IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 August 2023, accepted 19 September 2023, date of publication 22 September 2023, date of current version 28 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318324

APPLIED RESEARCH

Evaluation of Offline Reinforcement Learning for Blood Glucose Level Control in Type 1 Diabetes

PHUWADOL VIROONLUECHA[1], **ESTEBAN EGEA-LOPEZ**[1], **AND JOSE SANTA**[2]
[1]Department of Information and Communications Technologies, Universidad Politécnica de Cartagena (UPCT), 30202 Cartagena, Spain
[2]Department of Electronics, Computer Technology and Projects, Universidad Politécnica de Cartagena (UPCT), 30202 Cartagena, Spain

Corresponding author: Phuwadol Viroonluecha (phuwadol.viroonluecha@upct.es)

This work was supported in part by Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER), and Unión Europea (UE) under Grant TEC2017-84423-C3-2-P (ONOFRE-2); in part by MCIN/AEI/10.13039/501100011033 under Grant PID2020-112675RB-C41 (ONOFRE-3) and Grant RYC-2017-23823; and in part by ESF Investing in Your Future. The work of Phuwadol Viroonluecha was supported by AEI Formación de Personal Investigador (FPI) Predoctoral under Grant PRE2018-084260.

ABSTRACT Patients with Type 1 diabetes must closely monitor their blood glucose levels and inject insulin to control them. Automated glucose control methods that remove the need for human intervention have been proposed, and reinforcement learning has been used recently as an effective control method in simulation environments. However, its real-world application would require trial and error interaction with patients. As an alternative, offline reinforcement learning does not require interaction with humans and initial studies suggest promising results can be obtained with offline datasets, similar to classical machine learning algorithms. However, its application to glucose control has not yet been evaluated. In this study, we evaluated two offline reinforcement learning algorithms for blood glucose control and discussed their potential and shortcomings. We also evaluated the influence on training and performance of the method that generates the training datasets, as well as the influence of the type of trajectories used (single-method or mixed trajectories), the quality of the trajectories, and the size of the datasets. Our results show that one of the offline reinforcement learning algorithms evaluated, Trajectory Transformer, is able to perform at the same level as commonly used baselines such as PID and Proximal Policy Optimization.

INDEX TERMS T1D blood glucose control, offline reinforcement learning, transformer, artificial pancreas, machine learning.

I. INTRODUCTION

Type 1 diabetes (T1D) is an autoimmune system disorder involving the destruction of liver β cells of the pancreatic islets of Langerhans due to insulin deficiency. Without enough insulin, glucose cannot enter the cells to transform it into energy. People with T1D need to monitor their blood glucose (BG) levels regularly and take insulin to keep their blood sugar levels within a normal range. Higher (hyperglycemia) or lower (hypoglycemia) blood glucose levels can cause serious health problems such as blindness, kidney failure, or heart attack, so people with T1D must monitor their blood glucose levels and inject insulin to prevent them. There are several insulin delivery methods both manual and automated. The usual insulin delivery method to manage glucose levels is the basal-bolus (BB) regime, which involves taking insulin before meals and at bedtime. A continuous Glucose Monitor (CGM) is a device that measures human plasma glucose levels in real-time. A CGM typically consists of a small sensor that is inserted under the skin, a transmitter that sends the data to a receiver or smartphone, and an application or other interface that displays the glucose levels in real-time. Even combined with a CGM, the disadvantage of BB is the need for manual injection several times per day, which is a trouble, especially for children when they are at school [1].

As a solution, several methods for automated glucose control have been developed. Methods that completely remove the need for human intervention are usually called close-loop controls or artificial pancreas (AP). Those systems

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

additionally include an insulin pump and some method to regulate the injections, that is, a control algorithm. The control algorithms employed usually are predictive integral derivative controllers (PID) [2] and model predictive controllers (MPC) [3]. Both algorithms are effective and widely used [4]. In particular, PID controller is the most used both in commercial and research because of its simplicity and robustness [5]. But these methods are sensitive to external factors such as food intake, exercise, and illness, which affect the control effectiveness [6], [7].

Recently, machine learning (ML), including reinforcement learning (RL), has gained attention in diverse domains such as finance, robots, computer vision or language recognition. ML predictive models can be applied to time series data to understand changes in glycemic state and determine the amount of insulin to deliver. Reinforcement learning is a branch of ML that lets the agent learn by interacting with the environment, which in our case is the simulation of an artificial patient [8]. RL is being applied in diverse domains, including robotic rehabilitation [9], aircraft maintenance [10], and electric vehicle battery lifetime prediction [11]. The RL agent, gathers rewards from outcomes of the agent's action, which uses to learn to take better decisions. Thus, RL algorithms can use physiological data gathered from CGM systems to train the agent. However, this RL process, called online RL, requires extensive trial and error interaction with the environment, the real patient in this case, something that is obviously not safe at the moment. Therefore, online RL has been so far successfully used to automatically control BG [12], [13] but only in *in silico* tests and there is no clear way of bringing it to clinical trials because of the high risk involved when working on real patients.

In contrast, offline RL [14], a recent approach, could solve that problem. Offline RL requires only pre-obtained data to make an agent learn a policy for a particular environment. This data can come from real measurements taken from patients. Thus, this approach does not involve actual interaction with the environment (patient) during the training phase. Offline RL methods have been used in various applications such as marketing [15], web user interfaces [16], sport strategy planning [17], healthcare [18], [19], and T1D blood glucose control [4], [20]. Offline RL is particularly suitable for time series data, such as blood glucose data, due to its ability to learn from historical sequences and capture the temporal dependencies and patterns present in the data.

So, one key advantage of offline RL for blood glucose control is its ability to handle non-stationary environments. Blood glucose levels can vary significantly over time, and offline RL algorithms can adapt to these changes by learning from the entire historical sequence. Offline RL can leverage recurrent neural networks (RNNs) or transformers. These models can capture long-term dependencies and accurately represent the sequential nature of blood glucose measurements, leading to more accurate predictions and decision-making. This enables the agent to capture the dynamics of the underlying system and make appropriate decisions even in the face of changing blood glucose patterns.

The other key advantage is the ability of offline RL to avoid exploration by interaction with the environment. Since offline RL algorithms learn from a pre-collected dataset, which may have undergone extensive safety checks, the risk of dangerous or harmful actions during the learning process is reduced. This is particularly important in the context of blood glucose control, as patient safety is paramount.

On the other hand, there are some challenges in offline RL, including distribution mismatch, biased behavior, sample complexity, off-policy evaluation, and practical deployment. Addressing these challenges requires the development of robust algorithms, novel techniques for policy evaluation, and careful consideration of safety and deployment considerations. Indeed, the suitability of offline RL for BG control has only been started to be discussed in the literature [4], showing that certain offline RL algorithms may be a feasible alternative to online RL, a fact that has not been clearly established yet. Therefore, our first contribution in this paper is to show that *some* additional offline algorithms can actually perform at the level of online ones for BG control.

Moreover, the importance of the implementation details is recognized for both online RL and offline RL [12], in addition to the algorithms used: Online RL requires to design or select the state space, the reward function, and other factors while offline RL, in addition to those choices, requires careful selection of dataset trajectories, as we will discuss in the paper. That is, while for simple environments the states are clearly defined, for most of real problems, including BG control, this is actually a design decision. For example, as discussed in [12], one can use as input state just the last BG sample, or a sequence of past BG samples or a combination of past BG samples and injected insulin doses [13]. The design of the reward function is also a crucial step.

Therefore, a second contribution of our work is to explore and discuss part of this available design space. Our findings and *lessons learned* will be valuable for other researchers, enabling them to focus on other key aspects, which should save testing time, especially considering that training offline RL agents is a highly time-consuming and resource-intensive task.

In summary, the contributions of this work are:

- An evaluation of offline RL as a method for effective blood glucose (BG) control.
- An assessment of the potential and shortcomings of offline RL algorithms for data-driven BG control.
- A comparison of their performance against online RL and PID baselines.
- An exploration of several factors influencing the learning ability of offline RL agents, including the dataset size and its quality.
- Extensive evaluation of different dataset types, sizes and selection approaches.
- The identification of the importance of careful data selection for training offline RL agents.

To facilitate results repeatability, the trained agents, as well as the baseline data and the datasets generated for training for this paper, are available on the open science framework repository [21].

In the remainder of this paper, we first review glycemic control methods and related works. Afterwards, we describe our experimental setup and data generation and the results of our tests. Next we discuss our findings and potential next steps. Finally, we provide concluding remarks.

