IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 January 2023, accepted 12 March 2023, date of publication 20 March 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3259425

SURVEY

Reinforcement Learning Models and Algorithms for Diabetes Management

KOK-LIM ALVIN YAU[1], (Senior Member, IEEE), YUNG-WEY CHONG[2], XIUMEI FAN[3], CELIMUGE WU[4], (Senior Member, IEEE), YASIR SALEEM[5], (Senior Member, IEEE), AND PHEI-CHING LIM[6,7]

[1]Lee Kong Chian Faculty of Engineering and Science (LKCFES), Universiti Tunku Abdul Rahman (UTAR), Kajang, Selangor 47500, Malaysia
[2]National Advanced IPv6 Centre, Universiti Sains Malaysia (USM), Gelugor, Penang 11800, Malaysia
[3]School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, Shanxi 710048, China
[4]Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan
[5]Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB Ceredigion, U.K.
[6]Department of Pharmacy, Hospital Pulau Pinang, George Town, Penang 11090, Malaysia
[7]School of Pharmaceutical Sciences, Universiti Sains Malaysia (USM), Gelugor, Penang 11800, Malaysia

Corresponding authors: Kok-Lim Alvin Yau (yaukl@utar.edu.my) and Yung-Wey Chong (chong@usm.my)

This research was supported by Universiti Tunku Abdul Rahman (UTAR), the Publication Fund under Research Creativity and Management Office, Universiti Sains Malaysia, and the Universiti Sains Malaysia (USM) external grant (Grant Number: 304/PNAV/650958/U154).

ABSTRACT With the advancements in reinforcement learning (RL), new variants of this artificial intelligence approach have been introduced in the literature. This has led to increased interest in using RL to address complex issues in diabetes management. Using RL, a decision maker (or agent) observes decision-making factors (or state) from the dynamic operating environment, selects actions, and subsequently receives delayed rewards. The agent adapts its actions to changes in the operating environment to maximize its cumulative reward and improve system performance. This paper presents how various variants of RL have been used to improve diabetes management, such as a higher time in range during which the blood glucose level is within the normal range and a higher similarity between RL and physician's policies. Key highlights focus on the application of RL in diabetes management, including a taxonomy of the attributes of RL (e.g., roles and advantages), essential elements for training (e.g., data and simulators), representations of diabetes attributes in RL models, and variants of RL algorithms. In addition, this paper discusses open issues and potential future developments in the use of RL in diabetes management.

INDEX TERMS Actor-critic reinforcement learning, applied reinforcement learning, deep Q-network, deep reinforcement learning, diabetes, Markov decision process, multi-agent reinforcement learning, reinforcement learning.

I. INTRODUCTION

Diabetes mellitus, commonly known as diabetes, is a chronic metabolic disease. The number of diabetic patients has increased from 108 million in 1980 to 537 million in 2021, and it is expected to increase to 783 million by 2045 [1], [2], [3]. It is a concern that diabetes can lead to higher rates of intensive care unit (ICU) admissions and mortality, particularly among those also suffering from other noncommunicable and communicable diseases such as COVID-19 [4], [5].

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

A. AN OVERVIEW OF DIABETES

Diabetes occurs when a patient's pancreas: a) does not produce a sufficient level of insulin to convert glucose to glycogen for storage in liver and muscle cells to reduce the blood glucose level; and/ or b) produces more than a sufficient level of glucagon to convert glycogen to glucose to increase the blood glucose level [2]. Found in the pancreas, the β-cells synthesize and secrete insulin [6],

and the α-cells synthesize and secrete glucagon [7]. Both insulin and glucagon are hormones that maintain normal glucose homeostasis in the body to monitor and regulate the time-series blood glucose level, and their anomalous levels cause hyperglycemia and hypoglycemia conditions when the blood glucose level is higher- and lower-than-normal, respectively [8]. Typically, blood glucose levels above 180mg/dL are considered hyperglycemic, levels between 70mg/dL and 180mg/dL are considered normal, and levels below 70mg/dL are considered hypoglycemic [9], [10]. Both hyperglycemia and hypoglycemia can lead to both acute and chronic health complications. Long-term hyperglycemia can lead to cardiovascular disease [11], neuropathy [12], and retinal disease. Hypoglycemia can cause symptoms such as sweating, trembling, hunger, dizziness, and when the blood glucose level is too low in the brain, it can cause agitation, seizure, unconsciousness, coma, brain damage, and sudden death [13].

Two common types of diabetes are type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM). T1DM patients have autoimmune destruction of the β-cells causing absolute insulin deficiency. T2DM patients have β-cell failure and insulin resistance, whereby the liver and muscle cells do not respond well to insulin, causing insulin deficiency [14]. Both T1DM and T2DM affect the conversion of glucose to glycogen for storage, resulting in a higher level of blood glucose level. T2DM accounts for between 90% and 95% of all diabetes cases [15], and patients can suffer from both types simultaneously. While the root cause of diabetes remains unknown, a combination of factors affecting patients, including genetics, environment, and lifestyle, have been suggested [16]. Some of these factors are dynamic, such as when a patient changes lifestyle from time to time. For treatment, T1DM requires insulin injections or infusions. On the other hand, T2DM requires lifestyle and dietary interventions, and oral medications, and when these no longer work well enough, it also requires insulin injections or infusions. The traditional treatment approaches tend to consider a general patient population rather than individual patients, and ignore the dynamics of the operating environment including the patients.

B. AN OVERVIEW OF REINFORCEMENT LEARNING AND ITS APPLICATION TO DIABETES MANAGEMENT

In reinforcement learning (RL), an autonomous agent observes the factors that influence decision-making (known as states) in a dynamic operating environment, chooses actions based on its current understanding of the best course of action (known as policy), and receives feedback in the form of delayed rewards at a later time. The agent takes potential actions under different states in a trial-and-error manner in a large number of iterations so that the best-known policy converges to the optimal policy as time goes by. In other words, the agent learns the optimal policy that adapts to the dynamics of the operating environment through interactions with its operating environment as time goes by. This allows the agent to accumulate the maximum reward over time, which improves the overall performance of the system.

RL has been applied in a diverse range of healthcare applications, including recommending the right treatment [17] and medicine [18], and their dosage [18] and timing [19]. RL provides various advantages, particularly personalized care for different individuals rather than following general guidelines traditionally used for an average individual in treatments and medications. In addition to the capability of mimicking the physician's actions (e.g., prescriptions), RL learns the best possible course of action leading to good clinical outcomes in the long run in diabetes management. Equipped with RL, the system measures and receives inputs (e.g., the blood glucose level), and then estimates and recommends the insulin dosage for manual exogenous injection [20] or automated subcutaneous infusion [6] in a real-time closed-loop insulin-glucose system. The manual injection method uses a glucometer and it requires patients to perform a finger-prick test that takes one drop of blood from a finger at least four times a day (i.e., before each meal and sleeping). The automated infusion method uses a subcutaneous glucose sensor to measure and sample the blood glucose level every few minutes [21]. In [8] and [22], the sampling interval of the subcutaneous glucose sensor is five minutes due to the low dynamicity of the glucose-insulin-glucagon interaction.

C. CONTRIBUTIONS

This paper presents a review of the application of RL in diabetes management. including a taxonomy of the attributes of RL, essential elements for training using RL, representations of diabetes attributes in RL models, and RL algorithms. Furthermore, we present open issues and future directions for the application of RL in diabetes management to stimulate research interest in this area. Our paper focuses on the RL models and algorithms, so enabling technologies supporting diabetes management, such as secured communication, are not covered and have been presented in [23]. General reviews of the use of artificial intelligence have been presented in [24], [25], [26], [27], and [28], and a review of the representations of the RL model has been presented in [29]. Nevertheless, to the best of our knowledge, this paper is the first of its kind, presenting a comprehensive review of RL models and algorithms for diabetes management, including essential elements for training RL, such as data, system models, simulators, clinical studies, and so on.

D. ORGANIZATION OF THE PAPER

The rest of this paper is organized as follows. Section II presents background, including the Markov decision process (MDP) problem, the traditional RL algorithm, and a comparison between RL and other control algorithms. Section III presents a wide range of attributes related to applying RL in diabetes management, including the roles, advantages, data, system models, experiment designs, simulators, simulation

parameters and values, clinical studies, implementation, and performance measures. Section IV presents RL models, including the state, action, and reward representations. Section V presents RL algorithms and enhancements for diabetes management, including traditional RL, model-based RL, multi-agent RL, actor-critic RL, and deep RL. Section VI presents open issues. Finally, Section VII concludes this paper.

II. BACKGROUND

Diabetes management can be formulated as the MDP problem, and subsequently solved using RL. This section presents an overview of the MDP problem, the traditional RL algorithm, and a comparison between RL and other control algorithms to motivate the need for using RL in diabetes management. Table 1 presents a summary of general notations used in this paper.

A. THE MDP PROBLEM

The MDP problem is represented as a mathematical framework in the (S, A, P, R) tuple, where S represents the state space, A represents the action space, P represents the transition probability matrix, and R represents the reward function. In MDP, an autonomous agent makes sequential discrete-time decisions in the dynamic and noisy operating environment as time goes by [30]. Generally speaking, the MDP problem conforms to the decision-making process of physicians in diabetes management. Based on the state $s_t \in S$, the agent selects action $a_t \in A$ at time t, then it observes the next state s_{t+1} and receives the delayed reward $r_{t+1}(s_{t+1}) \in R : S \times A$ at the next time instant $t+1$. To collect more state information in diabetes management, the agent can perform state observation more frequently, such as a state observation every 3 minutes and an action selection every 120 minutes [14]. The state s_t transits to the next state s_{t+1} following the transition probability matrix $P(s_{t+1}|s_t, a_t)$, which represents the dynamics of the operating environment. The transition probability matrix satisfies the Markovian (or memoryless) property since a transition to the next state s_{t+1} depends on the current state s_t and action a_t only, rather than historical series of states and actions. The agent learns the optimal policy $\pi_t^* : S \rightarrow A$, which maps states in the state space S to optimal actions in the action space A, using the state s_t, action a_t, next state s_{t+1}, delayed reward $r_{t+1}(s_{t+1})$, and the transition probability matrix $P(s_{t+1}|s_t, a_t)$. Nevertheless, the transition probability matrix and the probability distribution of the reward function are generally unknown in reality.

B. THE TRADITIONAL RL ALGORITHM

The traditional RL algorithm solves the MDP problem without using the transition probability matrix, and it has been applied in diabetes management [14]. It is a model-free approach that allows the agent to discover through trial-and-error which actions are appropriate under different states by interacting with the operating environment as shown in Figure 1.

The agent's goal is to maximize its value function $V_{\pi_t}(s_t)$, which represents the goodness of being in a particular state s_t following the policy π_t. The value function $V_{\pi_t}(s_t)$, being the cumulative reward of a state s_t, is estimated by summing up the delayed and future (or discounted) rewards as time goes by as follows:

$$V_{\pi_t}(s_t) = E[r_{t+1}(s_{t+1}) + \gamma r_{t+2}(s_{t+2}) + \gamma^2 r_{t+3}(s_{t+3}) + \cdots] \quad (1)$$

TABLE 1. Summary of general notations and descriptions.

Notation	Description
$s_t \in S$	The state at time instant t. S is the set of possible states.
$a_t \in A$	The action at time instant t. A is the set of possible actions.
$P(s_{t+1}\|s_t, a_t)$	Transition probability from state s_t to state s_{t+1} under action a_t at time instant t.
$r_{t+1}(s_{t+1}) \in R$	Delayed reward under state s_{t+1} at time instant $t+1$. R is a reward function.
π_t	Policy at time instant t.
w	Weight.
θ_t	Network parameters at time instant t.
θ_t^-	Target network parameters at time instant t.
$V_{\pi_t}(s_t)$	Value function of state s_t following policy π_t at time instant t.
$Q_t(s_t, a_t)$	Q-value of the state-action (s_t, a_t) pair at time instant t.
$Q_t(s_t, a_t; \theta_t)$	Q-value of the state-action (s_t, a_t) pair under network parameters θ_t at time instant t.
$Q_t(s_t, a_t; \theta_t^-)$	Target Q-value of the state-action (s_t, a_t) pair under target network parameters θ_t^- at time instant t.
$A_{\pi_t}(s_t, a_t)$	Advantage of the state-action (s_t, a_t) pair following policy π_t at time instant t.
δ_i	Temporal difference of experience i,
y	Target.
$L(\theta_t)$	Loss function under network parameters θ_t at time instant t.
γ	Discount factor.
α	Learning rate.
ε	Exploration probability.
$\Gamma(k, \theta)$	Gamma distribution with the shape k and scale θ parameters.
$N_t(s)$	Number of visits to state s up to time instant t.
$N_t(s, a)$	Number of visits to state-action (s, a) pair up to time instant t.
$g_{i,t}$	The i^{th} observation of the blood glucose level observed at time instant t.
g_t	The blood glucose level at time instant t.
Δg_t	Glucose variability at time instant t.
$g_{\max,t}$	The highest measured blood glucose level up to time instant t.
$g_{\min,t}$	The lowest measured blood glucose level up to time instant t.
G_{target}	The target blood glucose level.
G_{hyper}	Hyperglycemia threshold.
G_{hypo}	Hypoglycemia threshold.
$N_{normal,k}$	Number of decision epochs in which the blood glucose level is within the normal range in day k.
$N_{hyper,k}$	Number of decision epochs in which hyperglycemia occurs in day k.
$N_{hypo,k}$	Number of decision epochs in which hypoglycemia occurs in day k.
I	Insulin infusion rate.
I_{basal}	Basal insulin secretion rate.
$I_{above-basal}$	Above-basal insulin secretion rate.

