Computers in Biology and Medicine 190 (2025) 110015

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A data-driven personalized approach to predict blood glucose levels in type-1 diabetes patients exercising in free-living conditions

Anas Neumann [a,b,1], Yessine Zghal [a,1], Marzia Angela Cremona [a,c,d], Adnene Hajji [a], Michael Morin [a,c], Monia Rekik [a,c,d,*]

[a] Université Laval - Department of Operations and Decision Systems, Faculty of Business Administration, Canada
[b] Polytechnique Montréal - Department of Mathematical and Industrial Engineering, Canada
[c] The Research Network on Cardiometabolic Health, Diabetes, and Obesity (CMDO), Canada
[d] University Hospital Center of Québec - Université Laval Research Center (CHUL), Canada

ARTICLE INFO

Keywords:
Type-1 diabetes
Blood glucose prediction
Artificial intelligence
Physical activity
Exercise
Free-living conditions

ABSTRACT

Objective: The development of new technologies has generated vast amount of data that can be analyzed to better understand and predict the glycemic behavior of people living with type 1 diabetes. This paper aims to assess whether a data-driven approach can accurately and safely predict blood glucose levels in patients with type 1 diabetes exercising in free-living conditions.

Methods: Multiple machine learning (XGBoost, Random Forest) and deep learning (LSTM, CNN-LSTM, Dual-encoder with Attention layer) regression models were considered. Each deep-learning model was implemented twice: first, as a personalized model trained solely on the target patient's data, and second, as a fine-tuned model of a population-based training model. The datasets used for training and testing the models were derived from the Type 1 Diabetes Exercise Initiative (T1DEXI). A total of 79 patients in T1DEXI met our inclusion criteria. Our models used various features related to continuous glucose monitoring, insulin pumps, carbohydrate intake, exercise (intensity and duration), and physical activity-related information (steps and heart rate). This data was available for four weeks for each of the 79 included patients. Three prediction horizons (10, 20, and 30 min) were tested and analyzed.

Results: For each patient, there always exists either a machine learning or a deep learning model that conveniently predicts BGLs for up to 30 min. The best performing model differs from one patient to another. When considering the best performing model for each patient, the median and the mean Root Mean Squared Error (RMSE) values (across the 79 patients) for predictions made 10 min ahead were 6.99 mg/dL and 7.46 mg/dL, respectively. For predictions made 30 min ahead, the median and mean RMSE values were 16.85 mg/dL and 17.74 mg/dL, respectively. The majority of the predictions output by the best model of each patient fell within the clinically safe zones A and B of the Clarke Error Grid (CEG), with almost no predictions falling into the unsafe zone E. The most challenging patient to predict 30 min ahead achieved an RMSE value of 32.31 mg/dL (with the corresponding best performing model). The best-predicted patient had an RMSE value of 10.48 mg/dL. Predicting blood glucose levels was more difficult during and after exercise, resulting in higher RMSE values on average. Prediction errors during and after physical activity (two hours and four hours after) generally remained within the clinical safe zones of the CEG with less than 0.5% of predictions falling into the harmful zones D and E, regardless of the exercise category.

Conclusions: Data-driven approaches can accurately predict blood glucose levels in type 1 diabetes patients exercising in free-living conditions. The best-performing model varies across patients. Approaches in which a population-based model is initially trained and then fine-tuned for each individual patient generally achieve the best performance for the majority of patients. Some patients remain challenging to predict with no straightforward explanation of why a patient is more challenging to predict than another.

* Correspondence to: Faculty of Business Administration, Department of Operations and Decision Systems, Université Laval, Office 2431, 2325 De la Terrasse St, Québec G1V 0A6, Canada.

E-mail addresses: anas.neumann@polymtl.ca (A. Neumann), yessine.zghal.1@ulaval.ca (Y. Zghal), marzia.cremona@fsa.ulaval.ca (M.A. Cremona), adnene.hajji@fsa.ulaval.ca (A. Hajji), michael.morin@fsa.ulaval.ca (M. Morin), monia.rekik@fsa.ulaval.ca (M. Rekik).