II. BACKGROUND AND RELATED WORK

A. T1D SIMULATION AND MODELS

For safety reasons, biomedical experiments with machine learning algorithms have been done and pre-evaluated *in silico* through computer simulation. Currently, there are several T1D simulators available, with both free and paid versions, as for instance, AIDA [22], Type 1 Diabetes Virtual Patient Population (T1D-VPP) [23], and the UVA/PADOVA Simulator [24]. AIDA is a free software simulating human plasma insulin and blood glucose for education and research purposes. T1D-VPP involves single (SH) and dual hormone (DH) mathematical models which generate a T1D diabetes virtual population of patients and model the effect of exercise in the glucoregulatory system.

The UVA/PADOVA simulator was originally developed in 2007. It is the first approved *in silico* T1D model by the United States Food and Drug Administration (FDA) [24]. The simulator provides virtual patients in three age groups: adults, adolescents, and children, with 10 patients per group in the free version. In this paper, we use *SimGlucose*, an open-source Python implementation of the UVA/PADOVA simulator [25], previously used in similar studies [8], [13], [26], [27], which can be seamlessly integrated with multiple machine-learning libraries.

B. METHODS FOR GLYCEMIC REGULATION

T1D conditions typically develop in children or young adults and require lifelong treatment with insulin injections. Several insulin regimes are used to control blood sugar. The traditional ones involve one or two injections per day. But patients must control their food intake to be constant throughout the three meals a day. Multiple daily injection therapy, or basal-bolus (BB), offers more flexibility in diet and dosage, but patients still need to control carbohydrate intake and insulin injections [28]. Automatic insulin pumps with integrated continuous glucose monitors (CGMs) have been developed to alleviate the burden of glycemic control and deliver optimal insulin according to current blood glucose levels, allowing patients to live independently without having to worry about delivering insulin. A system that does not requires any human intervention is usually called a closed-loop controller or Artificial Pancreas (AP). Currently, most of the commercially available insulin pumps use a PID (proportional-integral-derivative) algorithm to control blood sugar levels. A PID controller is a control system that uses feedback to adjust a system's output in order to achieve the desired outcome. In the context of blood glucose control, a PID controller is used to regulate the release of insulin in order to maintain a stable blood glucose level [29]. The proportional component of the PID controller adjusts the output based on the current error between the desired and actual blood glucose levels, while the integral component considers the accumulated error over time and the derivative part predicts future errors based on the current rate of change. By combining and tuning these three components, PID controllers can regulate blood glucose levels, but they usually have problems to adapt to disturbances in food intake and need to be customized to individual patients [13], [30].

ML is gaining momentum in AP research recently. ML algorithms can be used theoretically in the field of blood glucose control to develop systems that are able to automatically regulate blood glucose levels according to the individual needs. As other data-driven methods, the idea is to collect labeled data from CGMs and other devices and train a ML model. Through the training process, ML algorithms would ideally identify patterns and trends in order to learn how to predict blood glucose (BG) levels and adjust insulin levels accordingly. At this point, there are several alternatives. The first one is to use the ML model to just predict the expected BG level ahead of time and then use some other method to decide the insulin dose required to keep BG at the desired level. However, the human response to insulin is highly non-linear and it is also difficult to predict the response to the insulin injection. Therefore, another alternative is to learn that response with ML methods also. To this purpose, Reinforcement Learning (RL) could be used, since the ML agent directly learns the appropriate action (insulin doses) to take given a certain input state (the patient BG history). This is what traditional RL (also called *online RL*) does, by letting the agent interact with the environment, and receive a reward as a result of this interaction. By maximizing the cumulative rewards, the agent effectively learns how to adjust insulin levels. That is, by learning through trial and error, the agent could potentially develop effective strategies for maintaining healthy blood glucose levels over time. This method has successfully been used to automatically control BG levels in *in silico* trials, outperforming PID-based methods [12], [13].

However, the main drawback of this approach is that it is not clear at all how to apply it to real patients, that is, how to transfer the learning from the *in silico* environment to real patients. Although data (BG level, physical activity, etc.) can be automatically collected from real patients from electronic devices, RL agents still need to experiment with the patient response in order to learn.

To solve this issue, a more recent approach, called *offline reinforcement learning*, has emerged. In offline reinforcement learning, the agent is not able to receive any feedback from its environment during the learning process, and must instead learn only from previously collected data [14]. This means that the agent must learn to make decisions based on the information that is available (no exploration). Note that the main difference with other ML methods is that with offline RL the actions and rewards are also given as input data. For example, a typical supervised ML algorithm uses collected BG levels (as well as other context data) to train and is able to predict the next BG level, given a certain input BG history. On the contrary, to train an offline RL agent we need to use BG levels, actions taken and observed rewards, and, once trained, it is able to predict the required action, given a certain BG history as input.

The advantage is that it is useful in situations where it is not possible or practical to experiment with the environment, such as when working with historical data or in safety-critical environments. As a drawback, note that, although it removes the need to interact with the environment to learn, it still leaves open the question of how to collect the required states, actions and rewards for training, which is not obvious for many practical situations. In this paper, since we can collect those data from simulations, we put aside temporarily this question and focus on evaluating how effective is offline RL for BG control. Let us finally remark that the value of offline RL is that it is able to effectively generalize, that is, to apply the appropriate action to an input not previously seen in the training dataset. In other contexts, ML has proved to be very effective generalizing [31], but to the best of our knowledge, the generalizing performance of offline RL for BG control has only been started to be discussed in the literature [4]. Our goal in this paper is to evaluate it and discuss factors that may have an influence in the learning and prediction performance.

In particular, we evaluate the following offline RL algorithms: Decision Transformer **(DT)** [32] and Trajectory Transformer (TT) [33]. Both of them approach offline RL as a sequence modeling problem, that is, the agent is trained with sequences of observations, actions and rewards (*trajectories*) and its goal is to generate sequences that result in high rewards. We summarize their features:

Decision Transformer (DT) [32]:

- Trajectory representation: $\tau = \hat{R}_1, s_1, a_1, \hat{R}_2, s_2, a_2, \ldots, \hat{R}_T, s_T, a_T$
- Uses return-to-go ($\hat{R}_t = \sum_{t=1}^{T} r_t$) instead of rewards
- Input of DT is a subset of the trajectory τ consisting of the K most recent time steps

Trajectory Transformer (TT) [33]:

- Trajectory representation: $\tau = s_t^1, s_t^2, \ldots, s_t^N, a_t^1, a_t^2, \ldots, a_t^M, r_{t\,t=0}{}^{T-1}$
- Uses discretized states and actions as input, along with a scalar reward
- Augments the trajectory with return-to-go as in DT and employs a beam search algorithm for planning [34]

Both DT and TT uses as architecture for action prediction a transformer network. The transformer is a type of deep learning model that is designed to process sequential data which was introduced by Vaswani et al. in 2017 [35]. The transformer architecture is based on the idea of using self-attention mechanisms to process input data, rather than using traditional convolutional or recurrent layers. This allows the model to capture long-range dependencies in the data and to process the input sequence in parallel, which makes it faster and more efficient than many other types of models. A key aspect determining the performance of offline RL algorithms is the quality of the datasets used for training. In fact, their performance is usually validated separately according to the quality of the trajectories included in the dataset. For instance, the quality of the dataset can range from randomly (random dataset) generated trajectories to trajectories generated by the best-performing algorithm (expert dataset) or a mixture of them [32], [33].

C. RELATED WORK

Most of the current commercially available control algorithms for AP systems are based on PID or Model-Predictive-Control (MPC) [6], [7], [29], [36], [37], [38], [39]. PID and MPC controllers usually require the user to announce her meal intake and exercise activity, and so they work as an hybrid closed-loop system, [29]. PID and MPC are used in current FDA-approved products such as MiniMed systems, Control IQ, and Dexcom [5].

The main drawback of PID controllers is that usually do not handle well variability in food intake [13], [36]. Several improvements of the basic PID control have been put forward, such as insulin feedback (IF), which increases its efficiency [38], [39]. MPC controllers use a mathematical model to predict and control BG levels. It involves using a mathematical model of the patient's physiology to predict future blood glucose levels and optimizing a sequence of control actions over a specific time horizon. The process includes modeling the patient's dynamics, predicting glucose trajectories, formulating an optimization problem to minimize a cost function while satisfying constraints, implementing the first control action, and repeating the process in a receding horizon manner. MPC offers benefits such as dynamic adaptation and incorporation of safety constraints, but challenges include model accuracy, patient-specific parameters, and computational requirements. Di Ferdinando et al. [6] and Borri et al. [7] model the endogenous insulin delivery rate (IDR) with nonlinear differential difference equation (DDE). These models usually are applied to T2DM patients, since IDR cannot be neglected for them. Finally, overnight hypoglycemia is dealt by PID and MPC-based commercial products with Predictive Low Glucose Suspend (PLGS) technology [29], which predicts glucose concentration trends and suspends insulin delivery before hypoglycemia occurs.