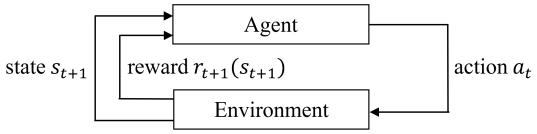

FIGURE 1. The closed-loop agent (e.g., the blood glucose regulator) interacts with the operating environment (e.g., the human body and the glucose variability) [30]. The agent selects actions (e.g., the insulin dosage) based on the state (e.g., the patient's clinical condition, such as the blood glucose level), and receives feedback in the form of the next state and delayed reward (e.g., the performance of time in range during which the blood glucose level is within the normal range).

which is equivalent to Equation (2) due to the recursive relationship:

$$V_{\pi_t}(s_t) = r_{t+1}(s_{t+1}) + \gamma V_{\pi_t}(s_{t+1}) \tag{2}$$

To do so, the agent learns the long-term reward of each state-action pair (s_t, a_t) and identifies the optimal action a_t^* under a particular state s_t in order to maximize rewards, or the value function $V_{\pi_t}(s_t)$, as time goes by. In Q-learning, which is a popular RL approach, the long-term reward, or the Q-value $Q_t(s_t, a_t)$, of each state-action pair is updated using the Q-function $Q_{t+1}(s_t, a_t)$ as follows:

$$\begin{aligned} Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[r_{t+1}(s_{t+1}) \\ + \gamma \max_{a \in A} Q_t(s_{t+1}, a) - Q_t(s_t, a_t)] \end{aligned} \tag{3}$$

where a higher learning rate $0 \leq \alpha \leq 1$ increases the significance of delayed and discounted rewards, and a higher discount factor $0 \leq \gamma \leq 1$ increases the significance of the discounted reward, which gives a longer-term view. The $r_{t+1}(s_{t+1}) + \gamma \max_{a \in A} Q_t(s_{t+1}, a) - Q_t(s_t, a_t)$ represents the temporal difference. Alternatively, the Q-function can be re-written as follows:

$$\begin{aligned} Q_{t+1}(s_t, a_t) = (1-\alpha) Q_t(s_t, a_t) + \alpha[r_{t+1}(s_{t+1}) \\ + \gamma \max_{a \in A} Q_t(s_{t+1}, a)] \end{aligned} \tag{4}$$

Without using the transition probability matrix $P(s_{t+1}|s_t, a_t)$, the agent must perform exploration and exploitation to learn the optimal policy π_t^*. The ε-greedy is a popular approach. Given a particular state s_t and the exploration probability $0 \leq \varepsilon \leq 1$, the agent chooses: a) a random action a_t with a small probability ε to learn the Q-value $Q_t(s_t, a_t)$ of state-action pairs (s_t, a_t) during exploration; and b) the best possible action a_t^* with probability $1-\varepsilon$ during exploitation. The best possible action a_t^* is chosen as follows:

$$a_t^* = \pi^*(s_t) = \arg\max_{a \in A} Q_t(s_t, a) \tag{5}$$

C. COMPARISON BETWEEN REINFORCEMENT LEARNING AND OTHER CONTROL ALGORITHMS

Traditionally, closed-loop control systems, including proportional–integral-derivative (PID), model predictive control (MPC), and iterative learning control (ILC), have been used to monitor and regulate the blood glucose level. In general, PID reacts to blood glucose levels outside the normal range, MPC addresses the delayed effects of meal ingestion and the pharmacological delay of insulin, and ILC addresses intra- and inter-patient glucose variabilities. A summary of the closed-loop control systems is presented in Table 2. RL offers three advantages compared to closed-loop control algorithms. First, while closed-loop control algorithms consider a general patient population under certain conditions, RL considers an individual patient, so unpredictable personal factors, such as stress, metabolic changes, and other diseases that cause changes in the blood glucose level (also known as glucose variability), are considered. Second, while MPC is more suitable for the deterministic environment, RL is suitable for the unpredictable and dynamic operating environment. Third, RL has lower computational complexity, so it is more efficient in online and real-time operations. Closed-loop control systems have been found to be inadequate in keeping blood glucose levels within the normal range after eating [31], [32], [33], [34], [35]. Nevertheless, the traditional RL approach has shortcomings which can be addressed by its variants and enhancements.

III. ATTRIBUTES OF RL IN DIABETES MANAGEMENT

From the perspective of diabetes management, this section presents the roles and advantages of RL. Then, it presents patient data, system models, simulators, simulation parameters and values, clinical studies, implementations, and experiment designs. Improved performance measures are also presented. Figure 2 presents a taxonomy of the attributes of RL in diabetes management.

A. ROLES OF RL IN DIABETES MANAGEMENT

In diabetes management, the unpredictability and dynamicity of the operating environment (e.g., the human body and the glucose variability) has urged RL to play a significant role in predictive tasks. The unpredictability and dynamics are attributed to a diverse range of factors, including the insulin dosage and time, meal size and time, the pharmacological delay of insulin, and so on. There are three main roles of RL in diabetes management as follows:

R.1 *Estimating the blood glucose level or the insulin dosage* is essential to monitoring and regulating the time-series blood glucose level continuously in a real-time closed-loop glucose-insulin system. The purpose is to maintain the blood glucose level within the normal range (or minimize the difference between the current and target glucose levels) to prevent postprandial hyperglycemia and hypoglycemia [6]. Estimating an accurate insulin dosage is a challenging task, even for physicians [23], due to three main challenges. First, glucose metabolism and insulin dosage have a nonlinear, complex, and time-varying relationship

TABLE 2. Comparison between RL and existing control algorithms for T1DM.

Approach	Description	Shortcoming
PID	A reactive approach that reacts only when the blood glucose level is outside the normal range instead of preventing it from happening during meal ingestion. Statistical methods adjust the insulin dosage (i.e., the insulin infusion rate) to match the blood glucose level.	Uses myopic feedback, which is affected by the pharmacological delay of insulin in its peak activation after an insulin infusion. So, it is inefficient against glucose variability even for patients with a strict lifestyle [31], [32].
MPC	A model-based approach that adjusts the insulin dosage (i.e., insulin infusion rate) to address the delayed effects of meal ingestion and the pharmacological delay of insulin, and prevent inappropriate insulin dosage that causes hypoglycemia [36].	First, the long-term prediction with multiple meal ingestion is inaccurate although the short-term prediction with single meal ingestion is accurate during the postprandial period [33]. Second, the performance is significantly affected by the accuracy of the glucose-insulin dynamics model, which considers a general patient population under certain conditions rather than an individual patient. So, unpredictable personal factors that cause glucose variabilities, such as stress, metabolic changes, and other diseases, are generally ignored [34].
ILC	An interactive approach that handles intra-patient (an individual) and inter-patient (a group of individuals) glucose variabilities through frequent data and feedback sampling to improve prediction iteratively [35].	Requires a high computing capability.
Traditional RL	A model-free approach that adjusts policy (e.g., the insulin dosage) to adapt to the dynamic operating environment (e.g., the blood glucose level) in order to maximize cumulative reward as time goes by.	First, the state space is large, which is also known as the curse of dimensionality. This increases complexity and computational time. Second, the convergence time, which is the time period incurred in searching for the optimal policy, is long.

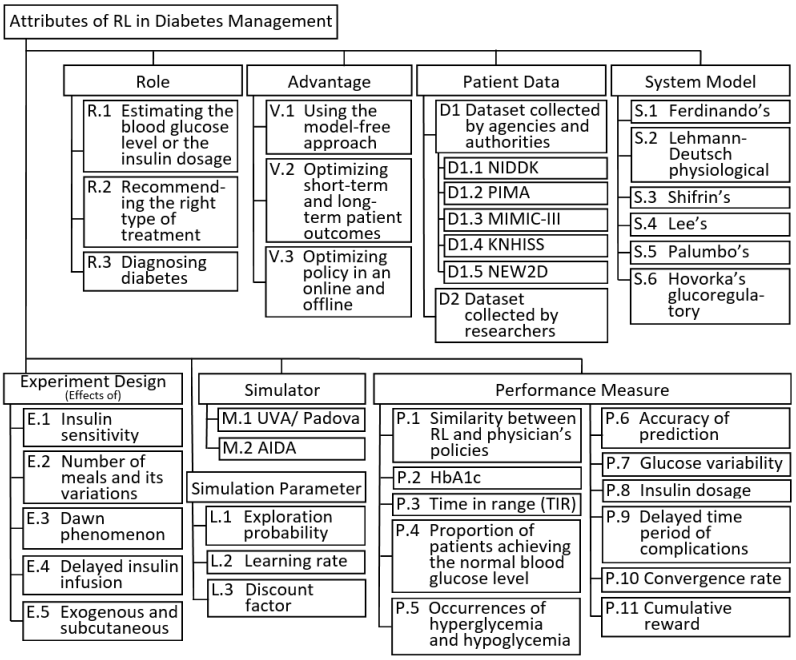

FIGURE 2. Attributes of RL in diabetes management.

affected by multiple factors, such as an individual's glucose metabolism and physical activity level. Second, various kinds of delays and inaccuracies have negative effects on prediction accuracy. One contributing factor is the measurement noise and delay caused by subcutaneous glucose sensors [37]. The sensors may have errors and hardware issues during calibration and normal operation, causing low stability and sensitivity. Understanding the characteristics of the consecutive sensor errors (e.g., highly interdependent) and noise (e.g., non-Gaussian or non-white) [38] is essential to detect and mitigate such errors and noise. Another contributing factor is the pharmacological delay of insulin in its peak activation, which is approximately 30 minutes for an insulin infusion and a few hours for an insulin injection depending on the types of insulin [6], [35]. In other words, the blood glucose level has a delayed response to insulin due to pharmacological delay. Third, endogenous characteristics (e.g., age and gender) and unpredictable exogenous events (e.g., meal ingestion, stress, and physical

activities) affect the postprandial blood glucose level. According to [39], there are at least 42 direct or indirect factors affecting the blood glucose level. To accurately estimate the blood glucose level or insulin dosage, RL must consider the three main challenges. To estimate the insulin dosage, RL must also consider insulin-on-board (IOB). IOB is the amount of bolus insulin that is still active from the previous bolus insulin injection, which must be considered if the injection is administered in less than 3.5 hours [10] to prevent hypoglycemia [40]. Subsequently, the estimated bolus insulin dosage is injected to regulate the blood glucose level.

R.2 *Recommending the right type of treatment* at each time interval, such as during each follow-up visit [26], [41]. The purpose is to enhance treatment effectiveness by providing the right medications and treatment regimen based on individual patients' clinical condition and medical history to maintain the blood glucose level within the normal range under comorbid conditions. The challenge is the complexity of treating the comorbid conditions because both hyperglycemia and hypoglycemia are known to cause other acute and chronic complications, such as cardiovascular disease and neuropathy (see Section I-A). RL must address this challenge to recommend a regimen comprised of multiple therapeutic classes for treating comorbid conditions, such as antihypertensive to treat high blood pressure, antihyperglycemic to treat glycemia, and lipid-lowering to treat cardiovascular disease [41]. Each therapeutic class has multiple pharmacological subclasses.

R.3 *Diagnosing diabetes* detects diabetes in its early stage. The purpose is to identify prediabetes accurately, with minimal false or delayed diagnoses [42]. In prediabetes, the right level of insulin is not produced, so the blood glucose level becomes higher-than-normal although it is still lower than that of diabetes [43]. The prediabetes stage has a higher risk of advancing to full-blown T2DM and other diseases, such as stroke and heart disease, so a healthy lifestyle with the right diet and body weight, and regular exercise help to prevent further escalation. The challenge is a large amount of data and information required to be processed to classify individuals accurately in diagnosis. RL has been applied to address this challenge by dynamically adjusting the weights of rules in a fuzzy rule-based method to enhance the consistency of rules applied [42].

B. ADVANTAGES OF RL IN DIABETES MANAGEMENT

In addition to addressing the shortcomings of the traditional closed-loop control systems (see Section II-C), namely PID, MPC, and ILC, the RL approach offers three main advantages in diabetes management as follows:

V.1 *Using the model-free approach*. The RL agent learns the appropriateness of different actions under different states through trial and error while interacting with the operating environment without the need for a model, such as the transition probability matrix, characterizing the dynamics of the operating environment. This is in contrast with traditional approaches, such as PID and MPC, that use mathematical glucose-insulin models comprised of ordinary and partial equations [44] to characterize the dynamics of the operating environment, particularly the time-varying physiological characteristics, such as the glucose, glucagon, and insulin levels [45]. Parameter estimation approaches, such as Bayesian inference [46] and deconvolution [47], have been used to estimate the coefficients of the models. Model-free RL has three main advantages when compared to model-based approaches. First, RL takes into account the long-term effects of treatments, while model-based approaches focus on short-term effects because they cannot model non-deterministic and time-varying effects. Second, RL does not need a large amount of prior knowledge (e.g., meal announcements and expert knowledge), including hidden factors such as comorbid diseases and the stress level, which is essential in developing mathematical models to cover a wide range of states (e.g., the glucose metabolism [48]) for characterizing the dynamics of the operating environment [14]. Instead, the RL agent adapts to the state by learning and selecting the optimal action that maximizes its cumulative reward, which considers hidden factors affecting the state and cumulative reward. Third, RL interacts with the real operating environment, while the model-based approach generally describes the real operating environment partially only [49].

V.2 *Optimizing short-term and long-term patient outcomes*. The RL agent makes sequential decisions that optimize delayed rewards (e.g., controlling the glucose level) and discounted rewards (e.g., improving the disease trajectory), leading to improvement in the overall short-term and long-term patient outcomes. Addressing short-term patient outcomes is essential to address transient metabolic responses, such as the rapid increase (decrease) of the postprandial blood glucose level that results in hyperglycemia (hypoglycemia). This is in contrast with traditional approaches, such as k-nearest neighbors and regression, that optimize short-term patient outcomes based on the current symptoms only [10].

V.3 *Optimizing policy in an online (or incremental) and offline manner*. The RL agent has the capability of learning in an online and offline manner. In general,

most RL approaches rely on online learning, which uses data provided to the agent in real time. The RL agent observes states and rewards from the operating environment, then uses the new information to learn in an incremental manner, contributing to a lower computational complexity. RL has also been shown to learn using offline learning based on a large amount of data sourced from databases in [50]. Multimodal data from multiple sources are collected and combined to provide a large dataset. Examples of data include: a) subpopulation information, such as environment, the susceptibility to diabetes, and responses to specific treatments; and b) personal information, such as patients' demographics, lifestyles, and clinical conditions (e.g., the blood glucose level) [50]. The offline learning approach enables the RL agent to better understand and treat diseases, which supports precision medicine [50]. Since RL performs online and offline learning approaches, it is in contrast with traditional approaches that perform offline learning only.

C. ESSENTIALS FOR TRAINING RL IN DIABETES MANAGEMENT

This section presents essentials for training RL in diabetes management, including patient data, system models, simulators, simulation parameters and values, clinical studies, implementation, and experiment designs.

1) PATIENT DATA

Patient data is collected for training RL agents. Datasets are either provided by agencies and authorities or collected by researchers conducting the investigations. While some investigations use more data and data types to increase the accuracy of prediction, others reduce the need for some data, such as medical check-up measurements (e.g., fasting plasma glucose, BMI, and blood pressure), for patients' convenience. Table 3 summarizes various data types collected from patients.

TABLE 3. Data types collected from patients and used in training RL algorithm.

Data types	Description
Demographic	Age and gender [13], [51]
Physical examination	Body mass index, temperature, blood pressure (i.e., systolic and diastolic), respiratory rate, and heart rate [13], [51]
Medical history/ condition	Duration of diabetes, past complications and cases (e.g., nephropathy (or kidney disease), hypertension, dyslipidemia, and hypoglycemia, and mortality [13], [51]
Lifestyle	Physical activity level, and habits (e.g., smoking) [13]
Previous prescriptions	Oral antidiabetic drugs (e.g., metformin, sulfonylureas or glinides, dipeptidyl peptidase-4, alpha-glucosidase inhibitors, sodium-glucose co-transporter 2 inhibitors, and thiazolidinediones), and insulin (e.g., basal, bolus, and premix) [13]
Lab test results	Potassium, chloride, sodium, hemoglobin, blood urea, HbA1c, serum creatinine, fasting blood glucose level, blood glucose level, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, total cholesterol, triglyceride, bicarbonate, Anion gap [13], [51]

a: DATASET COLLECTED BY AGENCIES AND AUTHORITIES (D.1)

Various datasets have been collected from diabetic patients with consent, and the datasets are useful for investigating diabetes management, particularly in training RL agents. Five different datasets have been used to train RL agents as follows:

D1.1 *The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dataset* was collected from patients within ten years, and it has been used to train RL agents [10]. The dataset consists of patient information (e.g., blood glucose level, weight, and carbohydrate), treatment, and diabetes-related complications [52].

D1.2 *The PIMA Indian women diabetic dataset* was collected from 768 patients, and it has been widely used to train RL agents [43]. The dataset consists of patient information, including age, blood pressure, BMI, skin thickness, blood glucose level, insulin level, diabetes pedigree function, and outcomes. Three attributes, including age, BMI, and blood glucose level, were selected to train RL agents in [43].