[1] Co-first authors: these authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2025.110015
Received 26 April 2024; Received in revised form 16 January 2025; Accepted 7 March 2025
Available online 31 March 2025
0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

Type-1 Diabetes (T1D)[2] is a chronic autoimmune condition that attacks and destroys the pancreatic β cells responsible for producing insulin [1,2]. As a result, People With Type-1 Diabetes (PWT1D) require continuous insulin injections to maintain glucose homeostasis and prevent hyperglycemia (Blood Glucose Level (BGL) > 180 mg/dL). Prolonged high blood glucose levels can lead to severe health complications including heart diseases, stroke, blindness, and kidney damage. PWT1D frequently experience episodes of hypoglycemia (BGL < 70 mg/dL), whether they manage their condition with Continuous Subcutaneous Insulin Infusion (CSII) or Multiple Daily Injections (MDI) [1]. Hypoglycemia, especially Nocturnal Hypoglycemia (NH), is a common effect of Physical Activity (PA) [3,4]. These episodes significantly impact quality of life and can result in serious physical complications [2,5,6]. Effective T1D management depends on a comprehensive understanding of each individual's glycemic behavior, which is influenced by factors such as PA, diet, stress, and lifestyle habits [2,7].

The development of new technologies such as insulin pumps, Continuous Glucose Monitor (CGM), and health monitoring applications, has generated vast amount of data that can be analyzed to better understand and predict the glycemic behavior of PWT1D. Consequently, data-driven methods are becoming the most used approaches for T1D management [8–13]. While regression models have been developed to predict future BGL and classification models to create alarm systems for anticipating episodes of hypoglycemia and hyperglycemia [14–17], significant challenges remain. Current blood glucose prediction models often lack accuracy during dynamic variations in BGL, which limits their practical use. This is especially the case when PA must be accounted for. To date, no universal method exists to measure the impact of PA on future blood glucose levels and their variability [18,19]. Recent studies, such as those by [7,20], and [4], have also highlighted inconsistencies with earlier experiments that primarily emphasized the effects of energy expenditure and intensity. Despite recent advancements in machine learning and deep learning, many researchers argue that existing methods remain inadequate for addressing the high inter- and intra-patient variability, particularly in Real World (free-living) Conditions (RWC) [21,22].

In this paper, we propose novel personalized data-driven approaches to predict future BGL of PWT1D in RWC. Our main goal is to anticipate exercise-induced hypoglycemia and hyperglycemia episodes, considering the patient's habits and living environment. We adapted and tested different data-driven approaches based on either Machine Learning (ML) – especially, XGBoost and Random Forest – or Deep Learning (DL) – LSTM, CNN-LSTM, and dual-encoder with attention layer – regression models. We implemented each DL model twice: firstly by training the model only on the data of the target patient and secondly by training the model on the data of all patients and then fine-tuning it to the target patient. Hyperparameter tuning was also performed to dynamically adapt the architecture and configuration of the models to each specific patient. To evaluate the performance of each model for each patient, we computed two metrics: the RMSE and the percentage of Predictions in clinically critical zones D and E (PDE). The PDE leverages the Clarke Error Grid (CEG), a graphical tool used in diabetes management to assess the clinical accuracy of blood glucose prediction systems by categorizing predictions into specific zones based on their potential impact on clinical decision-making [23,24].

We tested the proposed approaches on 79 real patients selected from the T1DEXI study [25], an observational, at-home research initiative gathering diverse data on structured and free-living exercise (aerobic, resistance, and interval training) for T1D adult patients. These data were collected over four weeks. The 79 selected patients used CSII pumps and Dexcom G6 CGM sensors. We designed our models specifically to exploit the data provided by the T1DEXI dataset as input features: a sequence of previous BGLs, the recent quantities of insulin injected, the Carbohydrate (CHO) intake, and the PA-related information (intensity, duration, steps, and heart rate).