ML has been used as a tool for the prediction of diabetes [40], [41], [42], [43], but also for glycemic control in an insulin pump, and such techniques are growing rapidly within the artificial pancreas research community. Most ML experiments are done *in silico*, through computer simulation. As CGM data are time series, non-linear autoregressive neural networks are used for BG prediction in [23], while [44], [45], [46], [47], [48] use recurrent neural networks (RNN) and long short-term memory (LSTM).

For BG control, RL has been increasingly tested, using multiple RL methods such as double score strategy [49], Q-learning [50], [51], [52], Deep Q-network (DQN) [53], Deep Deterministic Policy Gradient (DDPG) [54] and its improvement Twin Delayed DDPG (TD3) [55], Soft Actor-Critic [13], [56] and Proximal Policy Optimization [12]. These RL methods are called online RL, since the agent interacts with the environment to collect data. As an example, our previous work [12], shows a simple RL implementation strategy that outperforms PID with IF for BG control in *in silico* tests. The recent work of Yu et al. [57] uses a meta-RL framework called active RL with personalized embeddings (ARLPE). By learning a general meta-policy and then fine-tuning it to the particular patient, their results show very promising results. However, their results have only been tested for adult and adolescent cohorts, excluding the most difficult group to train, the children. In addition, it remains the question of how to actually do the fine tuning for real patients, which cannot be done on the simulator. That is in fact, the main problem of online RL. A potential alternative to alleviate this problem is to use a model-based RL approach, such as the recent one in [58], where a hypothesized insulin dose is simulated on a BG predictor before actually being injected to the patient. Its performance is good for simulations up to 12 hours and two meals but decreases in more realistic scenarios.

In summary, in spite of recent advances, online RL is not yet suitable for safety-critical environments, where interaction with the environment (the real patient) is not possible. Therefore, recently, researchers have paid more attention to offline RL. Offline RL is similar to online RL, but the offline RL agent does not need to interact and receive any new information from the environment during the learning process [14]. This means that the agent instead learns from previously collected data, which is safer and more useful for medical and healthcare research. Only a few works have evaluated the use of offline RL for BG control, such as [20], which uses Simulation-Augmented Batch RL (SABR), and [4], which applies and compares three offline RL techniques: Batch Constrained Deep Q-learning (BCQ), Conservative Q-learning (CQL) and Twin Delayed DDPG with Behavioural Cloning (TD3-BC). The work of Fox demonstrates how offline RL can reduce risks over two months and two years of evaluation. The work of Emerson et al. shows that TD3-BC outperformed PID across all patients. This is the work most similar to ours in this paper, but there are significant differences: first, we evaluate more recent offline RL algorithms (DT and TT), which have shown better results than the ones used by Emerson *et. al*. Second, their work only evaluates 9 patients, 3 from each of the three group ages available at SimGlucose, while we evaluate all the virtual patient population, 30 patients. Finally, their training dataset only contains 10^5 samples generated by PID for each patient, while our datasets contain 1 million sample per patient and have been generated with PID-IF and our previous online RL implementation. As we said, for offline RL it is key to evaluate the influence of the training dataset, so we have extensively evaluated this aspect by: trying different dataset sizes, using those two types of datasets, mixing them and selecting the best subset of trajectories.

III. MATERIAL AND METHODS

In this section we describe our evaluation of offline RL as a method for automatic BG control. We evaluate two offline RL algorithms, *Decision Transformer* [33] (DT) and *Trajectory Transformer* [32] (TT). Each of the algorithms have been trained with two different sets of datasets, one generated by our previous online RL BG controller, PPO-RNN [12], and another one generated from a PID-IF controller [12], [39]. We also use those methods as baselines for comparison. In the remaining of the paper, each combination is referred to as **Decis-PPO**, **Traj-PPO**, **Decis-PID-IF** and **Traj-PID-IF**, respectively. In addition, a dataset that mixes trajectories from both methods (**PPO-RNN** and **PID-IF**) is also used to evaluate both algorithms. As metrics used to determine whether the glycemic control algorithm works appropriately, we use the percentage in time in euglycemia or Time in Range (TIR). In both cases they refer to the time spent in the target glycemic level range between 70 and 180 mg/dL. Lower (hypoglycemia) and higher ranges (hyperglycemia) may cause short-term and long-term complications in T1D. Most diabetics should aim for a TIR of at least 70 percent of readings [4].

We first describe the baselines and the experimental setup and then discuss our evaluation results. Our general goal is to determine whether offline RL is a feasible method for automated BG control and how the quality and size of the datasets influence the learning process.

A. BASELINES

1) PROXIMAL POLICY OPTIMIZATION (PPO-RNN)

In a previous paper we proposed and evaluated a RL control based on the PPO [59] algorithm [12]. One key finding of our previous work was that we were able to successfully train the agents if we selected a proper observation frequency for each type of patient, different from the default 3-minute CGM samples. That is, instead of using the default frequency of the CGM sensor, observations were made every 45, 30 and 15 minutes for adults, adolescents and children respectively. In addition, a simple reward function, shown in eq. (1), was used.

$$reward = \begin{cases} 1, & \text{if } BG \in [70, 180]\ mg/dL. \\ 0, & \text{if } BG \in [10, 69] or [181, 1000]\ mg/dL. \\ -100, & \text{otherwise.} \end{cases} \tag{1}$$

With this implementation strategy, we showed that the PPO agent outperforms other control methods and is able to keep over 73% of time in euglycemia across all groups.

2) PROPORTIONAL INTEGRATIVE DERIVATIVE WITH INSULIN FEEDBACK (PID-IF)

In our previous paper [12] we also tested a PID control that aims to keep the BG level at a target point of 112.517 mg/dl, which is the zero-risk point in Clake's Risk Index. Note that PID-IF includes insulin feedback [38], [39]. Insulin feedback is an adjustment of insulin delivery that adapts to metabolism changes due to life activities and has been shown to improve the performance of PID controls. Therefore, formally, the trajectories used as input for both offline RL agents use $s_t^i = b_t$, $a_t^i = u_t$ and $r_t^i = rw_t$, where b_t, u_t, and rw_t are, respectively, the BG sample, the insulin units and the reward from eq. (1) at time t. In addition, offline RL uses additional input consisting of a terminal flag indicating whether the patient's BG is below 10 or above 1,000 mg/dL, and a timeout flag indicating whether the patient survived the full episode length. Thus, each sample in the dataset corresponds to an agent trajectory and consists of five vectors, with the mentioned data.

For our previous work we implemented the PID-IF control for the default observation frequency (OF) of 3 minutes for all patients. However, in this paper we want to combine PID-IF trajectories with PPO-RNN for training the offline RL agents. Since the PPO-RNN agents use different observation frequencies for each group age, as discussed previously, we have to adapt the PID parameters, proportional, derivative and integral constants, K_p, K_d, and K_i, for that particular frequencies. To accommodate different observation frequencies, the utilization of new OF values may necessitate the discovery of new PID parameters. Hence, in order to adapt to varying observation intervals and the corresponding insulin response based on age groups, the optimization framework Optuna [60] is employed to identify suitable PID parameters. After evaluating various optimization methods, including Tree-structured Parzen Estimator (TPE) [61], we determined that TPE exhibited the best performance. Subsequently, we conducted 1000 trials using TPE to identify the optimal PID parameters that resulted in the highest euglycemia percentage within a 10-day episode length. The optimal PID parameters for each patient are provided in Table 1.

In summary, in this work both baselines, PPO-RNN and PID-IF, use the same observation frequency; 15, 30, and 45 minutes for children, adolescents, and adults, respectively. Finally, meals were randomly generated by the Harris-Benedict algorithm [13] and used along in data generation for training and evaluation.

The baseline data as well as the datasets generated for training for this paper are available on the open science framework repository [62], in CSV format.

B. EXPERIMENTAL SETUP

We used the open-source implementations of TT and DT, available at [32], [33]. For training and evaluation, we used the SimGlucose: python framework based on the UVA/Padova simulator, with 30 virtual patients divided into three groups: adults, adolescents and children, with 10 subjects each [25]. The parameters of patients were obtained from the academic edition of the commercial UVA/PADOVA simulator version 2008, according to the developer [63]. This simulator is based on the Open AI Gym standard [64], which is compatible with RL algorithms and easy to adapt to various kinds of research. It also provides different types of CGM sensors, insulin pumps, and a random meal scheduler with noise. SimGlucose has been previously used in similar studies [8], [13], [26], [27]. We trained DT and TT with the datasets generated by our baselines previously described.