D1.3 *The medical information mart for intensive care III (MIMIC-III) version 1.4 dataset* was collected from 400,000 patients admitted to the ICU of the Beth Israel Deaconess Medical Center from 2001 to 2012, and it has been used to train RL agents [53]. The dataset consists of patient information, including demographics, mortality, and lab test results. The dataset has static variables (e.g., age, gender, chloride level, hemoglobin level, and mortality rate) and dynamic variables (e.g., blood glucose level, insulin level, temperature, creatinine, and bicarbonate). 190 diabetic patients were selected using the ICD-9 code 250.1 to train RL agents in [51].

D1.4 *The Korean national health insurance sharing service (KNHISS) dataset* was collected from 1 million patients, out of which 38,127 were diabetic patients, and it has been used to train RL agents in [50]. The dataset consists of patient information, including medical history, medications, the fasting plasma glucose groups (i.e., normal, prediabetes, or diabetes), and whether the patient has chronic complications (e.g., diabetic retinopathy and diabetic cataract) and acute complications (e.g., myocardial infarction and heart failure) caused by diabetes, a higher risk of developing diabetes, and up to four years of diabetes.

D1.5 *The newly diagnosed type 2 diabetic patients in China (NEW2D) dataset* was collected from 3,893 patients from June 2012 to February 2014, and it has been used to train RL agents [13]. The dataset consists of a wide range of patient information, including demographic, physical examination, medical history, lifestyle, previous prescriptions, and lab test results, as shown in Table 3.

There are criteria for choosing patients participating in data collection to ensure data consistency and exclude certain patients, such as those diagnosed with prediabetes. In [41], selected T2DM patients must fulfill three criteria: a) had been classified as diabetic patients with the E10-E14 codes based on the *international classification of diseases, tenth revision* (ICD-10) for at least twice; b) had anomalous HbA1c level (i.e., HbA1c $\geq$ 6.5%) for at least twice; and c) had T2DM prescription (other than acarbose and metformin) for at least once. In [13], selected T2DM patients from the NEW2D dataset (D1.5) must fulfill two criteria: a) each patient has at least two visits, namely the first and follow-up visits; and b) had anomalous HbA1c level during the two visits. In [50], based on the KNHISS dataset (D1.4), two criteria must be fulfilled, including: a) had been classified as diabetic patients based on ICD-10; and b) had an oral prescription to reduce the blood glucose level for at least 30 days. The selected data was from Jan 1, 2003 to December 31, 2013 or the date of death, whichever was earlier [50].

The dataset is cleaned by removing data entries that contains inappropriate, misleading, and redundant values, such as blood glucose levels and BMI values less than zero, as shown in [43]. The cleaned dataset becomes smaller, such as from 768 to 392 patients in the PIMA Indian women diabetic dataset (D1.2) [43]. Missing data may be inevitable. For instance, in [41], out of the patient data, 1% has missing blood pressure and BMI, 8% has missing HbA1c, and so on. One common approach to handling missing data is to replace it with values observed in previous encounters, as mentioned in [41]. However, if a large proportion of variables (e.g. more than half) in a patient data are missing, the data entry is removed, as suggested in [51].

b: DATASET COLLECTED BY RESEARCHERS (D.2)

Some researchers collect their own data for investigations. An example is [13] which collects data using case report forms in addition to using the existing NEW2D dataset (D1.5). Data is collected from around 2,800 patients who made first and follow-up visits (e.g., five visits every three months within one year) at 80 hospitals. Nevertheless, the number of participating patients reduced over time when they fail to return for follow-up visits within one year. The criteria for choosing patients participating in data collection includes: a) demographic information (i.e., age is at least 20 years old); b) medical history and event (e.g., time when patients were diagnosed, which is within the past six months); c) physical examination; d) prescriptions; and e) lifestyles (e.g., neither pregnant nor lactating). Measurements related to diabetes are collected, particularly the HbA1c level, which represents the average blood glucose level over the past 8 to 12 weeks [54]. The HbA1c level is categorized into high (i.e., HbA1c $>$ 9%), medium (i.e., 7% $\leq$ HbA1c $\leq$ 9%), and low (i.e., HbA1c $<$ 7%).

2) SYSTEM MODELS

A system model, which provides the operating environment, represents the glucose-insulin dynamics of an average virtual patient (AVP) in diabetes management. During training, an agent interacts with AVP(s) and learns the optimal policies. The system model must be realistic and complete, so the US Food and Drug Administration (FDA) encourages the use of its approved system models when running simulations, which helps to avoid using animals in preclinical tests [55]. The rest of this subsection explains various system models proposed for investigating RL in diabetes management.

S.1 *Ferdinando's system model*, which is created based on experiments and expert knowledge, comprises delayed differential equations characterized by blood glucose and insulin levels [56]. The insulin is absorbed into the blood following first-order dynamics (or the single delay model) under the skin. By adjusting the system model parameters, the model generates different clinical conditions (e.g., the rate of insulin-independent glucose uptake and the apparent distribution volume for glucose) to model different patients. Using the system model, the RL agent estimates IOB and selects the right insulin dosage during action selection [14].

S.2 *Lehmann-Deutsch physiological model* comprises multiple differential equations to represent the glucose-insulin dynamics of T1DM patients [57] and the variation of multiple parameters with time, including the plasma glucose concentration, glucose absorption, net hepatic glucose balance, volume of the distribution of glucose, the reference glucose level, the renal excretion of glucose, plasma insulin concentration, insulin in the remote compartment, the fractional disappearance rate of insulin, insulin distribution volume, and the exogenous insulin dosage [58]. Since generic parameters for the model reduce variability caused by measurement noise and inappropriate state estimations, the model uses different parameters specific to each patient.

S.3 *Shifrin's system model*, which is created based on experience, comprises a transition probability matrix to represent personal reactions to insulin dosage [10]. The model has a set of states (i.e., the blood glucose level, carbohydrate intake, and time duration since the last treatment) and a set of actions (i.e., insulin dosage). The transition from one state to another upon taking an action follows the transition probability matrix, which is complex due to different

types of hormones, carbohydrate intakes, and other processes.

S.4 *Lee's system model*, which is created based on experiments, comprises the gamma distribution $\Gamma(k, \theta)$ representing the insulin time-action profile [6], [59]. The profile represents the pharmacokinetics and pharmacodynamics properties of a patient's insulin analogues. k is the shape parameter and θ is a scale parameter, which is based on a patient's insulin activation peak time. The $\Gamma(k, \theta)$ distribution is min-max normalized to generate additional discount factors f_n and F_n (see Section V-H2). The peak of the distribution is identified using the euglycemic clamp test through simulation. In the simulation, a subcutaneous insulin pump infuses 0.2U (or 1,200pmol) insulin once every minute, providing a total of 6U insulin required to control the peak of the blood glucose level in a 30-minute postprandial period. Using the proportional integral controller [60], glucose is infused intravenously at a rate and time period determined based on the insulin time-action profile. The time duration from the insulin infusion to the peak of its action determines the θ value of the $\Gamma(k, \theta)$ distribution.

S.5 *Palumbo's system model*, which is created based on expert knowledge, comprises [61]: a) a delayed differential equation characterized by blood glucose and insulin levels; and b) a nonlinear function characterized by the insulin dosage (i.e., insulin infusion rate). Using the system model, the RL agent selects the right insulin dosage during action selection [62].

S.6 *Hovorka's glucoregulatory model*, which is created based on experience, comprises parameters sampled from a prior log-normal distribution and parameter correlations sampled from healthy individual data [63]. The parameters oscillate with random frequency and phase, which are caused by the glucose and insulin fluxes, to generate day-to-day glucose variability. The peak meal absorption time changes to generate slow and fast carbohydrate absorption.

In general, an existing system model is selected or a new system model is designed when a simulator (see the next subsection) is not used in simulation studies [62].

3) SIMULATORS

Simulation allows random actions to be selected and tested under different glucose-insulin dynamics in *in silico* trials without safety concerns. Another advantage is that simulation can be completed within a short time, allowing for a much longer simulated time. For simulating instantaneous changes in the blood glucose level (also known as glucose variability), the simulated time can be in the range of hours for a simulation step [64]. For simulating first and follow-up visits, the simulated time can be an interval of several weeks for a simulation step and several months for a simulation episode [65]. The numbers of simulation episodes and steps reduce when the convergence rate of the RL approach increases. Specifically, higher convergence rates reduce the time period for RL to search for the optimal policy. Some important initial values of simulation parameters include the blood glucose level (e.g., 8.85 mM [14]) and the insulin level (e.g., 58.95 pM [14]). To reduce the effects of different initial values, each simulation run is repeated (e.g., 20 times [51]), and then averaged. Two main simulators, which can be based on datasets (see Section III-C1), have been used in investigating RL in diabetes management as follows:

M.1 *The UVA/ Padova simulator* [66], which has been approved by the US FDA, has been used to simulate a closed-loop artificial pancreas in [6], [21], and [64]. The simulator incorporates processes related to glucose-insulin dynamics to simulate the effects of meal ingestion and insulin on the blood glucose level among AVPs. The processes include processing consumed food in the gut, producing glucose in the liver, absorbing exogenous insulin, and ingesting unannounced meals. There are three groups of AVPs with different levels of glucose metabolism and insulin dosage, including 100 adults, 100 adolescents, and 100 children [51], [66]. A subset of AVPs can be selected for simulation runs, such as 10 adults and 10 adolescents in [6], and 100 adults in [21]. Meal ingestion with random parameters helps to simulate patients' behavior in a realistic manner. Each AVP taking 2,000 unannounced meals are randomized using three different random seeds in [6]: a) the meal size is selected from a normal distribution with a mean of 65g and a standard deviation of 17g; b) the time period in between meals is selected from another normal distribution with a mean of 10 hours and a standard deviation of 1 hour; and c) a meal is skipped with a probability 0.05. Simglucose [67] is the python implementation of the UVA/ Padova simulator.

M.2 *The AIDA simulator* is an interactive educational simulator [68], [69], [70] that simulates a closed-loop artificial pancreas, and it has been used in [10], [57], and [71]. Similar to the UVA/ Padova simulator (M.1), the AIDA simulator incorporates processes related to glucose-insulin dynamics to simulate the effects of meal ingestion and insulin on the blood glucose level among AVPs. However, there are two main differences compared to the UVA/ Padova simulator [72]. The AIDA simulator considers insulin and glucagon dynamics, while the UVA/ Padova simulator does not consider glucagon dynamics. The AIDA simulator does not consider inter-patient (a group of individuals) glucose variabilities, while the UVA/ Padova simulator considers.

Apart from the popular simulators, *in silico* simulations have been written in MATLAB [73], [74] and Python [64]. Using MATLAB, a group of T1DM AVPs, which consist of 10 adults, 10 adolescents, and 8 children, are developed [75]. Each AVP takes three to four meals with random carbohydrate contents measured in grams at different times for ten days. Random uncertainties distributed between -50% and $+50\%$ in a uniform manner are introduced to account for errors made by patients in the carbohydrate estimations of meals. Using the Python toolkit called OpenAIGym [76], the RL agent is developed, tested, and analyzed under the operating environment provided by the UVA/ Padova simulator. The dynamic equations and parameters of UVA/ Padova are imported into OpenAIGym. The Optuna library is used to automate hyperparameter tuning to determine the best possible hyperparameters of RL and deep reinforcement learning (DRL), including the deep neural network parameters, by running a large number of simulations in a trial-and-error manner.

4) SIMULATION PARAMETERS AND VALUES

There are three main parameters requiring careful consideration when running RL algorithms as follows:

- L.1 Higher *exploration probability* $0 \leq \varepsilon \leq 1$ increases exploration and reduces exploitation. For increasing exploitation and improving stability, the exploration probability is set to small values, such as $\varepsilon = 0.02$ and $\varepsilon = 0.1$ [51]. In [65], the exploration probability of a state reduces gradually with $\varepsilon_t = N_0/(N_0 + N_t(s))$ as the number of visits $N_t(s)$ to the state s up to decision epoch t increases, where N_0 is a constant. This has been shown to improve the convergence rate.
- L.2 Higher *learning rate* $0 \leq \alpha \leq 1$ increases the step size of an update affecting the amount of changes of Q-values in RL and network parameters in DRL. For improving stability, the learning rate is set to small values, such as $\alpha = 0.02$ [51]. In [65], the learning rate of a state-action pair reduces gradually with $\alpha_t = 1/N_t(s, a)$ as the number of visits $N_t(s, a)$ to the state-action pair (s, a) up to decision epoch t increases. This has been shown to improve the convergence rate.
- L.3 Higher *discount factor* $0 \leq \gamma \leq 1$ increases the effect of the long-term reward, and hence improving the glycemic outcomes in the future [8]. The discount factor γ is generally set to a high value, such as $\gamma = 0.90$ and $\gamma = 0.95$ [51], to take account of delayed effects due to the pharmacological delay of insulin infusion.

5) CLINICAL STUDIES

Clinical studies of proposed RL solutions for diabetes management have been conducted, albeit the low number of such investigations in the literature compared to simulation. In [7], the clinical study investigates using RL to estimate the blood glucose level. The study involves 23 T1DM patients spanning five weeks. There are two scenarios: a) scenario 1 has fixed meal size and time; and b) scenario 2, which is more realistic, has random meal sizes and times, and random errors are introduced in the carbohydrate estimation of meals. Both simulation and clinical studies have been shown to achieve similar results. In [21], the clinical study investigates using RL to estimate and adjust the insulin dosage (i.e., the basal rate and the bolus insulin dosage), which is based on the insulin-to-carbohydrate ratio representing the carbohydrate contents that one unit of insulin covers. The study involves individual T1DM patients spanning fourteen weeks. The automated subcutaneous infusion is comprised of a subcutaneous glucose sensor and a subcutaneous insulin pump. The insulin dosage is adjusted based on: a) the blood glucose level of the day before; and b) the transfer entropy from the insulin to glucose signals. Incorporated with RL, both manual exogenous injection and automated subcutaneous infusion approaches have been shown to achieve closely similar performance.

6) IMPLEMENTATION

Details about the implementation of proposed RL solutions in diabetes management have been limited in the literature. For training, the agents' knowledge, such as Q-tables and policy networks (or deep neural networks in DRL), is trained on the cloud in the web server. For action selection, Q-tables and policy networks are transferred from the cloud to the AndroidAPS APP [77], which uses the Pytorch mobile library [78], [79], in smartphones [80]. Using Bluetooth, the AndroidAPS APP communicates with the subcutaneous glucose sensor to collect measurements and the subcutaneous insulin pump to infuse insulin. Data can be collected and sent to the cloud in the web server for training purpose. In [8], the RL approach is implemented using TensorFlow, which can be converted using TensorFlow Lite for use in smartphones [81], [82].

7) EXPERIMENT DESIGNS

Proposed RL approaches are investigated in simulations, clinical studies, and implementations to understand their performance achieved under simulated or real circumstances. In general, investigations show the effects of various aspects (e.g., insulin sensitivity and the number of meals) on RL performance. The effects of the following aspects are of interest:

- E.1 *Effects of insulin sensitivity* are investigated. Higher insulin sensitivity reduces the amount of insulin required to reduce the blood glucose level, so it increases the responsiveness of insulin infusion or injection. In [65], the peripheral tissue insulin sensitivity and the hepatic tissue insulin sensitivity vary from the nominal value randomly by either -10%, $+10\%$, or 0%. Variations can be up to $\pm 25\%$ [21] and $\pm 30\%$ [6]. Variations tend to be

higher in adults (e.g., ±30%) than adolescents (e.g., ±20%) [8].

