instances for training in one batch, d_{data} be the dimensionality of the training data, and k be the number of decision trees. Random forest and XGboost have a complexity of $O(kd_{data}n_{data}\log n_{data})$. In this study, $d_{data} = n_{feature}T_w$, the features of which are basic states, and T_w is the time window for training in the prediction models. Thus, the time complexity is $O(T_w k n_{feature} n_{data}\log n_{data})$.

For a neural network, the number of parameters W determines the time complexity, $O(W)$. Let m_i be the number of input units, m_h be the number of hidden units, and m_o be the

TABLE 4. Comparison between PID and SAC in blood glucose control.

Group	PID					SAC				
Patient	TIR	TAR level 1	TAR level 2	TBR level 1	TBR level 2	TIR	TAR level 1	TAR level 2	TBR level 1	TBR level 2
adolescent 1	0.686	0.307	0.149	0.007	0	0.517	0.439	0.097	0.043	0.007
adolescent 2	0.604	0.335	0.179	0.061	0.030	0.613	0.217	0.095	0.170	0.059
adolescent 3	0.807	0.188	0	0.005	0	0.616	0.345	0.073	0.038	0.069
adolescent 4	0.509	0.491	0.054	0	0	0.345	0.620	0.071	0.035	0.007
adolescent 5	0.835	0.165	0	0	0	0.825	0.132	0	0.043	0.016
adolescent 6	0.865	0.135	0	0	0	0.894	0.106	0	0	0
adolescent 7	0.530	0.470	0.108	0	0	0.670	0.321	0.052	0.009	0
adolescent 8	0.580	0.420	0	0	0	0.712	0.289	0.045	0	0
adolescent 9	0.710	0.290	0	0	0	0.266	0.734	0.012	0	0
adolescent 10	0.691	0.309	0	0	0	0.762	0.238	0	0	0
adult 1	0.641	0.297	0.049	0.063	0.002	0.590	0.410	0.052	0	0
adult 2	0.569	0.352	0.049	0.078	0.024	0.486	0.514	0.026	0	0
adult 3	0.609	0.391	0.066	0	0	0.898	0.102	0	0	0
adult 4	0.866	0.132	0	0.002	0	0.599	0.378	0.226	0.023	0
adult 5	0.767	0.233	0	0	0	0.837	0.155	0	0.009	0
adult 6	0.549	0.451	0.146	0	0	0.750	0.250	0.019	0	0
adult 7	0.425	0.563	0.290	0.012	0.002	0.351	0.505	0.304	0.144	0.036
adult 8	0.806	0.194	0	0	0	0.646	0.337	0	0.017	0
adult 9	0.726	0.222	0.021	0.052	0.009	0.592	0.340	0.109	0.068	0.007
adult 10	0.865	0.135	0	0	0	0.844	0.156	0	0	0
average	0.682 ±0.130	0.304 ±0.126	0.055 ±0.078	0.014 ±0.025	0.003 ±0.008	0.641 ±0.178	0.329 ±0.168	0.059 ±0.078	0.030 ±0.047	0.007 ±0.015
p-value	0.294	0.521	0.821	0.093	0.154					

The performance of PID and SAC was evaluated. Time per day within target glucose range (TIR), time below target glucose range (TBR), and time above target glucose range (TAR) were calculated. Threshold levels for TAR level 1 and TAR level 2 are 180 (mg/dL) and 250 (mg/dL), respectively. Threshold levels for TBR level 1 and TBR level 2 are 70 (mg/dL) and 54 (mg/dL), respectively. The Wilcoxon signed-rank test and the t-test were conducted to compare TIR, TAR and TBR between controls of PID and SAC.

number of output units. A feedforward neural network (FNN) has $W = m_i m_h + m_h m_o$, a recurrent neural network has $W = m_i m_h + {m_h}^2 + m_h m_o$, and a long short-term memory has $W = 4m_i m_h + 4{m_h}^2 + 3\, m_h + m_h m_o$, respectively.