Our results show that, while some fine-tuned deep learning models – particularly those based on LSTM and dual-encoder architectures – outperformed the other models for most patients, there is no straightforward explanation for the matching between specific patients and their best-performing model. The best hyperparameter configuration and architectural settings also vary from one patient to another. Our results confirm the benefits of adopting a personalized approach (model selection and configuration), tailored to the specificity of each patient. They also demonstrate that, for every patient, at least one algorithm provided sufficiently accurate predictions. Notably, the majority of the predictions made 30 min ahead fell in the clinically safe zones A and B of the CEG and almost no predictions fell into the critical zone E. The median and average RMSE achieved for predictions made 10 min ahead across the 79 patients were 6.99 mg/dL (0.39 mmol/L) and 7.46 mg/dL (0.41 mmol/L), respectively. For predictions made 30 min ahead the median and mean RMSE were 16.85 mg/dL (0.94 mmol/L) and 17.74 mg/dL (<1 mmol/L), respectively. Predicting BGLs proved to be more challenging during and after exercise. The average RMSE for predictions made 30 min ahead during aerobic, resistance, and interval training was 21.30 mg/dL, 16.45 mg/dL, and 18.46 mg/dL, respectively. These averages were 20.13 mg/dL, 19.49 mg/dL, and 20.56 mg/dL, respectively, during the two hours after an exercise and 19.41 mg/dL, 18.88 mg/dL, and 19.84 mg/dL, respectively, during the four-hours post exercise. Despite these higher prediction errors during and after PA, most errors occurred within the clinical safe zones of the CEG. Indeed, across all PA categories, the percentage of predictions falling into harmful zones D and E remained below 0.5% for all patients.

The remainder of this paper is organized as follows. Section 2 provides a review of recent scientific literature on data-driven algorithms designed for predicting future glucose levels in PWT1D, with a particular focus on algorithms trained and tested on data collected on patients exercising in free-living conditions. Section 3 describes the dataset used as well as our prediction methodology. Section 4 presents the experimental results evaluating the performance of our prediction models. In Section 5 we discuss the practical applicability of the proposed approaches and their limitations. Finally, Section 6 provides a summary of the findings and suggests potential avenues for future research.

2. Literature review

In the past few years, data-driven approaches have been intensively considered for T1D management. Regression models have been proposed to predict future BGL, while classification models have been employed as alarm systems to avoid hypoglycemia and hyperglycemia episodes [3,8,13,15,16,26,27]. The most common Prediction Horizon (PH) corresponds to 30 min [8,26], although predictions can sometimes reach up to two hours [2,18,20]. Most models are designed to predict a specific PH, while others can predict multiple horizons simultaneously [13,28–30]. Our literature review is divided into three parts. In the first part, we present the different designs of data-driven methods (including model architectures and the information used for predictions) recently proposed in the literature for T1D management. The second part focuses on the recent works that address PA and T1D. The third part discusses the training process and the prediction quality reported in the recent literature.

2.1. Data-driven methods: designs, architectures, and information used

A wide range of classical ML algorithms have been recently used in the context of T1D management: XGBoost [9,10,31], ARIMA with eXogenous input (ARIMAX) [11–13], Decision Trees (DTs), Naive Bayes (NB),

[2] A full list of acronyms is available at the end of the paper.

K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) (as reported in [8]). However, according to a recent systematic literature review [18], Artificial Neural Network (ANN) and DL models are the most used family of algorithms, being reported in 45% of the studied papers. They are also among the best-performing approaches according to several comparative studies [8,28]. Some ANN models are simple Feed-Forward Multi-Layered Perceptron (FF-MLP) considering the Softmax function when employed for classification [8,30,32]. However, due to the time-series nature of the input features (heart rate, BGL over time, etc.), several DL models are Recurrent Neural Network (RNN) and use LSTM cells [16,26,27,33,34] and Gated Recurrent Unit (GRU) [6,26,27,35]. Some papers demonstrate that feature extraction models like Temporal Convolutional Network (TCN) outperform RNN and LSTM models [19,36]. Other models also use encoder–decoder architectures with attention layers such as the well-known Transformer [6,17,35,37].