E.2 *Effects of the number of meals and its variations*, which can be single or multiple meals within a time period, are investigated. Generally speaking, meals are measured in grams of carbohydrate or mmol of the blood glucose level. The amount of carbohydrate in meals follow a meal schedule, and it is unannounced in most cases except a few [21]. For single-meal scenarios, a single meal is given to the patient. The preprandial and postprandial fasting periods are generally long (e.g., 9 hours [6]). For multiple-meal scenarios, multiple meals are given to the patient, which is more realistic. Different meal sizes and times have been used in experiments. In [6], three meals are given to patients, specifically 40g carbohydrate for breakfast at 8:00 a.m., 80g for lunch at 1:00 p.m., and 60g for dinner at 9:00 p.m. In [14], the glucose levels of the meals are 9, 10, and 10.5 mM for breakfast, lunch, and dinner, respectively. In [8], four meals are given to patients, specifically 70g carbohydrate at 7:00 a.m., 30g at 10:00 a.m., 110g at 2:00 p.m., and 90g at 9:00 p.m. Randomizations are introduced to meal size and time [21] to represent noise, such as advances and delays in meal time. The meal size is selected from a uniform distribution with a standard deviation of $[-30\%, +10\%]$ [8], in which the skewed range reflects that underestimation is more common than overestimation in reality [83]. The meal time is selected from a uniform distribution with a standard deviation of ±60 minutes [8] or ±10 minutes [6]). Variations of meal absorption (e.g., 30%) and carbohydrate bioavailability (e.g., 10%) are also set in [8].

E.3 *Effects of the dawn phenomenon* are investigated. The dawn phenomenon is an increase of up to 30mg/dL in the blood glucose level between 3:00 a.m. and 7:00 a.m. in the early morning on daily basis [6]. In some cases, the insulin sensitivity drops by up to 50% from its nominal value, and such significant change can happen within even 30 minutes causing hyperglycemia [84].

E.4 *Effects of delayed insulin infusion or injection*, which introduces some delays to insulin infusion or injection, such as when a patient is unable to administer insulin at the right time (e.g., a 6.5-minute delay in [14]), is investigated. In addition to the pharmacological delay of insulin in its peak activation after an insulin infusion or injection, the overall delay can lead to hyperglycemia. A delayed insulin infusion or injection can also cause hypoglycemia in the next dosage due to high IOB.

E.5 *Effects of the manual exogenous injection and automated subcutaneous infusion approaches* are investigated. To measure the blood glucose level, the manual exogenous injection approach requires patients to perform a finger-prick test at least four times a day. On the other hand, the automated subcutaneous infusion approach uses a subcutaneous glucose sensor to measure and sample the blood glucose level every few minutes [21]. Subsequently, the manual exogenous injection approach administers the estimated insulin four times a day [20]. On the other hand, the automated subcutaneous infusion approach uses a subcutaneous insulin pump to infuse insulin every short time interval, such as 1,200pmol (or 0.2U) insulin once every minute [8], [22] in a real-time closed-loop insulin-glucose system.

D. PERFORMANCE MEASURES

RL has been shown to improve various performance measures in the literature, including:

P.1 *Higher similarity between RL and physician's policies* shows consistency between policies made based on the RL agent's and clinical knowledge [13].

P.2 *Lower HbA1c level* slows down the progression of diabetes which helps to reduce the average blood glucose level within a time period (e.g., over the past 8 to 12 weeks [13]).

P.3 *Higher time in range (TIR)* increases the percentage of time duration (e.g., in the most recent 24 hours) in which the blood glucose level is within the normal range (between 70mg/dL and 180mg/dL) [8], [9], [85], [86]. The opposite is the GRADE (glycemic assessment diabetes equation) measure, which is the percentage of time duration in which the blood glucose level is outside the normal range at $<$70mg/dL (hypoglycemia) and $>$180mg/dL (hyperglycemia), particularly during meal ingestion [10]. GRADE can be calculated using a glucometer in the manual injection method or a subcutaneous glucose sensor in the automated infusion method. Higher TIR reduces GRADE. A range of thresholds between 40mg/dL [10] and 325mg/dL [57] have been used in the literature while showing an improvement on TIR. A smaller range of the selected thresholds has been widely used due to its capability in reducing the risk of long-term complications of diabetes despite being more challenging to be achieved.

P.4 *Higher proportion of patients achieving the normal blood glucose level* increases the number of patients achieving the normal blood glucose level [51].

P.5 *Lesser occurrences of hyperglycemia and hypoglycemia* reduce the number of times when the blood glucose level is outside the normal range.

P.6 *Higher accuracy of prediction* improves the effectiveness of insulin infusions and injections. The confusion matrix classifies the outcomes of insulin infusions into one of the four categories, namely true positive, false positive, true negative, and false negative as shown in Figure 3 [43].

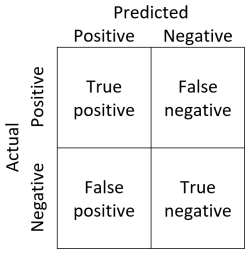

FIGURE 3. A confusion matrix.

P.7 *Lower glucose variability*, which can be measured based on standard deviation, reduces the fluctuations of the blood glucose level within a time period [10]. The estimated insulin dosage is more accurate when the blood glucose level is more stable.

P.8 *Lower insulin dosage* reduces the insulin dosage while achieving the normal blood glucose level, which helps to reduce side effects [51], such as weight gain.

P.9 *Longer delayed time period of chronic and acute complications* increases the average time period from diabetes diagnosis to the occurrence of chronic or acute complications caused by diabetes.

P.10 *Higher convergence rate*, which is the RL performance measure, reduces the time period in searching for the optimal policy.

P.11 *Higher cumulative reward*, which is the RL performance measure, improves the glycemic outcomes in diabetes management [64].

IV. RL MODELS FOR DIABETES MANAGEMENT

Different state, action, and reward representations have been proposed to perform different roles of diabetes management (see Section III-A) as seen in Table 4. Generally speaking, in diabetes management, the state represents the physiological condition. The action represents insulin dosage, and the medicine and treatment regimen. The reward represents the glycemic outcomes. The decision epoch is generally synchronized with events, such as meal ingestion and insulin infusions. The rest of this section presents various ways to design the state, action, and reward representations.

A. STATE REPRESENTATION APPROACHES

Each state has either a single or multiple sub-states representing decision-making factors, which can be observations, estimations, measurements, and so on. The rest of this subsection presents sub-state representations and approaches to simplify the state space.

1) SUB-STATE REPRESENTATIONS

In general, sub-states have been designed to represent short- and long-term factors, and glycemic condition and performance.

Sub-states can represent short- and long-term factors. Examples of short-term sub-states are [8]: a) the current blood glucose level (mg/dL) (e.g., measured using a glucometer or a subcutaneous glucose sensor); b) the carbohydrate contents of a meal (e.g., estimated using a mobile application); c) the bolus insulin dosage (e.g., administered using an insulin pump); and d) the glucagon dosage (e.g., also administered using an insulin pump) if it is used. Examples of long-term sub-states are [7]: a) the basal insulin dosage; and b) the current and past readings that describe trends, such as the mean insulin delivery and the postprandial error in the blood glucose level (e.g., the percentage of time duration in hypoglycemia and hyperglycemia) in the previous decision epoch.

Sub-states can capture the glycemic performance. Examples of sub-states are [35], [65]: a) the penalty $s_{1,t} = \max\{g_{\max,t} - G_{hyper}, 0\}$ that shows the extent of the highest measured blood glucose level up to decision epoch t, namely $g_{\max,t}$, is higher than the hyperglycemia threshold G_{hyper}; b) the penalty $s_{2,t} = \max\{G_{hypo} - g_{\min,t}, 0\}$ that shows the extent of the lowest measured blood glucose level up to decision epoch t, namely $g_{\min,t}$, is lower than the hypoglycemia threshold G_{hypo}; and c) the discretized percentages of time duration outside the normal range of the blood glucose level in hyperglycemia and hypoglycemia. Representing penalties in sub-states helps to capture the glycemic performance so that appropriate actions can be selected to minimize them.

2) APPROACHES FOR STATE REPRESENTATION

Various approaches have been used to improve the representations of states and sub-states, including determining the right number of states, simplifying sub-states with binary and numerical values, and collecting multiple observations and measurements in a single decision epoch.

Determining the right number of states helps to ensure there are a sufficient number of states for making the right decisions while improving the convergence rate. The Mclust package [87] determines the right number of states (e.g., 50 in [51]), which has the highest Bayesian information criterion value indicating the best possible clustering effect. Then, the Gaussian mixture model segregates the state space into states.

Simplifying sub-states with binary and numerical values reduces the state space. Examples of simplified sub-states, which represent the patient's clinical condition [50], are: a) $s_{chronic} = \{0, 1\}$ represents whether the patient has chronic complications (e.g., diabetic retinopathy, diabetic cataract, and foot disease) caused by diabetes; b) $s_{acute} = \{0, 1\}$ represents whether the patient has acute complications (e.g., myocardial infarction and heart failure) caused by diabetes; c) $s_{risk} = \{0, 1\}$ represents whether the patient has a higher risk of developing diabetes, which depends on multiple factors, such as age, family history, and BMI; d) $s_{period} = \{0, 1, 2\}$ represents whether the patient has up to four years of diabetes, between five and eight years, or more than eight years; and e) $s_{plasma} = \{0, 1, 2\}$ represents whether the patient has a

TABLE 4. Examples of RL models for diabetes management.

Role	Reference	Objective	State	Action	Reward	Iteration
Estimating the insulin dosage (R.1)	[14], [64]	To maintain the blood glucose level within the normal range.	Patient's clinical condition (i.e., the blood glucose level), consumed carbohydrate contents, and IOB.	Insulin dosage (e.g., the basal rate).	Reward increases with a higher TIR, a reduced difference between the current and target blood glucose level given by physicians, and lesser occurrences of hyperglycemia and hypoglycemia.	Every 120 minutes
Estimating the insulin dosage and time (R.1)	[86]	To maintain the blood glucose level within the normal range.	Patient's clinical condition (i.e., the blood glucose level and its time index) and meal ingestion (i.e., the carbohydrate contents and its time index)	Insulin dosage and infusion time.	Reward increases with the average score, which is higher when the blood glucose level is within the normal range.	Each meal time
Recommending the right type of treatment (R.2)	[13]	To improve the effectiveness of treatment.	Patient's clinical condition (i.e., demographics, physical examination, medical history, lifestyle, previous prescriptions, and lab test results)	Prescription pattern (i.e., the number of oral prescriptions and insulin dosage), subject to up to 4 oral antidiabetic drugs and 2 insulin injections.	Reward increases with a lower HbA1c level and a lesser occurrences of hypoglycemia in between visits.	Each follow-up visit

fasting plasma glucose that indicates normal, prediabetes, or diabetes.

Collecting multiple observations and measurements in a decision epoch improves the accuracy of state information. Observations can be collected at time instants $i = 1, \ldots, n \in t$ in a single decision epoch t, where n is the number of observations in the decision epoch t [10]. Specifically, the agent collects the i^{th} observations, namely $g_{i,t}$ (e.g., the blood glucose level), in decision epoch t. After collecting n observations, the agent selects the optimal action at the end of the decision epoch t. Given a weight $0 \leq w \leq 1$, the weighted blood glucose level $g_t = \sum_{i=1}^{i=n} g_{i,t} w^{n-i} / \sum_{i=1}^{i=n} w^{n-i}$ provides more significance to more recent observations.

B. ACTION REPRESENTATION APPROACHES

Each potential action has either a single or a set of multiple types of actions. For instance, an action set consists of the insulin dosage and other medications, which are essential to treat comorbid patients [10]. The action is generally discretized to reduce the action space.

The action can be either an actual or relative value. Examples of actual values are [8], [35]: a) the insulin dosage, particularly the basal rate is $a_t \in A = (\text{Suspension}, 0.5 \times I_{basal}, I_{basal}, 1.5 \times I_{basal}, 2 \times I_{basal})$, where I_{basal} is a unit of the basal rate; and b) the glucagon dosage is $a_t = 0.3\mu$g/kg subject to the rule of 1mg per day to ensure safety. An example of relative values is the changes in the insulin dosage $a_t \in A = (+1, 0, -1)$ representing the action to increase, maintain, or reduce the insulin dosage compared to the current action, subject to the rule of a $\pm 15\%$ change to ensure safety [7].

Rules are safety constraints that remove inappropriate actions from the set of potential actions so that only appropriate actions are selected for both exploration and exploitation. Rules are generally set based on physician's knowledge. As an example of a physician's knowledge, the potential actions (e.g., the insulin dosage) for children must be more constrained than those of adults [64]. Rules are also generally set based on the effects of the actions (or outcomes) which must adhere to strict medical policies in order to ensure patient safety. Appropriate actions ensure changes, including state, action, and reward, are acceptable and limited to a certain percentage. In [7], potential actions are selected so that the changes in the insulin-to-carbohydrate ratio and the basal rate are limited to $\pm 15\%$ and $\pm 20\%$, respectively, compared to the day before.

C. REWARD REPRESENTATION APPROACHES

In general, examples of rewards are short-term, long-term, comparative, tunable, and QLAY.

1) SHORT-TERM (INSTANTANEOUS) REWARDS

Instantaneous rewards provide short-term values that reflect the performance achieved by RL. The instantaneous rewards can be either a constant value or a difference between the current and previous states.

In [8], the instantaneous reward values are constants, and they are determined based on the blood glucose levels. The reward is $r_t(s_t) = 1$ when the blood glucose level is within the normal range. The reward is $r_t(s_t) = 0.1$ when the blood glucose level is at the borderline (i.e., between 70mg/dL and 90mg/dL, and between 140mg/dL and 180mg/dL). As the blood glucose level deviates from the normal range, the reward value reduces when the blood glucose level reduces from 70mg/dL to 30mg/dL and increases from 180mg/dL to 300mg/dL. The reward value is $r_t(s_t) = -1$ when the blood glucose level is lower than 30mg/dL and higher than 300mg/dL.

In [51], the instantaneous reward values are differences between the current and previous states. There are three main

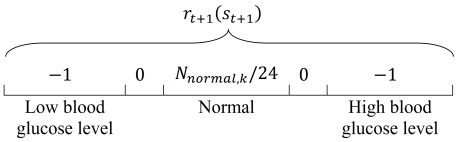

FIGURE 4. The reward value is based on the blood glucose level. For instance, the reward value is $r_{t+1}(s_{t+1}) = -1$ when the blood glucose level is high and low. The reward value for the buffer zones is zero.

states based on blood glucose levels, namely low, medium, and high. The reward value is $r_t(s_t) = 2$ when the blood glucose level changes from high to low, $r_t(s_t) = 1$ when changes from high to medium, $r_t(s_t) = -4$ when remains at high or low. In [7], a positive reward is given when the insulin dosage is lower than that of the previous day without causing an increase in the percentage of time duration in which the blood glucose level is outside the normal range.

2) LONG-TERM (AVERAGE) REWARDS

Average rewards provide long-term values that reflect the performance achieved by RL. In [14], the average reward values are determined based on the blood glucose levels. As shown in Figure 4, the reward value is $r_{t+1}(s_{t+1}) = N_{normal,k}/24$, where $N_{normal,k}$ represents the number of decision epochs in which the blood glucose level is within the normal range in the past 24 hours [14]. For the negative reward value $r_{t+1}(s_{t+1}) = -1$, the agent avoids such blood glucose levels which are fatal. In [35] and [88], the average cost (or negative reward) values are determined by the occurrences of hyperglycemia $N_{hyper,k}$ and hypoglycemia $N_{hypo,k}$ in day k. The cost value is $r_{t+1}(s_{t+1}) = (w_{hyper} \times N_{hyper,k}) + (w_{hypo} \times N_{hypo,k})$, where weight $w_{hypo} = 0.1$ is greater than $w_{hyper} = 0.01$ because of the need to prioritize the avoidance of hypoglycemia due to its more life-threatening consequences.