The dual attention network (DAN) used in this study had an encoder and a decoder based on LSTM to predict future glucose levels and construct extended states. Let n_{epoch} be the number of epoch for training one batch of data. In this study, $m_h \sim m_i$ and $m_o \sim m_i$. For the input to the DAN, $m_i = n_{feature}$. The time complexity for LSTM in DAN is $O(n_{epoch} n_{data} T_w {n_{feature}}^2)$. In addition, DAN used attention mechanisms, which have a time complexity of $O({T_w}^2\, n_{feature})$ in training one instance according to [59]. The time complexity for the attention mechanisms in DAN is $O(n_{epoch} n_{data} {T_w}^2\, n_{feature})$. Thus, the total time complexity of DAN is $O(n_{epoch} n_{data} (T_w + n_{feature}) T_w n_{feature})$.

The SAC model used in this study has value, Q, and policy networks which consisted of feedforward neural networks. Extended states were used as inputs, and sequential information was contained in the extended states. In these networks, $m_h \sim m_i$ and $m_o = 1$. The time complexity of each network in the SAC is $O(n_{epoch} n_{data} {n_{feature}}^2)$.

APPENDIX C
COMPARISON OF PREDICTION MODELS

We compared the performance of several prediction models which have been recently proposed and were adopted in this study, as shown in table 3. A model based on causal convolution [57] showed the best performance, which does not have an encoder and a decoder in structure. Random forest regressor (RFR) showed a slightly better performance than that of the XGBoost based model [85]. In a comparison of encoder-decoder based models, the dual attention network (DAN) showed a better performance than that of a sequential multi-output model (SeqMO) [56].

In consideration of the time complexity of prediction models, RFR can be a feasible model for online forecasting. In addition, an encoder and a decoder are needed to construct extended states to capture glucose dynamics from the perspective of interpretability, DAN can be a feasible model for interpretability.

APPENDIX D
COMPARISON BETWEEN PID AND SAC

We conducted additional analyses to compare the performance of PID and SAC. Time per day within target glucose range (TIR), time below target glucose range (TBR), and time above target glucose range (TAR) are useful metrics [86] in blood glucose control. TIR, TBR, and TAR are shown in table 4. The performance from TIR, TBR, and TAR did not show statistical differences between PID and SAC.

REFERENCES

[1] J. K. Snell-Bergeon and R. P. Wadwa, "Hypoglycemia, diabetes, and cardiovascular disease," *Diabetes Technol. Therapeutics*, vol. 14, no. S1, pp. S-51–S-58, Jun. 2012, doi: 10.1089/dia.2012.0031.

[2] C. J. Sumner, S. Sheth, J. W. Griffin, D. R. Cornblath, and M. Polydefkis, "The spectrum of neuropathy in diabetes and impaired glucose tolerance," *Neurology*, vol. 60, no. 1, pp. 108–111, Jan. 2003, doi: 10.1212/wnl.60.1.108.

[3] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, Nov. 2019, Art. no. 107843, doi: 10.1016/j.diabres.2019.107843.

[4] X. Zhuo, P. Zhang, L. Barker, A. Albright, T. J. Thompson, and E. Gregg, "The lifetime cost of diabetes and its implications for diabetes prevention," *Diabetes Care*, vol. 37, no. 9, p. 2557, 2014, doi: 10.2337/dc13-2484.

[5] R. N. Bergman, C. Cobellit, and G. Toffolo, "Minimal models of glucose/insulin dynamics in the intact organism: A novel approach for evaluation of factors controlling glucose tolerance," *Trans. Inst. Meas. Control*, vol. 3, no. 4, pp. 207–216, Oct. 1981, doi: 10.1177/014233128100300404.

[6] R. N. Bergman, "Minimal model: Perspective from 2005," *Hormone Res. Paediatrics*, vol. 64, no. 3, pp. 8–15, 2005, doi: 10.1159/000089312.

[7] R. Hovorka, F. Shojaee-Moradie, P. V. Carroll, L. J. Chassin, I. J. Gowrie, N. C. Jackson, R. S. Tudor, A. M. Umpleby, and R. H. Jones, "Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 282, no. 5, pp. E992–E1007, May 2002, doi: 10.1152/ajpendo.00304.2001.