TABLE 1. Optimal PID parameters obtained by optuna.

Patient	K_p	K_i	K_d
adolescent#001	-0.000291775	-1.42915E-07	-0.01999
adolescent#002	-0.000428201	-1.43021E-07	-0.00987
adolescent#003	-0.000187463	-6.29647E-08	-0.00785
adolescent#004	-0.000188523	-1.12114E-07	-0.00912
adolescent#005	-5.23529E-05	-1.76362E-07	-0.01109
adolescent#006	-8.65727E-10	-2.96707E-11	-0.01167
adolescent#007	-1.03457E-07	-8.77117E-08	-0.00846
adolescent#008	-3.34156E-10	-8.98967E-12	-0.00927
adolescent#009	-0.000118396	-1.73358E-07	-0.00774
adolescent#010	-2.237E-10	-5.3542E-12	-0.01215
adult#001	-0.000255779	-8.80847E-08	-0.01967
adult#002	-0.000762343	-1.35421E-07	-0.01966
adult#003	-4.93202E-10	-1.32181E-07	-0.01304
adult#004	-0.000187846	-1.10494E-07	-0.00892
adult#005	-0.000401528	-1.12032E-07	-0.01999
adult#006	-0.001015064	-1.02666E-06	-0.02417
adult#007	-0.002457841	-9.76956E-06	-0.0179
adult#008	-0.000164119	-1.23146E-07	-0.01839
adult#009	-0.0001885	-1.64768E-07	-0.01997
adult#010	-0.000165964	-3.62289E-08	-0.01791
child#001	-4.32616E-05	-4.99315E-07	-0.0012
child#002	-2.43848E-05	-1.19047E-08	-0.0063
child#003	-0.000114261	-2.2317E-08	-0.0019
child#004	-0.000122317	-9.84608E-07	-0.00171
child#005	-0.000144505	-2.35487E-08	-0.01025
child#006	-8.50475E-05	-4.07014E-07	-0.0017
child#007	-6.38112E-05	-7.54145E-08	-0.00464
child#008	-6.03971E-05	-1.14231E-07	-0.00226
child#009	-6.68974E-05	-1.83219E-07	-0.002
child#010	-8.80842E-06	-5.85201E-08	-0.00395

1) DATA GATHERING

Initially, we generated three groups of datasets for training the offline RL agents. Each dataset contains five features: observation, action, reward, terminal, and timeout. An observation is the current CGM state; an action is an amount of delivered insulin, and the reward is genereated according the reward function in eq. (1), described in [12]. A terminal is True when the patient's BG is under 10 or over 1,000 mg/dL, which is considered a *catastrohpic failure* and timeout is True when a patient *survived* for 10 days, that is, there was no catastrophic failure in the 10 days. In the first stage, we used the datasets generated from baselines - PPO and PID-IF. The size of each dataset is one million samples per patient, so we generated 30 million samples in total. The second stage considers a combination of PPO and PID-IF datasets, since we hypothesize that if we combine data from multiple sources, the agents may learn better. Thus, we sorted the datasets by the highest rewards and then mixed the datasets as follows: the first one with 80% samples from PPO and 20% from PID-IF 20%, and a second one with 50% of PPO and 50% of PID-IF. A new mixed dataset for each patient was generated. Finally, to test the influence of the dataset size in the learning process, in the final stage, we generated new datasets from the sorted baselines ones, by reducing the number of samples to one hundred thousand and ten thousand. In total, there are three groups of datasets for each patient: two baseline datasets, two combined datasets, and two reduced datasets, as shown in 1.

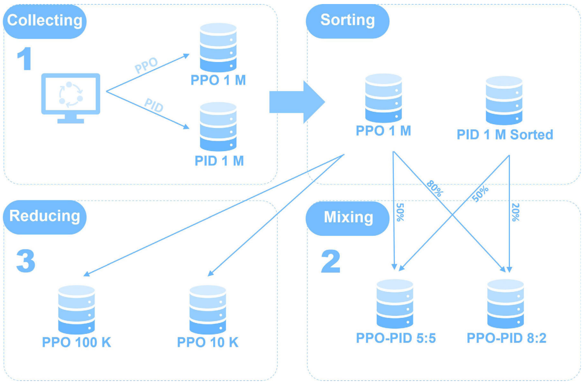

FIGURE 1. Schema of data generation from PPO and PID-IF algorithms.

2) TRAINING

We trained the offline RL agents for each patient and dataset with the original hyperparameters from its code repositories [32], [33]. Hyperparameter tuning has not been considered for this work because, first, we are concerned at this point about whether offline RL is a feasible method for BG regulation and the general factors that may have influence in the training process independently of the particular algorithm used; and second, because adding a hyperparameter optimization process on top at this stage was unfeasible due to the time and resources needed to test all the combinations of algorithms and datasets considered in this work. Once we have identified a promising algorithm We intend to perform a thorough hyperparameter tuning on it, using advanced methods such as the one in [65].

3) EVALUATION

We evaluated all the offline RL agents and dataset combinations, as well as the baselines, using 20 simulation replications with different seeds, per patient. Each replication is run for 10-days of simulation time, so each episode is 10-days long. The observation frequency is 45 minutes, 30 minutes, and 15 minutes, for adults, adolescents and children, respectively. The termination due to catastrophic failure (BG level under 10 or above 1,000 mg/dL) is identical to the one used in the training process. TIR or euglycemia fraction of time as well as hyperglycemia, hypoglycemia fractions and Clarke's risk index [66] are the metrics used for evaluation and comparison between DT and TT with different datasets.

IV. RESULTS

In the following sections we compare the different alternatives. A paired t-test has been done with the results for pairs of alternatives, between both offline RL and against the PPO and PID alternatives. In all the cases, it has been found that there are significant differences, with a p-value below 0.05, except for the combination of datasets.

A. EPISODE LENGTH

Our first test is to determine whether offline RL agents are able to avoid catastrophic failures. We simulate each virtual patient for a fixed duration of 10 days in order to compare the performance of different methods. Although blood glucose control is a continuous task that lasts indefinitely, this limited episode length allows us to assess and compare the percentages of euglycemia achieved by the different methods within a standardized timeframe.

Our results in Fig. 2 show that offline RL Trajectory and Decision Transformers cannot outperform PID-IF and online RL PPO-RNN, and cannot reach ten days as the baselines, which means that BG level reaches a value outside the 10-1000 mg/dL. **Traj-PPO** achieves the longest average episode length. It reaches an average episode length of over 4,000 steps (8 days of simulated time) in every age group. There are notable differences for each group and method, without a clear trend. In the following sections we look at the fraction of time spent at each state during the episode and discuss reasons for this behavior.

B. RISK INDEX AND GLYCEMIC STATES

We now compare the glycemic state, that is, the fraction of time spent in each BG range. In Fig. 3a, we show all methods for all age groups. **Traj-PPO** achieves the highest median euglycemia of offline RL methods. Its median and 75 percentile slightly outperform the PID-IF baseline. On the contrary, when trained with PID-IF trajectories, **Traj-PID-IF**, it exhibits a poor performance. The performance of the Decision Transformer is bad with all the datasets tested. The results show clearly that offline RL cannot learn properly how to control with PID-IF trajectories. In fact, **Decis-PID-IF** has the highest hyperglycemia fraction, while **Traj-PID-IF** has the highest hypoglycemia fraction. We can see in Fig. 3b the glycemic state by age group. **Traj-PPO** shows good performance across all age groups and even its median hyperglycemia in all groups is better than the original online PPO. However, its hypoglycemia median and 75 percentile are high and have a broad range, meaning that **Traj-PPO** implies a high low blood glucose risk, a serious concern in modern AP products. **Decis-PPO**, in its turn, shows unacceptable high ranges for both hypoglycemia in adults and hyperglycemia in adolescents and children.