3) COMPARATIVE REWARDS

Comparative rewards provide the difference between the blood glucose level and its target. According to the pancreatic cellular model [89], the insulin infusion rate $I = I_{basal} + I_{above-basal}$ is based on the basal insulin secretion rate I_{basal} and the above-basal insulin secretion rate $I_{above-basal} = f(g_t, \Delta g_t)$ controlled by the current glycemic conditions, where g_t represents the blood glucose level at time t and Δg_t represents the rate of change in the blood glucose level (also known as glucose variability) at time t. Following the pancreatic cellular model, the reward representation has two components [6]. First, the long-term reward $r_{t+1,long}(s_{t+1}) \propto -|g_t - G_{target}|$ represents the difference between the current blood glucose level g_t at time t and its target value, such as $G_{target} = 120$mg/dL [90]. This helps to regulate the long-term slow-changing basal insulin secretion rate I_{basal}. Second, the short-term reward $r_{t+1,short}(s_{t+1}) \propto |M_{target} \times (g_t - G_{target}) - \Delta g_t|$ represents the difference between glucose variability Δg_t at time t and $(g_t - G_{target})$, where M_{target} is the slope constant of the target glucose variability. This helps to regulate the short-term fast-changing above-basal insulin secretion rate $I_{above-basal}$. Both long- and short-term rewards are weighted using different constant values based on different levels of clinical risks, such as -1 for benign, -3 for potentially dangerous, and -5 for hazardous. For instance, the long-term reward is $r_{t+1,long}(s_{t+1}) \propto -3 \times |g_t - G_{target}|$ when the blood glucose level is outside the normal range, and $r_{t+1,long}(s_{t+1}) \propto -|g_t - G_{target}|$ when the blood glucose level is within the normal range.

TABLE 5. A summary of RL approaches applied in diabetes management.

RL approach	Description	Strength
Model-based RL	The agent interacts with a model of the operating environment and learns the best possible policy.	Increases the convergence rate.
Multi-agent RL	The distributed agent uses the local value function, which is a decomposition of the global value function.	Reduces system complexity.
Actor-critic RL	The agent has two components, namely critic and actor. The critic learns the value function (i.e., the values of state-action pairs) and the actor learns the optimal policy (i.e., the best possible state-action pairs).	Increases the convergence rate.
DQN	The agent has two neural networks, namely the main and target networks, and a replay memory. The main network learns the optimal policy. The target network approximates the target value to calculate the loss function during training. The replay memory stores and retrieves experiences for training.	Solves complex problems and addresses the curse of dimensionality.
Gaussian process approximation (GPA)	The agent provides the Gaussian probability distributions of states under uncertainties and noise.	Learns under continuous and high-dimensional state and action spaces.
Proximal policy optimization (PPO)	The agent learns separate components, such as policy, value function, and advantage, which are the decompositions of the Q-function.	Learns under continuous state and action spaces.

4) TUNABLE REWARDS

Reward functions and values can be tuned and adjusted by physicians based on personal clinical conditions, such as hypoglycemia and hyperglycemia thresholds, and insulin sensitivity, for achieving optimal outcomes [10]. The clinical conditions vary between patients. For instance, young patients tend to experience hyperglycemia faster than adults. The agent receives a positive (negative) reward for achieving a healthy (an unhealthy) blood glucose level. Negative rewards can be adjusted to penalize the occurrences of hyperglycemia and hypoglycemia.

5) THE QLAY REWARDS

The quality-adjusted life-year (QLAY) rewards represent the expected time in a healthy condition [50]. The reward is $r_{t+1}(s_{t+1}) = R^{WTP} \times \Pi_i(1 - d^i(s_{t+1})) - C^{MED}$, where

R^{WTP} represents the willingness to pay for a single perfectly healthy year without medical intervention and side effects of treatments, C^{MED} represents the medication cost, and $d^i(s_{t+1})$ represents the decrement factors s_{t+1}, including age, blood pressure, BMI, chronic diseases, acute diseases, risk, period, and fasting plasma glucose.

V. RL ALGORITHMS FOR DIABETES MANAGEMENT

Various RL approaches and enhancements have been proposed for diabetes management. Table 5 summarizes the RL approaches, which are described in the rest of this section. Table 6 summarizes the attributes of RL in diabetes management.

A. TRADITIONAL RL

The traditional RL approach (see Section II-B) has been applied in [43] and [62].

1) ZOHORA's TRADITIONAL RL APPROACH

In [43], the objective is to detect T2DM diabetes in its early stage (R.3). The state has three sub-states: a) the blood glucose level (from Level 1 at <115mg/dL to Level 5 at >150mg/dL); b) BMI (from Level 1 at <25.6 to Level 11 at >45); and c) age (from Level 1 at <24 to Level 6 at >45). The action represents a change in the state. The reward is $r_t(s_t) = +10$ for a positive patient's clinical condition and $r_t(s_t) = -100$ in other conditions. Patients with Q-values of more than +10 are categorized as diabetes, and those in between 0 and <+10 are categorized as prediabetes. Experiments investigation is based on the PIMA Indian women diabetic dataset (D1.2). The proposed solution has been shown to increase the accuracy of prediction (P.6) of T2DM patients with prediabetes and diabetes. Specifically, it achieves high values of true positive and true negative, a low value of false negative, and zero false positive in the confusion matrix (see Figure 3).

2) NOORI's TRADITIONAL RL APPROACH

In [62], the objective is to estimate the insulin dosage (R.1). The state represents the blood glucose and insulin levels. The action represents the insulin dosage. The reward represents the absolute difference between the current and target glucose level. Instead of using Equation (3), the SARSA algorithm, which updates the Q-function using the real past experience (or transition) $(s_t, a_t, r_{t+1}(s_{t+1}, a_{t+1}), s_{t+1}, a_{t+1})$ rather than an estimation of the long-term reward, is as follows:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[r_{t+1}(s_{t+1}) + \gamma Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)] \quad (6)$$

The SARSA algorithm has been shown to achieve a safer solution that minimizes negative reward [30]. Using the Palumbo's system model (S.5), experiments investigate the effects of insulin sensitivity (E.1). The proposed solution has been shown to increase TIR (P.3).

3) JAFAR's TRADITIONAL RL APPROACH

In [7], the objective is to estimate the insulin-to-carbohydrate ratio in order to determine the insulin dosage (R.1). The state has three sub-states: a) the delivered insulin dosage; b) the percentage of time duration in hyperglycemia; and c) the percentage of time duration in hypoglycemia. The action represents a change in the insulin dosage (i.e., increase, maintain, or reduce). The reward is a positive value when the insulin dosage is lower than that of the previous day without causing an increase in the percentage of time duration in which the blood glucose level is outside the normal range. Using the Hovorka's glucoregulatory model (S.6), experiments investigate the effects of meal sizes and times (E.2). Based on simulation and clinical studies, the proposed solution has been shown to increase TIR (P.3) and the proportion of patients achieving the normal blood glucose level (P.4).

B. MODEL-BASED RL

Model-based RL interacts with a model of the operating environment, rather than a real one, and learns the best possible course of actions. This reduces the need for a real state and delayed reward observed from the operating environment during training and action selection. Nevertheless, experiences comprised of state, action, and reward obtained through interactions with the real operating environment are still important for developing the model. Compared to the traditional model-free RL approach, the advantage of the model-based RL approach is that it can learn even without taking an action in the real operating environment, while traditional RL can learn only when taking an action in the real operating environment. Due to the greater opportunities to learn, model-based RL has been shown to increase efficiency and reduce the convergence time, which is more suitable for applications generating limited experiences and requiring real-time responses. Nevertheless, one shortcoming of the model-based RL approach is that the optimality of the learned policy is limited by the accuracy of the model. Other shortcomings are presented in (V.1).

1) LI's AND SHIFRIN's MODEL-BASED RL APPROACH

The objectives are to estimate the insulin dosage (R.1) in [51] and to recommend the right type of treatment (R.2) in [10]. The state represents the blood glucose level and carbohydrate contents in [10], in addition to other static and dynamic variables, such as age, gender, insulin level, and temperature, in [51]. The action represents the insulin dosage in [51] and [10], and additionally oral prescription in [10]. The reward value increases with increasing: a) the difference between the current blood glucose level g_t at time t and its target value $|g_t - G_{target}|$; and b) the average difference between the blood glucose level and the hypoglycemia threshold $(G_{hypo} - g_t)$ during hypoglycemia, where $g_t < G_{hypo}$. The model is the transition probability matrix $P(s_{t+1}|s_t, a_t)$, which represents

TABLE 6. Summary of the attributes of RL in diabetes management.

RL approach	Scheme	Role			Patient data				
		(R.1) Estimating the blood glucose level or the insulin dosage	(R.2) Recommending the right type of treatment	(R.3) Diagnosing diabetes	(D1.1) The NIDDK dataset	(D1.2) The PIMA Indian women diabetic dataset	(D1.3) The MIMIC-III dataset	(D1.4) The KNHISS dataset	(D1.5) The NEW2D dataset
Traditional	Zohora [43]			×		×			
	Noori [62]	×							
	Jafar [7]	×							
Model-based	Li [51]	×					×		
	Shifrin [10]		×		×				
Multi-agent	Li [51]	×					×		
Actor-critic	Sun [21]	×							
DQN	Liu [13]		×						×
	Zhu [8]	×							
GPA	Paula [57], [71]	×							
PPO	Lee [6]	×							

RL approach	Scheme	System model						Simulator		Simulation parameter		
		(S.1) Ferdinando's system model	(S.2) Lehmann-Deutsch physiological model	(S.3) Shifrin's system model	(S.4) Lee's system model	(S.5) Palumbo's system model	(S.6) Hovorka's glucoregulatory model	(M.1) The UVA/ Padova simulator	(M.2) The AIDA simulator	(L.1) Exploration probability	(L.2) Learning rate	(L.3) Discount factor
Traditional	Zohora [43]											
	Noori [62]					×						
	Jafar [7]						×					
Model-based	Li [51]									×		
	Shifrin [10]			×					×			
Multi-agent	Li [51]									×		
Actor-critic	Sun [21]							×				
DQN	Liu [13]											
	Zhu [8]							×				
GPA	Paula [57], [71]		×						×			
PPO	Lee [6]				×			×				

RL approach	Scheme	Experiment design (the effects of)				
		(E.1) Insulin sensitivity	(E.2) The number of meals and its variations	(E.3) The dawn phenomeno	(E.4) Delayed insulin infusion	(E.5) Manual exogenous injection and automated subcutaneous infusions
Traditional	Zohora [43]					
	Noori [62]	×				
	Jafar [7]		×			
Model-based	Li [51]					
	Shifrin [10]	×	×			
Multi-agent	Li [51]					
Actor-critic	Sun [21]	×	×	×		×
DQN	Liu [13]					
	Zhu [8]	×	×			×
GPA	Paula [57], [71]		×		×	
PPO	Lee [6]	×	×			

RL approach	Scheme	Performance measures										
		(P.1) Higher similarity between RL and physician's policies	(P.2) Lower HbA1c level	(P.3) Higher time in range (TIR)	(P.4) Higher proportion of patients with normal blood glucose level	(P.5) Lesser occurrences of hyperglycemia and hypoglycemia	(P.6) Higher accuracy of prediction	(P.7) Lower glucose variability	(P.8) Lower insulin dosage	(P.9) Longer delayed time period of chronic and acute complications	(P.10) Higher convergence rate	(P.11) Higher cumulative reward
Traditional	Zohora [43]						×					
	Noori [62]			×								
	Jafar [7]			×	×							
Model-based	Li [51]				×				×			
	Shifrin [10]			×								×
Multi-agent	Li [51]				×				×			
Actor-critic	Sun [21]			×		×		×				
DQN	Liu [13]	×	×			×						
	Zhu [8]			×								
GPA	Paula [57], [71]			×								
PPO	Lee [6]			×								

the dynamics of the operating environment [10], [51]. The transition probability matrix can be estimated based on real historical datasets and new running data, such as the physiological characteristics (e.g., the blood glucose and insulin levels) of a large number of patients in electronic medical records. Due to the dynamicity of the individual's glucose variability, the new running data is added to the historical dataset, replacing old data with new and novel data, as time goes by. The transition probability $P(s_{t+1}|s_t, a_t) = N_t(s_t, a_t, s_{t+1})/\sum_{s_{t+1}} N_t(s_t, a_t, s_{t+1})$, where $N_t(s_t, a_t, s_{t+1})$ is the number of visits to the (s_t, a_t, s_{t+1}) set. In [51], the experiments investigate the effects of different values of exploration probability (L.1) based on the MIMIC-III dataset (D1.3). The proposed solution has been shown to increase the proportion of patients achieving the normal blood glucose level (P.4) with lower insulin dosage (P.8). In [10], using the Shifrin's system model (S.3) and the AIDA simulator (M.2), experiments investigate the effects of insulin sensitivity (E.1) and meal sizes under different patients' weights (E.2) based on the NIDDK dataset (D1.1). The proposed solution has been shown to increase TIR (P.3) and cumulative reward (P.11).

C. MULTI-AGENT RL

Multi-agent RL decomposes the global value function $V_{\pi_t}(s_t)$ (see Equation (1)), which represents the goodness of being in a particular global state s_t following the global policy π_t, into separate value function $V^i_{\pi_t}(s^i_t)$ for distributed agent $i \in I$, where I is a set of agents [51].

1) LI's MULTI-AGENT RL APPROACH

In the multi-agent approach of [51], each distributed agent i observes its local state s^i_t and optimizes the local reward, comprised of delayed reward $r^i_{t+1}(s^i_{t+1})$ and the rewards of neighboring agents J, contributing to an optimal global Q-value. The global Q-function is the summation of the local Q-functions of agents I, specifically $Q_t(s_t, a_t) = \sum_{i=1}^{I} \alpha^i Q^i_t(s^i_t, a^i_t)$, where $\sum_i \alpha^i = 1$ is the linear weight factor [51]. The global action is $a_t = (a^1, a^2, \ldots, a^{|I|})$, and the global reward is $r_{t+1}(s_{t+1}) = \sum_i \alpha^i r^i_{t+1}(s^i_{t+1})$. More details about the Li's approach are presented in Section V-B1.

D. ACTOR-CRITIC RL

Actor-critic RL has two interactive and complementary components, namely critic and actor, which can be represented using neural networks to solve complex tasks as shown in Figure 5. The actor-critic RL has been shown to increase the convergence rate [35].

Each agent may have multiple actor-critic networks. In [35], there are: a) a basal actor-critic network that estimates the basal insulin dosage (i.e., the average basal rate); and b) the bolus actor-critic network that estimates the insulin-to-carbohydrate ratio, which represents the carbohydrate contents that one unit of bolus insulin covers in order to estimate the bolus insulin dosage required for an announced meal and its size.

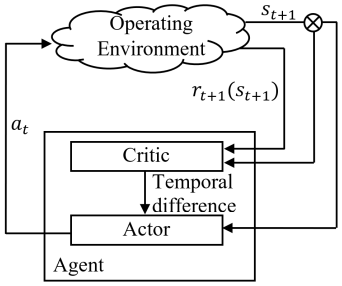

FIGURE 5. The actor-critic algorithm.