[8] C. Cobelli, C. D. Man, G. Toffolo, R. Basu, A. Vella, and R. Rizza, "The oral minimal model method," *Diabetes*, vol. 63, no. 4, pp. 1203–1213, Apr. 2014, doi: 10.2337/db13-1198.

[9] P. Vicini, A. Caumo, and C. Cobelli, "The hot IVGTT two-compartment minimal model: Indexes of glucose effectiveness and insulin sensitivity," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 273, no. 5, pp. E1024–E1032, Nov. 1997, doi: 10.1152/ajpendo.1997.273.5.E1024.

[10] C. Dalla Man, A. Caumo, R. Basu, R. Rizza, G. Toffolo, and C. Cobelli, "Minimal model estimation of glucose absorption and insulin sensitivity from oral test: Validation with a tracer method," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 287, no. 4, pp. E637–E643, Oct. 2004, doi: 10.1152/ajpendo.00319.2003.

[11] K. Thomaseth, A. Pavan, R. Berria, L. Glass, R. DeFronzo, and A. Gastaldelli, "Model-based assessment of insulin sensitivity of glucose disposal and endogenous glucose production from double-tracer oral glucose tolerance test," *Comput. Methods. Programs. Biomed.*, vol. 89, no. 2, pp. 132–140, Feb. 2008, doi: 10.1016/j.cmpb.2007.06.003.

[12] R. A. DeFronzo, J. D. Tobin, and R. Andres, "Glucose clamp technique: A method for quantifying insulin secretion and resistance," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 237, no. 3, p. E214, Sep. 1979, doi: 10.1152/ajpendo.1979.237.3.E214.

[13] T. Heise, E. Zijlstra, L. Nosek, S. Heckermann, L. Plum-Mörschel, and T. Forst, "Euglycaemic glucose clamp: What it can and cannot do, and how to do it," *Diabetes Obes. Metab.*, vol. 18, no. 10, pp. 962–972, Oct. 2016, doi: 10.1111/dom.12703.

[14] R. N. Bergman, Y. Z. Ider, C. R. Bowden, and C. Cobelli, "Quantitative estimation of insulin sensitivity," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 236, no. 6, p. E667, Jun. 1979, doi: 10.1152/ajpendo.1979.236.6.E667.

[15] E. Ferrannini and A. Mari, "How to measure insulin sensitivity," *J. Hypertension*, vol. 16, no. 7, pp. 895–906, Jul. 1998, doi: 10.1097/00004872-199816070-00001.

[16] F. C. Schuit, "Factors determining the glucose sensitivity and glucose responsiveness of pancreatic beta cells," *Hormone Res.*, vol. 46, no. 3, pp. 99–106, 1996, doi: 10.1159/000185004.

[17] C. D. Man, M. Campioni, K. S. Polonsky, R. Basu, R. A. Rizza, G. Toffolo, and C. Cobelli, "Two-hour seven-sample oral glucose tolerance test and meal protocol: Minimal model assessment of beta-cell responsivity and insulin sensitivity in nondiabetic individuals," *Diabetes*, vol. 54, no. 11, pp. 3265–3273, Nov. 2005, doi: 10.2337/diabetes.54.11.3265.

[18] H. Zisser, L. Robinson, W. Bevier, E. Dassau, C. Ellingsen, F. J. Doyle, and L. Jovanovic, "Bolus calculator: A review of four 'smart' insulin pumps," *Diabetes Technol. Therapeutics*, vol. 10, no. 6, pp. 441–444, Dec. 2008, doi: 10.1089/dia.2007.0284.

[19] C. Toffanin, H. Zisser, F. J. Doyle, 3rd, and E. Dassau, "Dynamic insulin on board: Incorporation of circadian insulin sensitivity variation," *J. Diabetes Sci. Technol.*, vol. 7, no. 4, pp. 928–940, Jul. 2013, doi: 10.1177/193229681300700415.