Actually, the risk index, evaluated in Fig. 4, provides a more summarized view of the relative danger of hyper and hypoglycemic states, and shows that the riskiest method when attending to hyperglycemia is **Decis-PID-IF**, while

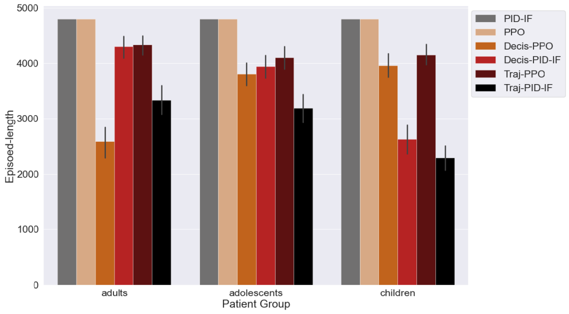

FIGURE 2. **Fraction of completed 10-day evaluation reached for each method and group.**

hypoglycemia is more frequent in adults, adolescents, and children when using **Decis-PPO**, **Traj-PPO**, and **Traj-PID-IF**, respectively. The information of the risk index is complemented by the percentage of time spent in severe hyperglycemia (>250 mg/dL) or hypoglycemia (<50 mg/dL) shown in Fig. 5. It can be seen that **Traj-PPO** spends slighly less time in severe hyperglycemia than the other methods, but more time in severe hypoglycemia, which is adequately captured by the risk indexes.

As summary from this section we can conclude that **Traj-PPO** provides a level of performance similar to online PPO and PID-IF, but it has serious issues with hypoglycemia, that is, tends to inject too much insulin. In the following sections we come back to this matter.

C. COMBINATION OF PPO AND PID-IF DATASETS

We compare the **Decis-PPO**, **Decis-PID-IF**, **Traj-PPO** and **Traj-PID-IF** with the combined datasets of PPO and PID-IF with two different ratios: eight to two (PP82) and five to five (PP55). In Table 2, we show the variation in percentage of the average episode length. We can see that the use of mixed datasets does not improve TT. On the contrary, it worsens its performance for all glycemic states. For DT, the mixed dataset slightly increases its performance for children and adolescents compared to **Decis-PID-IF**, and very clearly for adults. In Fig. 6 the global euglycemia in all methods is about the same level at 40%. However, the DT with both datasets performed well in avoiding hypoglycemia. TT has the same high and low glycemic risks. In terms of RI, from Fig. 7, we can see all DT and TT cases with mixed datasets range in 20-40 and they are outperformed only by the previous **Decis-PID-IF**.

In Table 3 we show the average daily dose of insulin injected by each method. As can be seen, there is a direct correlation, as expected, between the daily dose and the time spent at each glycemic state shown in Fig. 3a. Moreover, in Table 4 we show, in percentage, whether the catastrophic events of each method are due to hyperglycemia or hypoglycemia.

From these data, we see that the average insulin dose of **Traj-PPO** is higher than that of PPO and that all the catastrophic events of **Traj-PPO** are due to hypoglycemia, while in **Decis-PPO** they are practically balanced. When mixing the datasets, the proportion of catastrophic events due to hyperglycemia increases for all the methods.

With these tests our aim is to decide if the offline agents may improve their performance when trained with a ‘‘more’’ distributed dataset, that is, with a dataset with a potentially wider range of states and actions. Our results show that transformers cannot generalize adequately. We conclude that more care has to be put in selecting the trajectories for the datasets. For instance, when ordering the trajectories we just look at the highest rewards, but the average BG level of those trajectories is not taken into account. **Traj-PPO** only has catastrophic events due to hypoglycemia because it tends to keep patients on a low BG level. Due to our reward function, such kind of trajectories may have a reward which is high but equal to other trajectories that keep the patient on a higher BG level, which would be better. Such considerations have to be taken into account when creating the training datasets.

D. DATASET SIZE

The dataset size is important because one cannot realistically expect to collect samples from patients for years and so we want to test how much we can reduce the dataset to get good enough results. Interestingly, from Table 5, both DT and TT with 100k size have longer episode lengths than 1M size on average. This is due to the fact that we sorted and use only the best trajectories. And the average euglycemia percentage is almost the same level as the 1M dataset. The difference in euglycemia for TT is 0.47% and 1.8% for DT. While hyperglycemia between 100k and 1M datasets in TT decreases, in DT it increases by almost 10%. As a result, TT globally improves performance with 100k and has better RI than 1M, because it is able to reduce severe hypoglycemia, as shown in Fig. 5, but DT with 100K slightly decreases an already poor performance. Clearly, DT and TT are less effective when the amount of data was reduced to 10k. Both methods had a decrease of more than 10% in TIR and a significant increase in RI. Additionally, the computational time is also affected: a dataset comprising 1M samples required approximately 26 hours of training time per patient. For a dataset of 100k samples, this value was reduced to approximately 21 hours, and further decreased to around 19 hours for a dataset of 10k samples.

E. HYPERPARAMETER TUNING FOR DECISION TRANSFORMER

So far, we have shown that DT systematically performs worse than its alternatives. We have checked whether the DT bad performance is due to an incorrect selection of hyperparameters. DT takes a subset of the trajectory τ as input, specifically, the K most recent time steps. Each time step consists of three items: the return-to-go, state, and action. The default value $K = 20$ is considered in the algorithm implementation. Consequently, we conducted experiments by setting K to different values, namely 10, 50, and 100, and training two agents with different performance, using a PPO dataset comprising 100K samples. Table 6 illustrates the

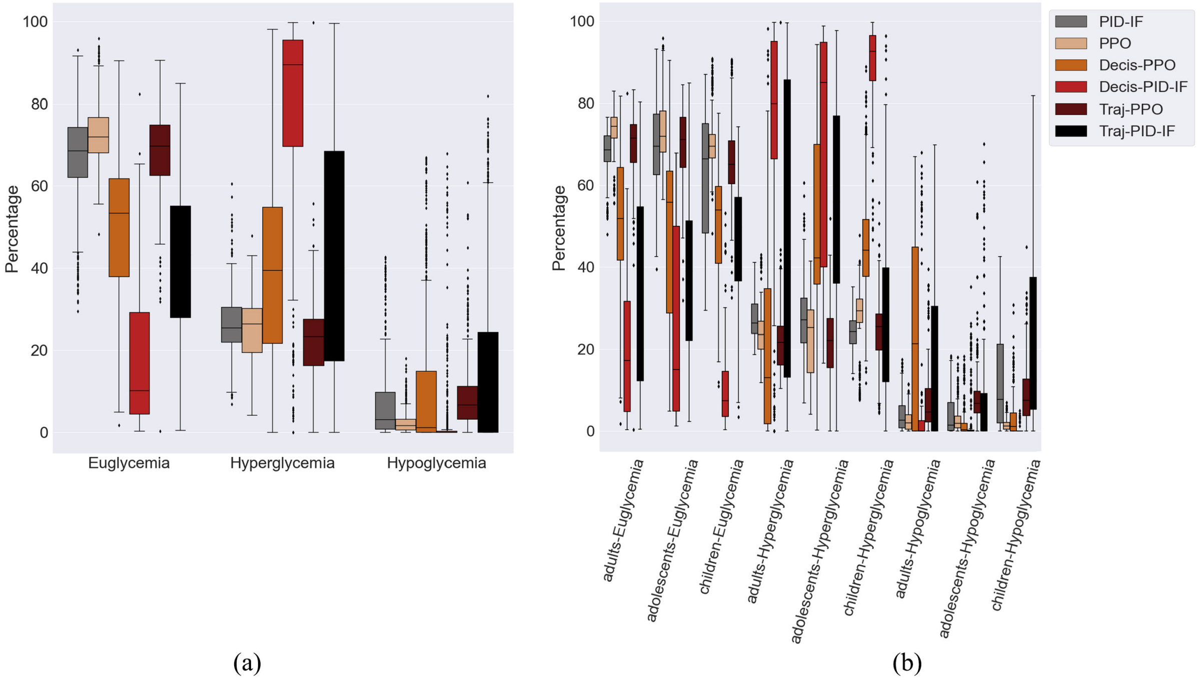

FIGURE 3. Comparative fraction of time spent in glycemic states. (a) global glycemic state. (b) glycemic state by group.

TABLE 2. Increase/reduction of completed 10-day episodes of mixed datasets reaced for each method and group.

Method	Group	PP55/PID	PP82/PID	PP55/PPO	PP82/PPO
Trajectory	children	-42.35%	-45.09%	-68.17%	-69.69%
	adolescents	-48.19%	-44.23%	-59.60%	-56.52%
	adults	-29.03%	-2.33%	-45.31%	-24.74%
Decision	children	20.13%	12.71%	-20.28%	-25.20%
	adolescents	1.92%	3.64%	5.68%	7.47%
	adults	4.01%	2.55%	72.90%	70.49%

TABLE 3. Average daily insulin dose.

Method	Average daily dose
Decis-PID-IF	6.9
Decis-PP-55	7.7
Decis-PP-82	8.5
Decis-PPO	9.5
PID-IF	10.7
PPO	10.9
Traj-PID-IF	9.7
Traj-PP-55	16.3
Traj-PP-82	16.0
Traj-PPO	13.3

TABLE 4. Type of catastrophic events by methods (boldface highlights the higher risk of each algorithm).