1) SUN's ACTOR-CRITIC RL APPROACH

In [21], the objective is to estimate the basal rate and three insulin-to-carbohydrate ratios (R.1). The state has two sub-states: a) $s_{1,t} = (1/N_{hypo,t})\sum(G_{hypo} - g_t)$ represents the average difference between the blood glucose level and the hypoglycemia threshold during hypoglycemia, where $g_t < G_{hypo}$ and $N_{hypo,t}$ is the number of occurrences of hypoglycemia up to decision epoch t. Similarly, $s_{2,t} = (1/N_{hyper,t})\sum(g_t - G_{hyper})$ represents the average difference between the blood glucose level and the hyperglycemia threshold during hyperglycemia, where $g_t > G_{hyper}$ and $N_{hyper,t}$ is the number of occurrences of hyperglycemia up to decision epoch t. The action represents the insulin dosage (i.e., the basal insulin rate) and the insulin-to-carbohydrate ratio. The cost (or negative reward) is $r_t(s_t) = (w_{hyper} \times s_{1,t}) + (w_{hypo} \times s_{2,t})$, where w_{hyper} and w_{hypo} weight the hyperglycemia and hypoglycemia features.

The critic learns the value function $V_{\pi_t}(s_t)$ of the current state s_t (see Equation (1)), which represents the cumulative reward including the delayed and discounted rewards for being in a state s_t. In other words, the critic evaluates the appropriateness of the current policy π_t for the state s_t. Then, the critic calculates and provides the temporal difference $r_{t+1}(s_{t+1}) + \gamma V_{\pi_t}(s_{t+1}) - V_{\pi_t}(s_t)$ to the actor. The actor updates its policy π_t using the policy gradient approach [64] $\pi_{t+1} = \pi_t - \alpha_t \nabla_\pi \bar{r}_t(\pi_t)$ to approximate the optimal policy as time goes by, where α_t represents the learning rate and $\nabla_\pi \bar{r}_t(\pi_t)$ represents the gradient of the average reward $\bar{r}_t(\pi_t)$, which is based on the temporal difference.

Using the UVA/ Padova simulator (M.1), experiments investigate the effects of insulin sensitivity (E.1), meal sizes and times (E.2), the dawn phenomenon (E.3), and the manual exogenous injection and automated subcutaneous infusion approaches (E.5). The proposed solution, which shows that both self-monitoring and continuous blood glucose monitoring approaches achieve comparable performance, has been shown to increase TIR (P.3) and reduce the risks and occurrences of hyperglycemia and hypoglycemia (P.5). In addition, the continuous blood glucose monitoring approach reduces glucose variability (P.7), contributing to a more stable blood glucose level.

E. DEEP Q-NETWORK

Compared to the traditional RL approach, deep Q-network (DQN) is a DRL approach comprised of neural networks for handling complex conditions, including the glucose-insulin dynamics. DQN has been shown to achieve outstanding achievements in a diverse range of applications, such as Atari games [91] and AlphaGo [92].

Figure 6 shows the DQN architecture. There are two identical neural networks which are fully connected in which a neuron connects to all neurons in the next layer through links, and each link has a weight. The main network, which is characterized by network parameters (or weight matrices) θ_t, approximates Q-values $Q_t(s_t, a_t; \theta_t) \approx Q_t^*(s_t, a_t)$ during action selection and training. The target network, which is a duplicate of the main network and is characterized by network parameters θ_t', approximates the target Q-value $Q_t(s_t, a_t; \theta_t^-)$ and calculate the target $y_t = r_{t+1}(s_{t+1}) + \gamma \max_a Q_t(s_{t+1}, a; \theta_t^-)$. Subsequently, the target y_t is used to calculate the loss function $L(\theta_t) = E_{s,a}[(y_t - Q_t(s, a; \theta))^2]$ for weight update during training. Each neural network is comprised of three types of layers. First, the *input layer* represents the state with each neuron representing a sub-state. So, the number of neurons in the input layer is based on the number of sub-states. In [88], the input layer has nine neurons receiving the blood glucose levels of the past eight time steps and the current insulin dosage. The sub-state values in the input layer are normalized since they have different ranges and units, such as the blood glucose level $10 \leq G \leq 400$mg/dl and the current insulin dosage $0 \leq U \leq 600$pmol/min. Second, multiple *hidden layers* with a number of neurons in each layer (e.g., 2 layers with 128 neurons in each layer [13])), and each neuron uses an activation function, such as hyperbolic tangent, log-sigmoid, and leaky ReLU [13], [88]. Third, the *output layer* represents the actions with each neuron representing the Q-value of the respective action under the input state. So, the number of neurons in the output layer is based on the number of potential actions. In [88], the output layer has ten neurons representing the estimations of the blood glucose level in the next ten time steps. The Q-values of the output layer are denormalized.

As shown in Figure 6, there is a replay memory for experience replay. During training, through back-propagation, the main network optimizes the network parameters θ_t by minimizing the loss function $L(\theta_t)$. The network parameters θ_t are updated using stochastic gradient descent $\theta \leftarrow \theta - \alpha \nabla_{\theta_t} L(\theta_t)$. During normal operation, through feed-forwarding, the input layer receives a state, and the output layer generates the Q-values of all potential actions.

Compared to the traditional RL approach, the advantage of DQN is that it addresses the curse of dimensionality, or the large state space, in traditional RL. Nevertheless, the shortcomings of DQN are that it requires a high computing capability and has a longer convergence time, which is based on the size of the state space [86].

1) LIU's DRL APPROACH

In [13], the objective is to recommend the right type of treatment during each follow-up visit (R.2). The state represents a wide range of decision factors, including demographic, physical examination, medical history, lifestyle, previous prescriptions, and lab test results, as shown in Table 3. The action represents the number of oral antidiabetic drugs and insulin dosage. The reward value increases when the HbA1c level reduces or is less than 7%, and a penalty is imposed when hypoglycemia occurs. Generally speaking, the number of oral antidiabetic drugs and insulin dosage increase when the mean and median HbA1c level increase. The DQN architecture has dilated recurrent neural networks [93] that consider long-term dependencies in the prediction of future blood glucose levels. Clinical studies are based on the NEW2D dataset (D1.5). The proposed solution has been shown to increase the similarity between RL and physician's policies (P.1), and reduce the HbA1c level (P.2) and the occurrences of hyperglycemia and hypoglycemia (P.5).

2) ZHU's DRL APPROACH WITH PRIORITIZED EXPERIENCE REPLAY

In [8], the objective is to estimate the insulin and glucagon dosages (R.1). The state represents the blood glucose level, estimated carbohydrate contents, insulin dosage (i.e., basal and bolus insulin dosages), and glucagon dosage. The action represents the basal insulin dosage (i.e., basal insulin rate). The reward value is a constant value for different ranges of blood glucose levels, which is higher when the blood glucose level is within the normal range and lower otherwise. The DQN approach uses a two-step framework and the prioritized experience replay approach.

In the two-step framework, there are two main steps. First, the long-term generalized training is performed to train a DQN model comprised of dilated recurrent neural networks without safety constraints [94] using the multi-dimensional time-series data. The outcome is a DQN model of meal ingestion, and the blood glucose and hormone levels, for an average individual. The DQN model provides initialization to improve the initial performance. Second, personalized learning is performed through transfer learning to fine-tune the DQN model using a short-term dataset with safety constraints so that the DQN model is tailored for an individual with personal characteristics. During the personalized training, some of the earlier layers of the DQN model are retained, and the rest of the layers are updated to avoid over-fitting. This is because the earlier layers contain general features useful for the general patient population, such as suspending insulin infusion during hypoglycemia.

In the prioritized experience replay approach, each experience has a selection probability $P_i = p_i^\phi / \sum_i p_i^\phi$ with $p_i = |\delta_i| + c$, where p_i represents the priority of the experience i, δ_i represents the temporal difference of experience i, c is a small positive constant, and $0 \leq \phi \leq 1$ represents

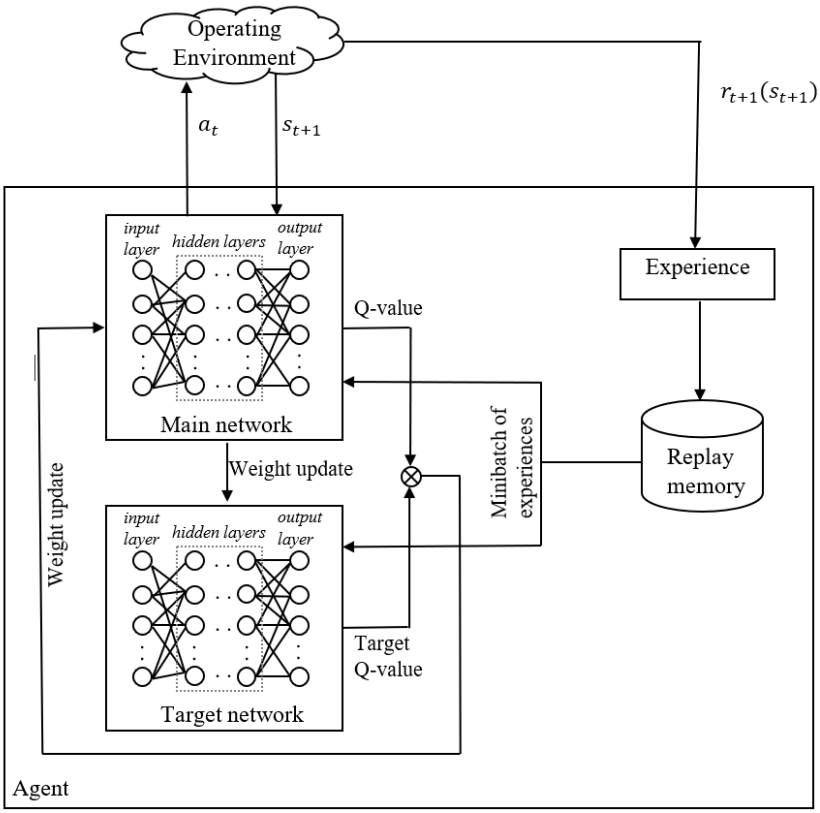

FIGURE 6. DQN architecture has a main network, a target network, and a replay memory.

the level of using prioritization. So, experiences with higher temporal differences δ_i, which indicates their importance, are selected from the replay memory to provide a minibatch of experiences for training. The prioritized experience replay approach has been shown to increase the convergence rate even when the number of experiences in the replay memory is limited.

Using the UVA/ Padova simulator (M.1), experiments investigate the effects of insulin sensitivity (E.1), meal sizes and times (E.2), the manual exogenous injection and automated subcutaneous infusion approaches (E.5). The proposed solution has been shown to increase TIR (P.3).

F. THE GAUSSIAN PROCESS APPROXIMATION

The Gaussian process approximation (GPA) is a data-driven, non-parametric, and probabilistic approach that performs function approximation and regression to provide the Gaussian probability distributions of states under uncertainties and noise, particularly the glucose-insulin dynamics affected by the inaccuracy of the blood glucose level measurements and the pharmacological delay of insulin that can vary significantly between patients and days. This helps to learn the optimal policy under continuous and high-dimensional state and action spaces.

1) PAULA's GPA APPROACH

In [57] and [71], the objective is to estimate the insulin dosage (R.1). The state represents the current blood glucose level and the insulin dosage in the previous time step. The action represents the insulin dosage. The Gaussian reward function is based on the blood glucose level with the width of the glucose band being the normal blood glucose level. There are three separate Gaussian process models, which are characterized by their respective mean and covariance functions, that provide three different estimations: a) the next state based on different current state-action pairs; b) the value functions of different states $V_{\pi_t}(s_t)$; and c) the Q-functions of different state-action pairs $Q_t(s_t, a_t)$. There are two main steps. First, using a general dataset, the agent learns a general optimal policy π_t^* for an average individual based on the Ito's stochastic model, which models the glucose variability and glucose-insulin dynamics, in an offline manner [71]. The Bayesian active learning approach selects and explores the relevant part of the state space out of the potential states. Second, using real experiences of a patient and starting with the optimal general policy π_t^*, the agent learns a personalized policy π_t^c. The agent updates the three separate Gaussian process models to capture individual dynamics, including the glucose-insulin dynamics and the insulin sensitivity of an individual. Subsequently, the agent updates Q-values based on the estimated Q-functions and value functions. Using

the Lehmann-Deutsch physiological model (S.2) and the AIDA simulator (M.2), experiments investigate the effects of meal sizes and times (E.2) and the pharmacological delay of insulin (E.4). The proposed solution has been shown to increase TIR (P.3).

G. PROXIMAL POLICY OPTIMIZATION

Proximal policy optimization (PPO) decomposes the Q-function into separate components, such as policy, value function, and advantage, to solve problems with the continuous state and action spaces [95], such as the blood glucose and insulin levels, which are continuous in nature. PPO has single or multiple neural networks, and each neural network provides specific estimations. For instance, in [14], a single neural network provides value function $V_{\pi_t}(s_t)$ and advantage $A_{\pi_t}(s_t, a_t)$ used to calculate the Q-value of a state-action pair $Q_t(s_t, a_t) = A_{\pi_t}(s_t, a_t) + V_{\pi_t}(s_t)$, where the advantage $A_{\pi_t}(s_t, a_t)$ is a quadratic function of the nonlinear components of a state.

1) LEE's PROXIMAL POLICY OPTIMIZATION

In [6], the objective is to estimate the insulin dosage (R.1). The state represents the current glycemic conditions, namely the blood glucose level and the glucose variability with measurement errors and noise, and IOB. The state is normalized by the running mean and standard deviation through the normalization layer as shown in Figure 7. The action, which is selected using the policy network, represents the insulin dosage. The reward is based on the difference between the current blood glucose level and its target value, and glucose variability (see Section IV-C3). As shown in Figure 7, the PPO approach has three multiple neural networks: a) the policy network maps states s_t to actions a_t; b) the short-term value function network estimates the short-term return $r_{t+1,short}(s_{t+1})$ and its average $\bar{r}_{t+1,short}(s_{t+1})$; and c) the long-term value function network estimates the long-term return $r_{t+1,long}(s_{t+1})$ and its average $\bar{r}_{t+1,long}(s_{t+1})$. PPO selects actions to minimize advantage $A_t = (r_{t+1,short}(s_{t+1}) - \bar{r}_{t+1,short}(s_{t+1})) + \theta(r_{t+1,long}(s_{t+1}) - \bar{r}_{t+1,long}(s_{t+1}))$, where θ is a scale parameter that indicates the significance of the long-term return and its average. The policy π_t is updated after every single simulated meal. Using the Lee's system model (S.4) and UVA/ Padova simulator (M.1), experiments investigate the effects of insulin sensitivity (E.1) and the number of meals (E.2). The proposed solution has been shown to increase TIR (P.3).

H. OTHER ENHANCEMENTS

This section presents other enhancements which can be incorporated into most RL approaches.

1) LIMITING THE POTENTIAL ACTION SET FOR DIFFERENT STATES

Limiting the potential action set for different states addresses the curse of dimensionality, or the large state space, in traditional RL. In [50], the objective of the contextual bandit model [50] is to recommend the right type of treatment over a period of one year (R.2). This helps to maintain the blood glucose level within the normal range at the early stage and prevent complications [96]. The state represents eight factors, including blood glucose level, age, and BMI. The action represents prescriptions, which can be monotherapy (e.g., metformin) at the initial stage, and dual therapy or triple therapy at the later stage. The mean reward for each state-action pair is $Q_t(s_t, a_t) = \sum_{i=0}^{t} r_i(s_i, a_i)/t$. Consequently, there are a large number of states (i.e., 1,296) and actions (i.e., 15). The three-armed bandit model selects three best potential actions for each state. Based on the KNHISS dataset (D1.4), the proposed solution has been shown to increase the convergence rate (P.10), the similarity between RL and physician's policies (P.1), and the average time period from diabetes diagnosis to the occurrence of chronic or acute complications caused by diabetes (P.9).