[20] H. R. Murphy, G. Rayman, K. Lewis, S. Kelly, B. Johal, K. Duffield, D. Fowler, P. J. Campbell, and R. C. Temple, "Effectiveness of continuous glucose monitoring in pregnant women with diabetes: Randomised clinical trial," *BMJ*, vol. 337, Sep. 2008, Art. no. a1680, doi: 10.1136/bmj.a1680.

[21] U. Holzinger, J. Warszawska, R. Kitzberger, M. Wewalka, W. Miehsler, H. Herkner, and C. Madl, "Real-time continuous glucose monitoring in critically ill patients: A prospective randomized trial," *Diabetes Care*, vol. 33, no. 3, pp. 467–472, Mar. 2010, doi: 10.2337/dc09-1352.

[22] G. Aleppo, K. J. Ruedy, T. D. Riddlesworth, D. F. Kruger, A. L. Peters, I. Hirsch, R. M. Bergenstal, E. Toschi, A. J. Ahmann, V. N. Shah, M. R. Rickels, B. W. Bode, A. Philis-Tsimikas, R. Pop-Busui, H. Rodriguez, E. Eyth, A. Bhargava, C. Kollman, and R. W. Beck, "REPLACE-BG: A randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes," *Diabetes Care*, vol. 40, no. 4, pp. 538–545, Apr. 2017, doi: 10.2337/dc16-2482.

[23] S. Weinzimer, K. Miller, R. K. Beck, D. Xing, R. Fiallo-Scharer, L. K. Gilliam, C. Kollman, L. Laffel, N. Mauras, K. Ruedy, W. Tamborlane, and E. Tsalikian, "Effectiveness of continuous glucose monitoring in a clinical care environment: Evidence from the juvenile diabetes research foundation continuous glucose monitoring (JDRF-CGM) trial," *Diabetes Care*, vol. 33, no. 1, pp. 17–22, Jan. 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19837791/#affiliation-1, doi: 10.2337/dc09-1502.

[24] B. P. Kovatchev, E. Renard, C. Cobelli, H. C. Zisser, P. Keith-Hynes, S. M. Anderson, S. A. Brown, D. R. Chernavvsky, M. D. Breton, L. B. Mize, and A. Farret, "Safety of outpatient closed-loop control: First randomized crossover trials of a wearable artificial pancreas," *Diabetes Care*, vol. 37, no. 7, pp. 1789–1796, Jul. 2014, doi: 10.2337/dc13-2076.

[25] D. M. Maahs, B. A. Buckingham, J. R. Castle, A. Cinar, E. R. Damiano, E. Dassau, J. H. DeVries, F. J. Doyle, S. C. Griffen, A. Haidar, and L. Heinemann, "Outcome measures for artificial pancreas clinical trials: A consensus report," *Diabetes Care*, vol. 39, no. 7, pp. 1175–1179, Jul. 2016, doi: 10.2337/dc15-2716.

[26] R. Gondhalekar, E. Dassau, and F. J. Doyle, "Periodic zone-MPC with asymmetric costs for outpatient-ready safety of an artificial pancreas to treat type 1 diabetes," *Automatica*, vol. 71, pp. 237–246, Sep. 2016, doi: 10.1016/j.automatica.2016.04.015.

[27] C. M. Ramkissoon, B. Aufderheide, B. W. Bequette, and J. Vehí, "A review of safety and hazards associated with the artificial pancreas," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 44–62, 2017, doi: 10.1109/rbme.2017.2749038.

[28] M. E. Wilinska, L. J. Chassin, C. L. Acerini, J. M. Allen, D. B. Dunger, and R. Hovorka, "Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes," *J. Diabetes Sci. Technol.*, vol. 4, no. 1, pp. 44–132, Jan. 2010, doi: 10.1177/193229681000400117.

[29] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator: New features," *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 26–34, 2014, doi: 10.1177/1932296813514502.

[30] R. Visentin, E. Campos-Náñez, M. Schiavon, D. Lv, M. Vettoretti, M. Breton, B. P. Kovatchev, C. D. Man, and C. Cobelli, "The UVA/Padova type 1 diabetes simulator goes from single meal to single day," *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 273–281, 2018, doi: 10.1177/1932296818757747.