Method	Hyper%	Hypo%
Decis-PID-IF	80.10%	19.90%
Decis-PP-55	84.18%	15.82%
Decis-PP-82	82.70%	17.30%
Decis-PPO	51.68%	48.32%
Traj-PID-IF	58.11%	41.89%
Traj-PP-55	25.11%	74.89%
Traj-PP-82	35.76%	64.24%
Traj-PPO	0.00%	100.00%

impact of different K values on performance. As can be seen, the performance remains quite consistent, particularly when comparing K=10 and K=20.

V. DISCUSSION

In our previous paper, [12] and similar works [13], PID and PPO agents performed considerably well for BG control in the T1D simulator, so our hypothesis was that offline RL with these datasets should have comparable performance. Our results show that at least **Traj-PPO** has a performance similar to that of online PPO in most of the metrics, which is promising, since the main goal of this work is to determine whether offline RL can be a realistic alternative for data-driven BG control, before attempting clinical trials with real patient data. Our results also agree quantitatively with the work of [4], which shows a similar level of performance, although tested with fewer patients and different algorithms. Our evaluation also shows that training offline RL is not straightforward: neither all the algorithms tested nor the datasets used were equally effective in learning. It suggests that a better understanding of the influence of different data aspects and careful planning and design of the data-gathering is still necessary before collecting real-patient data for further tests, which is a complex and time-consuming task.

More research is needed to correct some of the observed deficiencies of offline RL methods. Most importantly, to prevent the inability to achieve full episode length without catastrophic failures. Unlike the baselines, the average

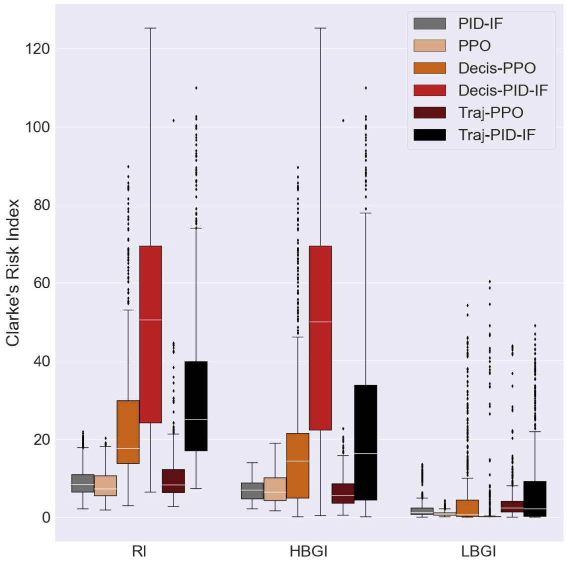

FIGURE 4. Comparative fraction by global risk index.

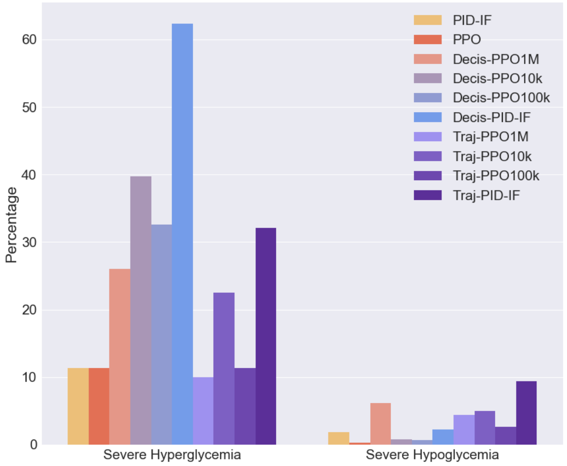

FIGURE 5. Percentage of time spent in severe hyperglycemia (>250 mg/dL) or hypoglycemia (<50 mg/dL) per method.

episode length cannot reach the full 10 days, even though the best one, **Traj-PPO**, reaches almost 9 days globally.

In Table. 4, the catastrophic event of **Traj-PPO** is 100% due to hypoglycemia, while no catastrophic hyperglycemia occurred. Thus, additional research is needed to ensure that **Traj-PPO** is able to avoid hypoglycemia and thus able to achieve the full episode length and higher TIR. A direct next step is to further improve the quality of the training dataset to avoid hypoglycemic trajectories, as discussed below.

From our results, it is also clear that DT is not able to deliver good performance in this task, showing unacceptable high hyperglycemia levels in some groups. A simple reason may be that we have not optimized the DT hyperparameters, in particular, the minibatch sequence length, to which DT is sensitive for several tasks. However, the preliminary tests conducted to check if that was the case, seem to rule it out. There may be the need for deeper adaptations, such as pretraining or architectural changes, which have been shown to improve the basic DT performance [67], [68], [69], [70]. We leave the improvement of DT behavior as future work. Training with the **PID-IF** dataset did not yield satisfactory results for any of the algorithms. It seems that **PID-IF** generates too many out-of-distribution samples, that is, actions that move the state to not previously seen states which degrade the performance [14].

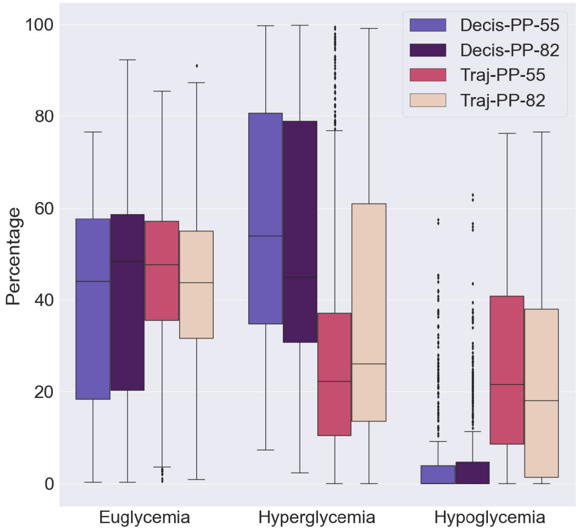

FIGURE 6. Comparative fraction of time spent of mixed datasets in global glycemic state.

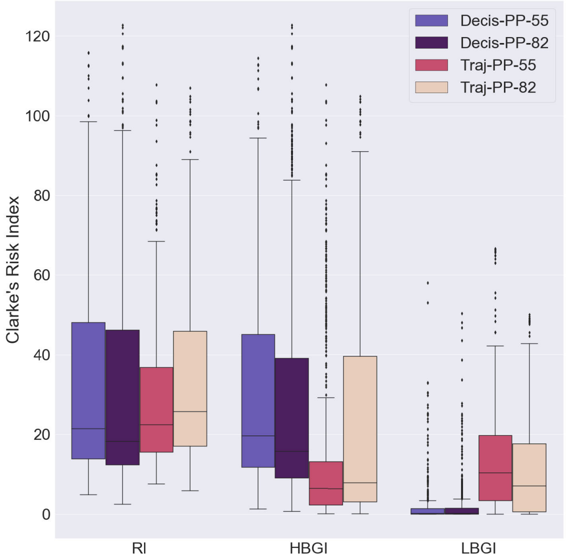

FIGURE 7. Comparative fraction of global risk index of mixed dataset.

We sorted data by reward and length of the episode, then combined sorted PID and PPO datasets to determine if we can improve the learning process of the offline RL. Unfortunately, just a crude mixing, even with sorted trajectories, is not enough to improve the performance. It was partially effective with DT, slightly improving an already quite bad performance. It suggests that it may have potential but our results also imply that it is actually the quality of the datasets what actually brings the improvements.

TABLE 5. Evaluation of influence of dataset size (boldface highlights the best performance of each algorithm).

Method	Trajectory transformers			Decision transformers		
Dataset size	10k	100k	1M	10k	100k	1M
Episode Length	81.50 ± 2.55	96.03 ± 1.28	87.37 ± 2.31	74.92 ± 2.90	86.69 ± 2.23	71.85 ± 3.05
Euglycemia	52.14 ± 1.36	67.80 ± 0.91	68.27 ± 0.84	40.79 ± 1.73	48.68 ± 1.72	50.48 ± 1.45
Hyperglycemia	36.00 ± 1.59	24.29 ± 0.91	22.47 ± 0.77	54.34 ± 1.88	48.84 ± 1.82	38.75 ± 2.04
Hypoglycemia	11.84 ± 0.70	7.91 ± 0.56	9.26 ± 0.77	2.85 ± 0.40	2.47 ± 0.31	10.76 ± 1.42
Risk index	20.24 ± 1.27	9.75 ± 0.50	10.41 ± 0.60	31.71 ± 1.97	24.39 ± 1.67	24.08 ± 1.45

TABLE 6. Hyperparameter K tuning for Decision Transformer (boldface highlights the best performance of each algorithm).