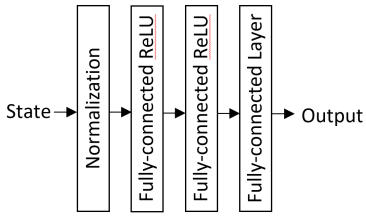

FIGURE 7. An abstract model for the architecture of the three similar networks in PPO [6].

2) MODELING THE INFUSED INSULIN ACTIVITIES USING DISCOUNT FACTOR

In general, the pharmacokinetics and pharmacodynamics properties of individual insulin analogs mimic the natural pattern of insulin release in the body, whereby the infused insulin is activated rapidly, peaked, and then reduced gradually over time following the insulin time-action profile [59]. The discount factor γ has been applied to estimate the individual insulin analogs, particularly the remaining effect of the infused insulin.

In [6], the insulin activity is represented by short- and long-term returns. The short-term return is given by $\sum_{i=0}^{t-1} \gamma^i f_i r_{t+1,short}(s_{t+1}) + \gamma^n f_n \bar{r}_{t+1,short}(s_{t+n})$ and the long-term return is given by $\sum_{i=0}^{t-1} \gamma^i F_i r_{t+1,long}(s_{t+1}) + \gamma^n F_n \bar{r}_{t+1,long}(s_{t+n})$, where the insulin becomes inactive after n steps. The short-term discount factor $0 \leq f_n \leq 1$ and long-term discount factor $0 \leq F_n \leq 1$ are based on the insulin time-action profile, represented by the probability density function and the cumulative density function of a gamma distribution, respectively. The proposed solution has been shown to increase TIR (P.3).

3) DEFINING MULTIPLE REWARD FUNCTIONS FOR PROBLEMS WITH MULTIPLE SUB-ACTIONS

Each action set can have more than a single sub-action. In [86], each action set has a pair of sub-actions, namely the

insulin dosage and infusion time. The insulin infusion time can be adjusted, which can be administered either before or after each meal ingestion. The algorithm uses coarse-grained and fine-grained searches to identify the optimal action set.

The coarse-grained search uses two different reward functions to avoid local optima: a) the *hard-reward* value increases when the number of times the blood glucose level is within the normal range after each meal time increases; and b) the *soft-reward* value increases when glucose variability reduces. The hard-reward function selects the optimal insulin dosage during normal operation. However, when the selected insulin dosage at different infusion times yields the same hard reward, the soft-reward function selects the optimal insulin infusion time given the optimal insulin dosage. The hard- and soft-reward functions have been shown to avoid local optima. The selected optimal pair of sub-actions improve convergence and reduce the convergence time.

The fine-grained search uses a heuristic algorithm to perform an efficient exploration to reduce the convergence time. Specifically, it explores a set of potential actions to limit the search space for an optimal pair of sub-actions. The potential actions represent the next directions for optimizing the greedy sub-actions, which mimic the adjustments made by physicians.

The proposed solution has been shown to increase time in range (P.3).

VI. OPEN ISSUES

This section presents open issues which can be investigated to address the shortcomings of RL in diabetes management.

A. INTEGRATING HUMAN INTELLIGENCE INTO RL

Traditionally, artificial intelligence, including RL, operates without human intervention. Nevertheless, in diabetes management, stakeholders including physicians, nurses, pharmacists, and the patients themselves, provide valuable human intelligence. Examples of human intelligence include the physician's knowledge and policies, the patient's inputs (e.g., the announced meal ingestion, and the clinical conditions and symptoms not detected by subcutaneous sensors). Human intelligence can be integrated into RL in three ways. First, in the human-AI approach, the human selects safe and appropriate single or multiple potential actions, and then AI selects and performs the best possible action. This helps to increase the convergence rate during exploration, particularly at the early stage of the operation. Second, in the AI-human approach, AI selects and performs the best possible action, and then the human verifies the action, such as revising the reward value of the selected action to further increase (reduce) the reward of appropriate (inappropriate) actions. Third, in the joint AI and human approach (see Figure 8), the state, action, and reward of both human intelligence and AI are combined. One or more points of interaction between the human and the AI agent are possible. Consider a human h and an agent m, the state combines the human state $s_{h,t}$ and the agent state $s_{m,t}$; the selected action is either the human action $a_{h,t}$ or the agent action $a_{m,t}$; and the reward combines the human reward $r_{h,t}(s_{h,t})$ and the agent reward $r_{m,t}(s_{m,t})$. Further investigation can be pursued to investigate and improve ways of collaboration between humans and AI in human-centered diabetes management.

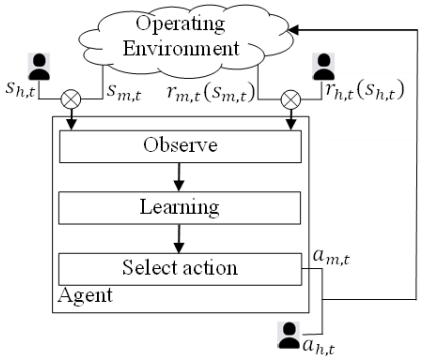

FIGURE 8. Points of interaction between human and RL.

B. COMPLEMENTING LEARNING WITH DATASETS COLLECTED IN THE PAST AND EXPERT KNOWLEDGE

In contrast to the traditional supervised and unsupervised machine learning approaches, RL uses on-policy learning that selects actions and undergoes learning (or training) at the same time. Learning is based on the feedback (i.e., the next state and delayed reward) received from the operating environment. However, exploring random actions on patients with and without safety constraints can be both dangerous and inefficient. The feedback may be affected by measurement errors and noise. On one hand, the safety constraints may not be sufficient to address all kinds of clinical conditions. On the other hand, the safety constraints may be too strict to learn effectively. Using datasets collected in the past and expert knowledge can minimize the reliance on feedback received from the operating environment and increase learning efficiency, and hence the convergence rate. Further investigation can be pursued to investigate and improve ways of complementing learning with datasets and expert knowledge, rather than relying on the feedback received from the operating environment only.

C. CONSIDERING THE EFFECTS OF GLUCOSE-INSULIN-GLUCAGON DYNAMICS

In general, being a model-free approach, the RL agent learns the appropriateness of different actions under different states without the need for a model, such as the transition probability matrix. The model characterizes the glucose-insulin-glucagon dynamics (e.g., the pharmacological delay of insulin and meal ingestion) and clinical conditions (e.g., the blood glucose level). The lack of understanding of the operating environment (e.g., individual patients) leads to various assumptions on the characteristics, effects, and timings of various dynamics; on the other hand, understanding

the dynamics of an individual patient helps to address the complexity and efficiency of the RL model while improving performance.

In the actor-critic RL approach in [88], the critic stores the blood glucose level of the past eight time steps and the current insulin dosage to estimate the value function $V_{\pi_t}(s_t)$. Subsequently, the actor estimates the insulin dosage in the next ten steps. The right number of steps in the past and the future can be affected by the dynamics of the operating environment. In [88], the action representation is $a_t = 0.3 \times V_{\pi_t}(s_t) - 4.4\, E_t$, where E_t represents the extent to which the blood glucose level is outside the normal range [88]. The static values, such as 0.3 and 4.4, in the action representation can be affected by the dynamics of the operating environment.

Further investigation can be pursued to investigate and improve ways of understanding the individual patient's dynamics, then incorporate the findings when determining the designs and hyperparameters of RL models and algorithms.

D. OTHER OPEN ISSUES

Further investigation can also be made to address the following open issues:

- Incorporating physicians' medical knowledge into algorithms.
- Using the trend or predicted blood glucose level to perform early interventions as safety precautions.
- Using historical states and actions (e.g., the blood glucose levels and complications in the past) to increase safety and robustness against measurement errors and sensor noise.
- Tailoring new rules for an individual patient. For instance, the HbA1c target is less stringent for patients who are elderly (e.g., 7.5% for patients aged 65 to 75, and <8.0% for patients aged 76 and above [97]) and who have experienced hypoglycemia.
- Introducing the dynamics of RL models. For instance, the decision epochs, such as the time intervals between visits to physicians, are unequal.
- Incorporating new aspects, such as physical activities, comorbid diseases, prescriptions, the body conditions (e.g., pregnancy and puberty), and races and ethnicities, in simulation. The variation of these aspects in an individual patient can be considered, such as the physical activities during weekdays and over the weekend. Although these factors affect glucose variability and glucose-insulin dynamics, they have not been incorporated in FDA-approved simulators, such as UVA/Padova [66]. Incorporating the new aspects helps to avoid the overestimation of the benefits of RL in diabetes management.
- Performing clinical trials. Although RL has been proposed for a diverse range of healthcare solutions, there have only been a few clinical trials [65].

VII. CONCLUSION

This paper presents a review of the application of reinforcement learning (RL) to improve three main roles in diabetes management, including estimating the blood glucose level (or the insulin dosage), recommending the right type of treatment, and diagnosing diabetes. RL offers three main advantages: a) using the model-free approach; b) optimizing short- and long-term patient outcomes; and c) optimizing policy in an online and offline manner. This enables RL to address the challenges of diabetes management, which revolves around the individual patient's glucose variability and glucose-insulin dynamics, including the pharmacological delay of insulin. Various aspects of training RL in diabetes management are presented, including patient data, system models, simulators, simulation parameters, clinical studies, implementation, experiment designs, and performance measures. A diverse range of RL approaches has been proposed, including the traditional RL, model-based RL, multi-agent RL, actor-critic RL, deep Q-network, the Gaussian process approximation, and the proximal policy optimization. Based on RL, this paper discusses how diabetes can be managed using appropriate representations (i.e., state, action, and reward) and various enhanced algorithms. RL has been shown to increase the similarities with physician's policies, time in range, the proportion of patients achieving the normal blood glucose level, prediction accuracy, and the delayed time period of complications, as well as reduce the HbA1c level, glucose variability, occurrences of hyperglycemia and hypoglycemia, and insulin dosage. RL performance measures, such as the convergence rate and cumulative reward, have also been shown to improve. Finally, this paper presents open issues for further investigation.

REFERENCES

[1] N. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. Da Rocha Fernandes, A. W. Ohlrogge, and B. I. D. F. Malanda, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018.

[2] *World Health Organization*. Accessed: Nov. 11, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] *IDF Diabetes Atlas 2021—10th Edition*. Accessed: Jan. 18, 2023. [Online]. Available: https://diabetesatlas.org/atlas/tenth-edition/

[4] R. Pal, M. Banerjee, U. Yadav, and S. Bhattacharjee, "Clinical profile and outcomes in COVID-19 patients with diabetic ketoacidosis: A systematic review of literature," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 6, pp. 1563–1569, Nov. 2020.

[5] N. Chamorro-Pareja, S. Parthasarathy, J. Annam, J. Hoffman, C. Coyle, and P. Kishore, "Letter to the editor: Unexpected high mortality in COVID-19 and diabetic ketoacidosis," *Metabolism*, vol. 110, Sep. 2020, Art. no. 154301.

[6] 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.

[7] A. Jafar, A. E. Fathi, and A. Haidar, "Long-term use of the hybrid artificial pancreas by adjusting carbohydrate ratios and programmed basal rate: A reinforcement learning approach," *Comput. Methods Programs Biomed.*, vol. 200, pp. 1–11, Mar. 2021.

[8] 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.

[9] L. J. Klaff, R. Brazg, K. Hughes, A. M. Tideman, H. C. Schachner, P. Stenger, S. Pardo, N. Dunne, and J. L. Parkes, "Accuracy evaluation of contour next compared with five blood glucose monitoring systems across a wide range of blood glucose concentrations occurring in a clinical research setting," *Diabetes Technol. Therapeutics*, vol. 17, no. 1, pp. 8–15, Jan. 2015.

[10] M. Shifrin and H. Siegelmann, "Near-optimal insulin treatment for diabetes patients: A machine learning approach," *Artif. Intell. Med.*, vol. 107, pp. 1–21, Jul. 2020.

[11] 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.

[12] 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.

[13] Z. Liu, L. Ji, X. Jiang, W. Zhao, X. Liao, T. Zhao, S. Liu, X. Sun, G. Hu, M. Feng, and G. Xie, "A deep reinforcement learning approach for type 2 diabetes mellitus treatment," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Oldenburg, Germany, Nov. 2020, pp. 1–9.

[14] M. A. Raheb, V. R. Niazmand, N. Eqra, and R. Vatankhah, "Subcutaneous insulin administration by deep reinforcement learning for blood glucose level control of type-2 diabetic patients," *Comput. Biol. Med.*, vol. 148, pp. 1–10, Sep. 2022.

[15] L. Ferrara, M. Joksimovic, and S. D'Angelo, "Could polyphenolic food intake help in the control of type 2 diabetes? A narrative review of the last evidence," *Current Nutrition Food Sci.*, vol. 18, no. 9, pp. 785–798, Nov. 2022.

[16] D. J. Hunter, "Gene-environment interactions in human diseases," *Nature Rev. Genet.*, vol. 6, pp. 287–298, Apr. 2005.

[17] A. Jalalimanesh, H. S. Haghighi, A. Ahmadi, and M. Soltani, "Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning," *Math. Comput. Simul.*, vol. 133, pp. 235–248, Mar. 2017.

[18] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Optimal adaptive control of drug dosing using integral reinforcement learning," *Math. Biosci.*, vol. 309, pp. 131–142, Mar. 2019.

[19] L.-F. Cheng, N. Prasad, and B. E. Engelhardt, "An optimal policy for patient laboratory tests in intensive care units," in *Proc. Pasific Symp. Biocomput.* Singapore: World Scientific, Nov. 2018, pp. 320–331.

[20] M. H. Lim, W. H. Lee, B. Jeon, and S. Kim, "A blood glucose control framework based on reinforcement learning with safety and interpretability: In silico validation," *IEEE Access*, vol. 9, pp. 105756–105775, 2021.

[21] 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.

[22] B. Kovatchev, "A century of diabetes technology: Signals, models, and artificial pancreas control," *Trends Endocrinol. Metabolism*, vol. 30, no. 7, pp. 432–444, Jul. 2019.

[23] Z. Ying, Y. Zhang, S. Cao, S. Xu, and M. Ma, "OIDPR: Optimized insulin dosage via privacy-preserving reinforcement learning," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 5, pp. 1–18, May 2021.

[24] A. Coronato, M. Naeem, G. D. Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101964.

[25] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Comput. Surveys*, vol. 55, no. 1, pp. 1–36, Jan. 2023.

[26] S. Liu, K. Y. Ngiam, and M. Feng, "Deep reinforcement learning for clinical decision support: A brief survey," 2019, *arXiv:1907.09475*.

[27] A. A. Abdellatif, N. Mhaisen, Z. Chkirbene, A. Mohamed, A. Erbad, and M. Guizani, "Reinforcement learning for intelligent healthcare systems: A comprehensive survey," 2021, *arXiv:2108.04087*.

[28] N. Sharma and A. Singh, "Diabetes detection and prediction using machine learning/IoT: A survey," in *Proc. Int. Conf. Adv. Informat. Comput. Res. (ICAICR)*, Shimla, India, 2018, pp. 471–479.

[29] M. Tejedora, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artif. Intell. Med.*, vol. 104, pp. 1–13, Apr. 2020.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, U.K.: Bradford Books, 2018.