[31] F. Chee, T. L. Fernando, A. V. Savkin, and V. V. Heeden, "Expert PID control system for blood glucose control in critically ill patients," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 4, pp. 419–425, Dec. 2003, doi: 10.1109/titb.2003.821326.

[32] G. Marchetti, M. Barolo, L. Jovanovic, H. Zisser, and D. E. Seborg, "An improved PID switching control strategy for type 1 diabetes," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 857–865, Mar. 2008, doi: 10.1109/TBME.2008.915665.

[33] 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, and M. E. Wilinska, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiol. Meas.*, vol. 25, no. 4, pp. 905–920, Aug. 2004, doi: 10.1088/0967-3334/25/4/010.

[34] L. Magni, D. M. Raimondo, L. Bossi, C. D. Man, G. De Nicolao, B. Kovatchev, and C. Cobelli, "Model predictive control of type 1 diabetes: An in silico trial," *J. Diabetes Sci. Technol.*, vol. 1, no. 6, pp. 804–812, Nov. 2007, doi: 10.1177/193229680700100603.

[35] M. Messori, G. P. Incremona, C. Cobelli, and L. Magni, "Individualized model predictive control for the artificial pancreas: In silico evaluation of closed-loop glucose control," *IEEE Control Syst. Mag.*, vol. 38, no. 1, pp. 86–104, Feb. 2018, doi: 10.1109/MCS.2017.2766314.

[36] Q. Sun, M. V. Jankovic, J. Budzinski, B. Moore, P. Diem, C. Stettler, and S. G. Mougiakakou, "A dual mode adaptive basal-bolus advisor based on reinforcement learning," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2633–2641, Nov. 2019, doi: 10.1109/jbhi.2018.2887067.