Subject	K	EpLength	Euglycemia	Hyperglycemia	Hypoglycemia	RI
adolescent#001	Original (K=20)	100%	87.43±1.13	2.21±0.33	10.36±1.02	3.33±0.15
	10	100%	86.97±1.17	2.49±0.42	10.55±0.9	3.4±0.17
	50	100%	82.3±1.09	2.99±0.4	14.71±0.97	4.18±0.14
	100	100%	85.81±0.98	3.77±0.43	10.42±0.95	3.59±0.17
adolescent#002	Original (K=20)	92%	55.67±0.93	40.97±0.97	3.36±0.72	16.39±0.43
	10	100%	56.02±0.72	42.39±1.02	1.59±0.58	16.78±0.4
	50	95%	52.39±1.46	45.96±1.62	1.65±0.58	18.12±0.69
	100	100%	49.04±1.69	49.99±1.86	0.97±0.57	20.19±0.76

In fact, the importance of having good trajectories is obvious: if the dataset size is reduced but only the best trajectories are kept, the performance can be even improved. The average TIR in the 100k-sample dataset is at a value similar to the 1M-sample datasets. The episode length is increased because of sorting trajectories and keeping the best ones, which can be seen in the results obtained from combining datasets and dataset size reduction. However, offline RL algorithms can not learn from datasets when the size is down to 10k samples. We have found a good trade-off with a dataset size of 100k samples, which also agrees with the work of [4] and [20]. But we may further improve the results by filtering appropriately the datasets, that is keeping the best ones, and removing the trajectories with undesirable characteristics. For instance, removing the trajectories that result in high hypoglycemic and hyperglycemic fractions, even if they have a good accumulated reward. This can be done by shrinking the target TIR, for example, to be in the range of 90-100 mg/dL. Alternatively, we can redesign the reward function to punish more hypoglycemia and high hyperglycemia.

Although the offline RL with Transformers architecture does not outperform clearly the baselines, the main advantage of offline RL is that it does not require interaction with the environment, as compared to online RL, which needs to interact with the patient to collect data for training. Offline RL emerges therefore as a safer and promising alternative for RL, being a practical application of automated and customized glycemic control.

The next phase of research is further optimizing current methods to adapt the algorithms to learn how to better control blood sugar to normal levels and to make it more effective. However, a potential solution in which the patient or caregiver simply collects its own CGM data over time and converts it into a customized training dataset for offline RL still leaves multiple open questions. In particular, for the best performing model, we have used datasets generated from a simulated environment and from an optimal agent that was previously trained also on a simulated environment. But for real patients, to generate the dataset we would need to collect their CGMs and insulin doses, delivered according to the insulin regime the patients use, which is assumed to be not optimal in the first place. And, since exploration is not possible in offline RL [14], we can only expect marginal improvements over the patient actual insulin regime. A potential avenue, tested in this paper, is to generate the training datasets by mixing trajectories from different sources. For instance, from real patient data and an optimized agent from a simulated environment customized to the patient class. Our results with mixed datasets in this paper have not been satisfactory, so mixing deserves further attention.

Only when those issues and others have been clarified, we can expect to conduct clinical trials with healthcare professionals to collect datasets, and test and evaluate them in real patients.

VI. CONCLUSION

In this paper we have carried out a thorough evaluation of two recent offline RL algorithms for automated BG control of T1D patients. We have evaluated the influence on training and performance of the method that generated the datasets, as well as the influence of the type of trajectories used (single-method or mixed trajectories), the quality of the trajectories and the size of the datasets, and compared it with typically used baselines: PID and online RL methods.

Our results show that a Trajectory offline RL trained with a previous optimal PPO agent data performs at the level of the baselines, which supports that offline RL can be a realistic alternative for data-driven BG control. However, we have also shown several shortcomings of the tested methods, discussing potential avenues for improvement and next steps.

ACKNOWLEDGMENT

(Phuwadol Viroonluecha, Esteban Egea-Lopez, and Jose Santa contributed equally to this work.)

REFERENCES

[1] A. Singh. (Oct. 2022). *A Basal-Bolus Injection Regimen Involves Taking a Number of Injections Through the Day*. [Online]. Available: https://www.diabetes.co.uk/insulin/basal-bolus.html

[2] A. J. Barnes and R. W. Jones, "Pid-based glucose control using intraperitoneal insulin infusion: An in silico study," in *Proc. 14th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, 2019, pp. 1057–1062.

[3] C. Berget, S. Lange, L. Messer, and G. P. Forlenza, "A clinical review of the t: Slim x2 insulin pump," *Exp. Opinion Drug Del.*, vol. 17, no. 12, pp. 1675–1687, Dec. 2020.

[4] H. Emerson, M. Guy, and R. McConville, "Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes," *J. Biomed. Informat.*, vol. 142, Jun. 2023, Art. no. 104376.

[5] T. Biester, M. Tauschmann, A. Chobot, O. Kordonouri, T. Danne, T. Kapellen, and K. Dovc, "The automated pancreas: A review of technologies and clinical practice," *Diabetes, Obesity Metabolism*, vol. 24, no. S1, pp. 43–57, Nov. 2021.

[6] 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.

[7] A. Borri, G. Pola, P. Pepe, M. D. D. Benedetto, and P. Palumbo, "Symbolic control design of an artificial pancreas for type-2 diabetes," *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 5, pp. 2131–2146, Sep. 2022.

[8] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101836.

[9] S. Pareek, H. J. Nisar, and T. Kesavadas, "AR3n: A reinforcement learning-based assist-as-needed controller for robotic rehabilitation," *IEEE Robot. Autom. Mag.*, early access, Jun. 21, 2023, doi: 10.1109/MRA.2023.3282434.

[10] J. Lee and M. Mitici, "Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics," *Rel. Eng. Syst. Saf.*, vol. 230, Feb. 2023, Art. no. 108908.

[11] A. Namdari, M. A. Samani, and T. S. Durrani, "Lithium-ion battery prognostics through reinforcement learning based on entropy measures," *Algorithms*, vol. 15, no. 11, p. 393, Oct. 2022.

[12] P. Viroonluecha, E. Egea-Lopez, and J. Santa, "Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning," *PLoS ONE*, vol. 17, no. 9, Sep. 2022, Art. no. e0274608.

[13] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, "Deep reinforcement learning for closed-loop blood glucose control," in *Proc. Mach. Learn. Healthcare Conf.*, 2020, pp. 508–536.

[14] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.

[15] Y. Zhang, B. Tang, Q. Yang, D. An, H. Tang, C. Xi, X. Li, and F. Xiong, "Bcorle(λ): An offline reinforcement learning and evaluation framework for coupons allocation in e-commerce market," in *Advances in Neural Information Processing Systems*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 20410–20422.

[16] P. A. Apostolopoulos, Z. Wang, H. Wang, C. Zhou, K. Virochsiri, N. Zhou, and I. L. Markov, "Personalization for web-based services using offline reinforcement learning," 2021, *arXiv:2102.05612*.

[17] X. Chen, J.-Y. Jiang, K. Jin, Y. Zhou, M. Liu, P. J. Brantingham, and W. Wang, "Reliable: Offline reinforcement learning for tactical strategies in professional basketball games," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2022, pp. 3023–3032, doi: 10.1145/3511808.3557105.

[18] S. Tang and J. Wiens, "Model selection for offline reinforcement learning: Practical considerations for healthcare settings," in *Proc. 6th Mach. Learn. Healthcare Conf.*, vol. 149, K. Jung, S. Yeung, M. Sendak, M. Sjoding, and R. Ranganath, Eds. Aug. 2021, pp. 2–35.

[19] C. Shiranthika, K.-W. Chen, C.-Y. Wang, C.-Y. Yang, B. H. Sudantha, and W.-F. Li, "Supervised optimal chemotherapy regimen based on offline reinforcement learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4763–4772, Sep. 2022.

[20] I. Fox, "Machine learning for physiological time series: Representing and controlling blood glucose for diabetes management," Ph.D. dissertation, Division Comput. Sci. Eng., Univ. Michigan, Ann Arbor, MI, USA, Jul. 2020.

[21] P. Viroonluecha, E. Egea-Lopez, and J. Santa, "Offline reinforcement learning for type 1 diabetes," Jan. 2023. [Online]. Available: osf.io/gj783

[22] N. Resalat, J. E. Youssef, N. Tyler, J. Castle, and P. G. Jacobs, "A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model," *PLoS ONE*, vol. 14, no. 7, Jul. 2019, Art. no. e0217301.