[31] J. A. Gantt, K. A. Rochelle, and E. P. Gatzke, "Type 1 diabetic patient insulin delivery using asymmetric PI control," *Chem. Eng. Commun.*, vol. 194, no. 5, pp. 586–602, Jan. 2007.

[32] G. Marchetti, M. Barolo, L. Jovanovic, H. Zisser, and D. S. Seborg, "A feed forward feed-back glucose control strategy for type 1 diabetes mellitus," *J. Process Control*, vol. 18, pp. 149–162, 2008.

[33] S. Del Favero, D. Bruttomesso, F. Di Palma, G. Lanzola, R. Visentin, A. Filippi, R. Scotton, C. Toffanin, M. Messori, S. Scarpellini, P. Keith-Hynes, B. P. Kovatchev, J. H. DeVries, E. Renard, L. Magni, A. Avogaro, C. Cobelli, and on behalf of the AP@home Consortium, "First use of model predictive control in outpatient wearable artificial pancreas," *Diabetes Care*, vol. 37, no. 5, pp. 1212–1215, May 2014.

[34] F. Ståhl and R. Johansson, "Diabetes mellitus modeling and short-term prediction based on blood glucose measurements," *Math. Biosci.*, vol. 217, no. 2, pp. 101–117, Feb. 2009.

[35] 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. 116–125, Feb. 2013.

[36] M. W. Percival, Y. Wang, B. Grosman, E. Dassau, H. Zisser, L. Jovanović, and F. J. Doyle, "Development of a multi-parametric model predictive control algorithm for insulin delivery in type 1 diabetes mellitus using clinical parameters," *J. Process Control*, vol. 21, no. 3, pp. 391–404, Mar. 2011.

[37] M. Breton and B. Kovatchev, "Analysis, modeling, and simulation of the accuracy of continuous glucose sensors," *J. Diabetes Sci. Technol.*, vol. 2, no. 5, pp. 853–862, 2008.

[38] A. Facchinetti, G. Sparacino, and C. Cobelli, "Modeling the error of continuous glucose monitoring sensor data: Critical aspects discussed through simulation studies," *J. Diabetes Sci. Technol.*, vol. 4, no. 1, pp. 4–14, Jan. 2010.

[39] G. Yee, "Bright spots & landmines: The diabetes guide I wish someone had handed me," *Clin. Diabetes*, vol. 35, no. 5, pp. 356–357, Dec. 2017.

[40] C. Toffanin, H. Zisser, F. J. Doyle, 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.

[41] H. Zheng, I. O. Ryzhov, W. Xie, and J. Zhong, "Personalized multimorbidity management for patients with type 2 diabetes using reinforcement learning of electronic health records," *Drugs*, vol. 81, no. 4, pp. 471–482, Mar. 2021.

[42] F. Mansourypoor and S. Asadi, "Development of a reinforcement learning-based evolutionary fuzzy rule-based system for diabetes diagnosis," *Comput. Biol. Med.*, vol. 91, pp. 337–352, Dec. 2017.

[43] M. F. Zohora, M. H. Tania, M. S. Kaiser, and M. Mahmud, "Forecasting the risk of type II diabetes using reinforcement learning," in *Proc. Joint 9th Int. Conf. Informat., Electron. Vis. (ICIEV) 4th Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, Kitakyushu, Japan, Aug. 2020, pp. 1–6.

[44] C. Cobelli, C. Dalla 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.

[45] 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.

[46] 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.

[47] 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.

[48] D. Zeevi et al., "Personalized nutrition by prediction of glycemic responses," *Cell*, vol. 163, no. 5, pp. 1079–1094, Nov. 2015.

[49] M. Eghbali-Zarch, R. Tavakkoli-Moghaddam, F. Esfahanian, A. Azaron, and M. M. Sepehri, "A Markov decision process for modeling adverse drug reactions in medication treatment of type 2 diabetes," *Proc. Inst. Mech. Engineers, H, J. Eng. Med.*, vol. 233, no. 8, pp. 793–811, Jun. 2019.

[50] S. H. Oh, J. Park, S. J. Lee, S. Kang, and J. Mo, "Reinforcement learning-based expanded personalized diabetes treatment recommendation using South Korean electronic health records," *Exp. Syst. Appl.*, vol. 206, pp. 1–12, Nov. 2022.

[51] T. Li, Z. Wang, W. Lu, Q. Zhang, and D. Li, "Electronic health records based reinforcement learning for treatment optimizing," *Inf. Syst.*, vol. 104, pp. 1–11, Feb. 2022.

[52] Diabetes Control and Complications Trial Research Group, "The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus," *New England J. Med.*, vol. 329, no. 14, pp. 977–986, Sep. 1993.

[53] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.
[54] D. M. Nathan, H. Turgeon, and S. Regan, "Relationship between glycated haemoglobin levels and mean glucose levels over time," *Diabetologia*, vol. 50, no. 11, pp. 2239–2244, Sep. 2007.
[55] P. Palumbo, G. Pizzichelli, S. Panunzi, P. Pepe, and A. De Gaetano, "Model-based control of plasma glycemia: Tests on populations of virtual patients," *Math. Biosci.*, vol. 257, pp. 2–10, Nov. 2014.
[56] M. D. Ferdinando, P. Pepe, S. D. Gennaro, and P. Palumbo, "Sampled-data static output feedback control of the glucose-insulin system," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 3626–3631, 2020.
[57] M. De Paula, G. G. Acosta, and E. C. Martínez, "On-line policy learning and adaptation for real-time personalization of an artificial pancreas," *Exp. Syst. Appl.*, vol. 42, no. 4, pp. 2234–2255, Mar. 2015.
[58] E. D. Lehman and T. Deutsch, "A physiological model of glucose-insulin interaction in type 1 diabetes mellitus," *J. Biomed. Eng.*, vol. 14, pp. 235–242, May 1992.
[59] 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.
[60] B. W. Bequette, "Glucose clamp algorithms and insulin time-action profiles," *J. Diabetes Sci. Technol.*, vol. 3, no. 5, pp. 1005–1013, Sep. 2009.
[61] P. Palumbo, S. Panunzi, and A. De Gaetano, "Qualitative behavior of a family of delay-differential models of the glucose-insulin system," *Discrete Continuous Dyn. Syst. B*, vol. 2, no. 2, pp. 399–424, 2007.
[62] A. Noori, M. A. Sadrnia, and M. B. N. Sistani, "Glucose level control using temporal difference methods," in *Proc. Iranian Conf. Electr. Eng. (ICEE)*, Tehran, Iran, May 2017, pp. 895–900.
[63] 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.
[64] A. Mackey and E. Furey, "Artificial pancreas control for diabetes using TD3 deep reinforcement learning," in *Proc. 33rd Irish Signals Syst. Conf. (ISSC)*, Cork, Ireland, Jun. 2022, pp. 1–6.
[65] M. C. Serafini, N. Rosales, and F. Garelli, "Long-term adaptation of closed-loop glucose regulation via reinforcement learning tools," *IFAC-PapersOnLine*, vol. 55, no. 7, pp. 649–654, 2022.
[66] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator," *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 26–34, Jan. 2014.
[67] J. Xie. (2018). *Simglucose v0.2.1*. [Online]. Available: https://github.com/jxx123/simglucose
[68] L. Avila and E. Martínez, "Behavior monitoring under uncertainty using Bayesian surprise and optimal action selection," *Exp. Syst. Appl.*, vol. 41, no. 14, pp. 6327–6345, Oct. 2014.
[69] L. Avila and E. Martínez, "An active inference approach to on-line agent monitoring in safety–critical systems," *Adv. Eng. Informat.*, vol. 29, no. 4, pp. 1083–1095, Oct. 2015.
[70] E. D. Lehmann, C. Tarín, J. Bondia, E. Teufel, and T. Deutsch, "Development of AIDA v4.3b diabetes simulator: Technical upgrade to support incorporation of lispro, aspart, and glargine insulin analogues," *J. Electr. Comput. Eng.*, vol. 2011, pp. 1–17, Jan. 2011.
[71] M. De Paula, L. O. Ávila, and E. C. Martínez, "Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes," *Appl. Soft Comput.*, vol. 35, pp. 310–332, Oct. 2015.
[72] P. Colmegna and R. S. Pena, "Analysis of three T1DM simulation models for evaluating robust closed-loop controllers," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 371–382, Oct. 2014.
[73] B. P. Kovatchev, M. Breton, C. D. Man, and C. Cobelli, "In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes," *J. Diabetes Sci. Technol.*, vol. 3, no. 1, pp. 44–55, 2009.
[74] S. D. Patek, B. W. Bequette, M. Breton, B. A. Buckingham, E. Dassau, F. J. Doyle III, J. Lum, L. Magni, and H. Zisser, "In silico preclinical trials: Methodology and engineering guide to closed-loop control in type 1 diabetes mellitus, *J. Diabetes Sci. Technol.*, vol. 3, pp. 269–282, Mac. 2009.
[75] E. Daskalaki, P. Diem, and S. G. Mougiakakou, "Personalized tuning of a reinforcement learning control algorithm for glucose regulation," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Osaka, Japan, Jul. 2013, pp. 3487–3490.
[76] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," 2016, *arXiv:1606.01540*.
[77] *AndroidAPS*. Accessed: Jan. 1, 2023. [Online]. Available: https://androidaps.readthedocs.io
[78] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 8026–8037.
[79] *PyTorch Mobile*. Accessed: Jan. 1, 2023. [Online]. Available: https://pytorch.org/mobile
[80] A. Chakrabarty, S. Zavitsanou, T. Sowrirajan, F. J. Doyle III, and E. Dassau, "Getting IoT-ready," in *Artificial Pancreas: Current Situation and Future Directions*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 29–57, Accessed: Jan. 1, 2023.
[81] 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. 2019.
[82] K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 603–613, Feb. 2020.
[83] A. S. Brazeau, H. Mircescu, K. Desjardins, C. Leroux, I. Strychar, J. M. Ekoé, and R. Rabasa-Lhoret, "Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes," *Diabetes Res. Clin. Pract.*, vol. 99, no. 1, pp. 19–23, Jan. 2013.
[84] T. B. O'Neal and E. E. Luther, *Dawn Phenomenon*. St. Petersburg, FL, USA: StatPearls Publishing, 2017.
[85] R. W. Beck, R. M. Bergenstal, T. D. Riddlesworth, C. Kollman, Z. Li, A. S. Brown, and K. L. Close, "Validation of time in range as an outcome measure for diabetes clinical trials," *Diabetes Care*, vol. 42, no. 3, pp. 400–405, Mar. 2019.
[86] Z. Wang, Z. Xie, E. Tu, A. Zhong, Y. Liu, J. Ding, and J. Yang, "Reinforcement learning-based insulin injection time and dosages optimization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Shenzhen, China, Jul. 2021, pp. 1–8.
[87] L. Scrucca, M. Fop, and T. B. Murphy, "Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models," *R J.*, vol. 8, no. 1, pp. 205–233, Aug. 2016.
[88] E. Daskalaki, L. Scarnato, P. Diem, and S. G. Mougiakakou, "Preliminary results of a novel approach for glucose regulation using an actor-critic learning based controller," in *Proc. UKACC Int. Conf. Control*, Coventry, U.K., 2010, pp. 1–5.
[89] M. G. Pedersen, G. M. Toffolo, and C. Cobelli, "Cellular modeling: Insight into oral minimal models of insulin secretion," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 298, no. 3, pp. E597–E601, Mar. 2010.
[90] S. K. Garg, S. A. Weinzimer, W. V. Tamborlane, B. A. Buckingham, B. W. Bode, T. S. Bailey, R. L. Brazg, J. Ilany, R. H. Slover, S. M. Anderson, and R. M. Bergenstal, "Glucose outcomes with the in-home use of a hybrid closed-loop insulin delivery system in adolescents and adults with type 1 diabetes," *Diabetes Technol. Therapeutics*, vol. 19, no. 3, pp. 155–163, Mar. 2017.
[91] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, and Y. Chen, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.
[92] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 1–7.
[93] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *J. Healthcare Informat. Res.*, vol. 4, no. 3, pp. 308–324, Apr. 2021.
[94] J. Chen, K. Li, P. Herrero, T. Zhu, and P. Georgiou, "Dilated recurrent neural network for short-time prediction of glucose concentration," in *Proc. 3rd Int. Workshop Knowl. Discovery Healthcare Data*, 2010, pp. 1–5.
[95] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[96] X. Cai, X. Gao, W. Yang, X. Han, and L. Ji, "Efficacy and safety of initial combination therapy in treatment-Naïve type 2 diabetes patients: A systematic review and meta-analysis," *Diabetes Therapy*, vol. 9, no. 5, pp. 1995–2014, Oct. 2018.
[97] A. E. Lee et al., "Improving care in older patients with diabetes: A focus on glycemic control," *Permanente J.*, vol. 20, no. 3, pp. 51–56, Sep. 2016.

KOK-LIM ALVIN YAU (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical and electronics engineering from Universiti Teknologi Petronas, Malaysia, in 2005, the M.Sc. degree in electrical engineering from the National University of Singapore, in 2007, and the Ph.D. degree in network engineering from the Victoria University of Wellington, New Zealand, in 2010. He is currently a Professor with the Lee Kong Chian Faculty of Engineering and Science (LKCFES), Universiti Tunku Abdul Rahman (UTAR), Malaysia. He is also a Researcher, a Lecturer, and a Consultant in applied artificial intelligence and reinforcement learning.

YUNG-WEY CHONG is currently a Senior Lecturer with the National Advanced IPv6 Centre of Excellence, Universiti Sains Malaysia, where she has been a Faculty Member since 2012. Her research interests include Industry Revolution 4.0, ranging from embedded systems, wireless communications, cloud computing, and artificial intelligence. She is a Committee Member of SOI Asia (www.soi.asia), a project that utilizes satellite-based internet to support interactive multimedia communications between partner universities.

XIUMEI FAN received the bachelor's degree from Tianjin University, China, in 1989, and the Ph.D. degree from Beijing Jiaotong University, China, in 2002. She was a Professor with the Beijing Institute of Technology, from 2004 to 2013. She is currently a Professor with the Xi'an University of Technology and the Shaanxi Province Hundred Talents Program. Her main research interests include vehicular networks, mobile internet, edge computing, and various aspects of broadband wireless networks.

CELIMUGE WU (Senior Member, IEEE) received the M.E. degree from the Beijing Institute of Technology, China, in 2006, and the Ph.D. degree from The University of Electro-Communications, Japan, in 2010. He is currently a Professor with the Graduate School of Informatics and Engineering, The University of Electro-Communications. His current research interests include vehicular ad hoc networks, sensor networks, intelligent transport systems, the IoT, and mobile cloud computing.

YASIR SALEEM (Senior Member, IEEE) received the B.S. degree in information technology from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2012, the dual M.Sc. degree in computer science by research from Sunway University, Malaysia, and Lancaster University, U.K., in 2015, and the Ph.D. degree from Institut Mines-Telecom, Telecom SudParis, and Institut Polytechnique de Paris, France, in 2020. He has been a Lecturer in computer science with Aberystwyth University, U.K., since January 2022. His research interests include the Internet of Things, smart cities, vehicular networks, intelligent transportation systems, wireless networks, cognitive radio networks, and privacy preservation.

PHEI-CHING LIM received the B.Pharm. and M.Sc. degrees in clinical pharmacy from Universiti Sains Malaysia, in 2006 and 2016, respectively. She is currently pursuing the Ph.D. degree. She is also a Pharmacist with the public state hospital, namely, Hospital Pulau Pinang. She is actively involved in clinical trials and research, especially in diabetes care.