[37] 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 J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1223–1232, Apr. 2021, doi: 10.1109/JBHI.2020.3014556.
[38] S. Lee, J. Kim, S. W. Park, S.-M. Jin, and S.-M. Park, "Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 536–546, Feb. 2021, doi: 10.1109/JBHI.2020.3002022.
[39] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, "Deep reinforcement learning for closed-loop blood glucose control," in *Proc. Mach. Learn. Healthcare Conf.*, vol. 126, 2020, pp. 508–536.
[40] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, and S. Petersen, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: 10.1038/nature14236.
[41] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *J. Intell. Robot. Syst. Theory Appl.*, vol. 86, no. 2, pp. 153–173, May 2017, doi: 10.1007/s10846-017-0468-y.
[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
[43] A. Caumo and C. Cobelli, "Hepatic glucose production during the labeled IVGTT: Estimation by deconvolution with a new minimal model," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 264, no. 5, pp. E829–E841, May 1993, doi: 10.1152/ajpendo.1993.264.5.E829.
[44] G. Sparacino and C. Cobelli, "A stochastic deconvolution method to reconstruct insulin secretion rate after a glucose stimulus," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 5, pp. 512–529, May 1996, doi: 10.1109/10.488799.
[45] P. Magni, G. Sparacino, R. Bellazzi, and C. Cobelli, "Reduced sampling schedule for the glucose minimal model: Importance of Bayesian estimation," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 290, no. 1, pp. E177–E184, Jan. 2006, doi: 10.1152/ajpendo.00241.2003.
[46] G. Jiang and B. B. Zhang, "Glucagon and regulation of glucose metabolism," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 284, no. 4, pp. E671–E678, Apr. 2003, doi: 10.1152/ajpendo.00492.2002.
[47] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting subcutaneous glucose concentration in humans: Data-driven glucose modeling," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 246–254, Feb. 2009, doi: 10.1109/TBME.2008.2005937.
[48] D. J. Drucker, "The biology of incretin hormones," *Cell Metab.*, vol. 3, no. 3, pp. 153–163, Mar. 2006, doi: 10.1016/j.cmet.2006.01.004.
[49] R. N. Bergman, G. Bortolan, C. Cobelli, and G. Toffolo, "Identification of a minimal model of glucose disappearance for estimating insulin sensitivity," *IFAC Proc. Volumes*, vol. 12, no. 8, pp. 883–890, Sep. 1979, doi: 10.1016/S1474-6670(17)65505-8.
[50] H. Kirchsteiger, R. Johansson, E. Renard, and L. D. Re, "Continuous-time interval model identification of blood glucose dynamics for type 1 diabetes," *Int. J. Control*, vol. 87, no. 7, pp. 1454–1466, Jul. 2014, doi: 10.1080/00207179.2014.897004.
[51] T. M. Wallace, J. C. Levy, and D. R. Matthews, "Use and abuse of HOMA modeling," *Diabetes Care*, vol. 27, no. 6, pp. 1487–1495, Jun. 2004, doi: 10.2337/diacare.27.6.1487.
[52] M. Matsuda and R. A. DeFronzo, "Insulin sensitivity indices obtained from oral glucose tolerance testing: Comparison with the euglycemic insulin clamp," *Diabetes Care*, vol. 22, no. 9, pp. 1462–1470, Sep. 1999, doi: 10.2337/diacare.22.9.1462.
[53] E. J. Knobbe and B. Buckingham, "The extended Kalman filter for continuous glucose monitoring," *Diabetes Technol. Therapeutics*, vol. 7, no. 1, pp. 15–27, Feb. 2005, doi: 10.1089/dia.2005.7.15.
[54] D. J. Albers, M. Levine, B. Gluckman, H. Ginsberg, G. Hripcsak, and L. Mamykina, "Personalized glucose forecasting for type 2 diabetes using data assimilation," *PLOS Comput. Biol.*, vol. 13, no. 4, Apr. 2017, Art. no. e1005232, doi: 10.1371/journal.pcbi.1005232.
[55] E. Monte-Moreno, "Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques," *Artif. Intell. Med.*, vol. 53, no. 2, pp. 127–138, Oct. 2011, doi: 10.1016/j.artmed.2011.05.001.
[56] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, and J. Wiens, "Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1387–1395.
[57] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 414–423, Feb. 2020, doi: 10.1109/jbhi.2019.2931842.
[58] A. C. Miller, N. J. Foti, and E. Fox, "Learning insulin-glucose dynamics in the wild," in *Proc. Mach. Learn. Healthcare Conf.*, vol. 126, 2020, pp. 172–197.
[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6000–6010.
[60] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2627–2633.
[61] S. D. Patek, M. D. Breton, Y. Chen, C. Solomon, and B. Kovatchev, "Linear quadratic gaussian-based closed-loop control of type 1 diabetes," *J. Diabetes Sci. Technol.*, vol. 1, no. 6, pp. 834–841, Nov. 2007, doi: 10.1177/193229680700100606.
[62] F. Chee, A. V. Savkin, T. L. Fernando, and S. Nahavandi, "Optimal H_∞ insulin injection control for blood glucose regulation in diabetic patients," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 10, pp. 1625–1631, Oct. 2005, doi: 10.1109/TBME.2005.855727.
[63] S. U. Acikgoz and U. M. Diwekar, "Blood glucose regulation with stochastic optimal control for insulin-dependent diabetic patients," *Chem. Eng. Sci.*, vol. 65, no. 3, pp. 1227–1236, Feb. 2010, doi: 10.1016/j.ces.2009.09.077.
[64] G. M. Steil, "Algorithms for a closed-loop artificial pancreas: The case for proportional-integral-derivative control," *J. Diabetes Sci. Technol.*, vol. 7, no. 6, pp. 1621–1631, Nov. 2013, doi: 10.1177/193229681300700623.
[65] A. L. Alshalalfah, G. B. Hamad, and O. A. Mohamed, "Towards safe and robust closed-loop artificial pancreas using improved PID-based control strategies," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 8, pp. 3147–3157, Aug. 2021, doi: 10.1109/TCSI.2021.3058355.
[66] A. Mirzaee, M. Dehghani, and M. Mohammadi, "Robust LPV control design for blood glucose regulation considering daily life factors," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101830, doi: 10.1016/j.bspc.2019.101830.
[67] Y. Wang, J. Zhang, F. Zeng, N. Wang, X. Chen, B. Zhang, D. Zhao, W. Yang, and C. Cobelli, "'Learning' can improve the blood glucose control performance for type 1 diabetes mellitus," *Diabetes Technol. Therapeutics*, vol. 19, no. 1, pp. 41–48, Jan. 2017, doi: 10.1089/dia.2016.0328.
[68] C. Diuk, A. Cohen, and M. L. Littman, "An object-oriented representation for efficient reinforcement learning," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 240–247.
[69] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.
[70] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1861–1870.
[71] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–23.
[72] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery RL: Safe reinforcement learning with learned recovery zones," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4915–4922, Jul. 2021, doi: 10.1109/LRA.2021.3070252.
[73] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101836, doi: 10.1016/j.artmed.2020.101836.
[74] E. Daskalaki, P. Diem, and S. G. Mougiakakou, "An actor-critic based controller for glucose regulation in type 1 diabetes," *Comput. Methods. Programs. Biomed.*, vol. 109, no. 2, pp. 25–116, Feb. 2013, doi: 10.1016/j.cmpb.2012.03.002.
[75] E. Daskalaki, P. Diem, and S. G. Mougiakakou, "Model-free machine learning in biomedicine: Feasibility study in type 1 diabetes," *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0158722, doi: 10.1371/journal.pone.0158722.
[76] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 3145–3153.