[23] M. Asad, U. Qamar, and M. Abbas, "Blood glucose level prediction of diabetic type 1 patients using nonlinear autoregressive neural networks," *J. Healthcare Eng.*, vol. 2021, pp. 1–7, Feb. 2021.

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

[25] J. Xie. *Simglucose v0.2.1*. (2018). [Online]. Available: https://github.com/jxx123/simglucose

[26] S. Goel, S. Sharma, and R. Tripathi, "Predicting diabetes using CNN for various activation functions: A comparative study," in *Proc. 10th Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Moradabad, India, Dec. 2021, pp. 665–669.

[27] A. Huang, L. Leqi, Z. Lipton, and K. Azizzadenesheli, "Off-policy risk assessment in contextual bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23714–23726.

[28] T. Kirkgöz, M. Eltan, S. B. Kaygusuz, Z. Y. Abali, D. Helvacioglu, T. S. Menevse, B. G. Tosun, T. Güran, A. Bereket, and S. Demircioglu, "Efficacy of the novel degludec/aspart insulin co-formulation in children and adolescents with type 1 diabetes: A real-life experience with one year of IDegAsp therapy in poorly controlled and non-compliant patients," *J. Clin. Res. Pediatric Endocrinology*, vol. 14, no. 1, pp. 10–16, Mar. 2022.

[29] M. J. Schoelwer and M. D. DeBoer, "Artificial pancreas technology offers hope for childhood diabetes," *Current Nutrition Rep.*, vol. 10, no. 1, pp. 47–57, Jan. 2021.

[30] P. D. Ngo, S. Wei, A. Holubová, J. Muzik, and F. Godtliebsen, "Control of blood glucose for type-1 diabetes by using reinforcement learning with feedforward algorithm," *Comput. Math. Methods Med.*, vol. 2018, pp. 1–8, Dec. 2018.

[31] S. Cui, H. Tseng, J. Pakela, R. K. T. Haken, and I. E. Naqa, "Introduction to machine and deep learning for medical physicists," *Med. Phys.*, vol. 47, no. 5, pp. e127–e147, May 2020.

[32] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–14.

[33] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–14.

[34] C. Meister, T. Vieira, and R. Cotterell, "Best-first beam search," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 795–809, Dec. 2020.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[36] M. K. Bothe, L. Dickens, K. Reichel, A. Tellmann, B. Ellger, M. Westphal, and A. A. Faisal, "The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas," *Exp. Rev. Med. Devices*, vol. 10, no. 5, pp. 661–673, Sep. 2013.

[37] S. Trevitt, S. Simpson, and A. Wood, "Artificial pancreas device systems for the closed-loop control of type 1 diabetes," *J. Diabetes Sci. Technol.*, vol. 10, no. 3, pp. 714–723, May 2016.

[38] L. M. Huyett, E. Dassau, H. C. Zisser, and F. J. Doyle, "Design and evaluation of a robust PID controller for a fully implantable artificial pancreas," *Ind. Eng. Chem. Res.*, vol. 54, no. 42, pp. 10311–10321, Jun. 2015.

[39] C. C. Palerm, "Physiologic insulin delivery with insulin feedback: A control systems perspective," *Comput. Methods Programs Biomed.*, vol. 102, no. 2, pp. 130–137, May 2011.

[40] M. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[41] H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri, and O. P. Verma, "Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction," *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3073–3087, May 2021.

[42] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Apr. 2020.

[43] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review," *Diabetology Metabolic Syndrome*, vol. 14, no. 1, p. 196, Dec. 2022.

[44] 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.

[45] M. F. Rabby, Y. Tu, M. I. Hossen, I. Lee, A. S. Maida, and X. Hei, "Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, p. 101, Mar. 2021.

[46] W. Wang, M. Tong, and M. Yu, "Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization," *IEEE Access*, vol. 8, pp. 217908–217916, 2020.

[47] T. Zhu, L. Kuang, K. Li, J. Zeng, P. Herrero, and P. Georgiou, "Blood glucose prediction in type 1 diabetes using deep learning on the edge," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Daegu, South Korea, May 2021, pp. 1–5.

[48] X. Lu and R. Song, "A hybrid deep learning model for the blood glucose prediction," in *Proc. IEEE 11th Data Driven Control Learn. Syst. Conf. (DDCLS)*, Chengdu, China, Aug. 2022, pp. 1037–1043.

[49] 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, 2021, pp. 1–8.

[50] 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, Mar. 2021, Art. no. 105936.

[51] 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.

[52] S. Ahmad, A. Beneyto, I. Contreras, and J. Vehi, "Bolus insulin calculation without meal information. A reinforcement learning approach," *Artif. Intell. Med.*, vol. 134, Dec. 2022, Art. no. 102436.

[53] 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 Informat.*, vol. 25, no. 4, pp. 1223–1232, Apr. 2021.

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

[55] 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.

[56] 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.

[57] X. Yu, Y. Guan, L. Yan, S. Li, X. Fu, and J. Jiang, "ARLPE: A meta reinforcement learning framework for glucose regulation in type 1 diabetics," *Exp. Syst. Appl.*, vol. 228, Oct. 2023, Art. no. 120156.

[58] S. D. Giorno, F. D'Antoni, V. Piemonte, and M. Merone, "A new glycemic closed-loop control based on dyna-Q for type-1-diabetes," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104492.

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

[60] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 2623–2631.

[61] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–9.

[62] P. Viroonluecha, E. Egea-Lopez, and J. Santa. (Nov. 2022). *Virtual T1DM Blood Glucose*. [Online]. Available: https://osf.io/zurvk

[63] J. Xie. (Mar. 2022). *How did you Obtain the Parameters in Vpatient_Params.csv?* [Online]. Available: https://github.com/jxx123/simglucose/issues/26

[64] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," 2016, *arXiv:1606.01540*.

[65] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases," *Comput. Biol. Chem.*, vol. 97, Apr. 2022, Art. no. 107619.

[66] W. Clarke and B. Kovatchev, "Statistical tools to analyze continuous glucose monitor data," *Diabetes Technol. Therapeutics*, vol. 11, no. S1, pp. 45–54, Jun. 2009.

[67] S. G. Konan, E. Seraj, and M. Gombolay, "Contrastive decision transformers," in *Proc. 6th Annu. Conf. Robot Learn.*, Jun. 2022, pp. 1–11.

[68] L. Meng, M. Wen, Y. Yang, C. Le, X. Li, W. Zhang, Y. Wen, H. Zhang, J. Wang, and B. Xu, "Offline pre-trained multi-agent decision transformer: One big sequence model conquers all starcraftii tasks," 2021, *arXiv:2112.02845*.

[69] M. Xu, Y. Shen, S. Zhang, Y. Lu, D. Zhao, J. Tenenbaum, and C. Gan, "Prompting decision transformer for few-shot policy generalization," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 24631–24645.

[70] A. Kazemi, A. Abzaliev, N. Deng, R. Hou, D. Liang, S. A. Hale, V. Pérez-Rosas, and R. Mihalcea, "Adaptable claim rewriting with offline reinforcement learning for effective misinformation discovery," 2022, *arXiv:2210.07467*.

PHUWADOL VIROONLUECHA received the Ph.D. degree in information and communications technologies from the Technical University of Cartagena, Spain, in 2023. His current research interests include the application of machine learning to improve blood glucose control in type I diabetic patients, application of data science, artificial intelligence, and the Internet of Things (IoT) technologies in e-health and industry.

ESTEBAN EGEA-LOPEZ received the Telecommunications Engineering degree from the Polytechnic University of Valencia (UPV), Spain, in 2000, the master's degree in electronics from the University of Gävle, Sweden, in 2001, and the Ph.D. degree in telecommunications from Universidad Politécnica de Cartagena (UPCT), Spain, in 2006. He is currently an Associate Professor with the Department of Information Technologies and Communications, UPCT. His research interests include vehicular, ad-hoc and wireless sensor networks, and RFID.

JOSE SANTA received the M.S. degree in computer engineering from the University of Murcia, Spain, in 2004, the M.S. degree in advanced information and telematics technologies, in 2008, and the Ph.D. degree in computer engineering with European Mention from the University of Murcia, in 2009. A great part of his research work, both before and after his Ph.D., is about intelligent transportation systems, mobile communications, next-generation networks, cyber-physical systems, and the Internet of Things (IoT), with special emphasis on real prototypes and evaluation. He is currently a Senior Research Fellow (Ramn y Cajal) with the Department of Electronics, Computer Technology and Projects, Technical University of Cartagena.