[77] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140, doi: 10.1371/journal.pone.0130140.
[78] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.
[79] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.
[80] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.
[81] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 833–840.
[82] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1283–1292.
[83] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *Proc. Conf. Artif. Intell. (AAAI)*, vol. 32, no. 1, 2018, pp. 2095–2102.
[84] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, 2014, pp. 2155–2159.
[85] E. A. Pustozerov, A. S. Tkachuk, E. A. Vasukova, A. D. Anopova, M. A. Kokina, I. V. Gorelova, T. M. Pervunina, E. N. Grineva, and P. V. Popova, "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus," *IEEE Access*, vol. 8, pp. 219308–219321, 2020, doi: 10.1109/ACCESS.2020.3042483.
[86] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, E. Bosi, B. A. Buckingham, W. T. Cefalu, K. L. Close, and C. Cobelli, "Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019, doi: 10.2337/dci19-0028.

MIN HYUK LIM (Graduate Student Member, IEEE) received the B.S. degree in physics and the M.D. degree from Seoul National University, in 2007 and 2011, respectively. He is currently pursuing the Ph.D. degree with the Department of Biomedical Engineering, Seoul National University College of Medicine. His research interests include hybridization of machine learning and classical control, recommendation systems, interpretability, and explainability of the algorithm for clinical applications.

WOO HYUNG LEE received the M.D. degree from Seoul National University College of Medicine, in 2011, the Rehabilitation degree from Seoul National University Hospital, and the Ph.D. degree in biomedical engineering from Seoul National University College of Medicine, in 2020. He completed his residency training in 2016 and a fellowship in 2017. As a Junior Member of the Faculty, he has been an Assistant Professor with the Department of Rehabilitation Medicine, Seoul National University Hospital, since 2020. His research interests include motor recovery, pediatric neurorehabilitation, and swallowing rehabilitation.

BYOUNGJUN JEON received the B.S. degree in biophysics from the University of Washington, in 2014. He is currently pursuing the Ph.D. degree in interdisciplinary program in bioengineering from Seoul National University. His research interests include medical robots, machine learning, and drug screening.

SUNGWAN KIM (Senior Member, IEEE) received the B.S. degree in electronics engineering and the M.S. degree in control and instrumentation engineering from Seoul National University (SNU), in 1985 and 1987, respectively, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 1993. Since 2010, he has been a Professor with the Department of Biomedical Engineering, SNU College of Medicine. Prior to joining SNU, he worked as a Senior Aerospace Engineer with the National Aeronautics and Space Administration (NASA) Langley Research Center, USA. He is an Associate Fellow of AIAA.