Both alarm systems and prediction models use different data types as input. Previous BGL collected by CGMs is the main exploited data [18]. The number of BGL points used for a prediction, or the length of the BGL sequence, depends on the capability of the proposed model. The second most used data is the quantity of insulin injected. Most datasets and papers distinguish between fast-acting (5 min to four hours) and slow-acting (up to 24 h) insulin [29,38]. The third most frequently used data is related to PA. Several studies demonstrate the impact of PA on blood glucose concentration and variability. They emphasize the importance of considering PA-related features for BGL prediction, especially for patients who exercise frequently [39–41]. Data related to PA is mainly extracted from smartwatches that collect several metrics like heart rate (bpm), skin temperature, number of steps, or moving speed. Some papers also differentiate between aerobic exercises and High-Intensity Interval Training (HIIT), and between trained and untrained patients [22]. The fourth most exploited type of data is CHO intakes, usually measured in grams. Some papers also distinguish between dietary nutrients and energy sources: monosaccharides, starches, lipids, proteins, or fibers [29]. Finally, fewer models consider stress-related data [20].

The use of multivariate inputs yields some technical challenges. First, data are expressed in different measurement units and magnitudes and usually need a standardization step. They are also collected at different frequencies. For instance, while CGM and heartbeats are continuously collected every five minutes, other PA data (exercise type, intensity, duration, etc.) and CHO intakes are collected punctually as discrete events with an exact timestamp. Besides, even for data extracted at the same frequency, the different sensors are often not synchronized. In this case, a resampling step is mandatory [3,6,19]. Moreover, insulin doses, PA, and CHO intakes do not instantly affect BGLs. For example, the slow insulin impacts BGL for one to 24 h, but the effect of meals lasts for three to four hours [34]. This motivates the use of three absorption curves, namely the Insulin-On-Board (IOB), the Carbohydrates-On-Board (COB), and the Exercise-On-Board (sometimes Activity-On-Board) (EOB) [3,33]. The model proposed by [34] uses an autonomous channel network to consider the time-dependent scale discrepancy, incompletion, and redundancy of the different time-series variables. Each channel has a multi-lag structure to extract features, has an adapted length, and is used for a different type of input. The authors also use IOB and COB physiological models from [42,43] as pre-processing steps.

Data-driven models can be either population-based or personalized. Population-based models are trained on the data of several PWT1D. They are intended to be robust and perform well for different patients [2,44]. Yet, they sometimes fail to adapt to individuals with new blood glucose patterns. Their performance can be improved by using patient clustering techniques [2,45]. Personalized models are trained on the data of a single patient. They enable circumventing the inter-patient variability [13,14,16]. Although they may perform well for the studied patient, they risk overfitting [2]. Besides, obtaining enough RWC data to train a personalized model is difficult in practice: CGMs typically collect 288 data points per day. Hence, according to Langarica et al. [46], personalized models are harder to obtain and less frequently considered in the literature. Yet, several personalized approaches such as Muñoz-Organero et al. [33], Faccioli et al. [13], and Iacono et al. [16] have been proposed. To reduce the quantity of data needed, pre-trained population-based models are often used to perform transfer learning and fine-tuning [22].

2.2. PA and T1D management

Several recent works focus on PA and the effect of exercising on BGLs, during and after exercise sessions. De Paoli et al. [47] studied six real adults exercising in aerobic (paddle/bicycle, belly dance, and eight-a-side football) and anaerobic (gym, sailing, and home workouts) sessions. They observed that aerobic sessions mostly caused hypoglycemia, while anaerobic exercises yielded hyperglycemia during exercise. Regardless of the tested approaches, BGL was harder to predict (larger RMSE) during aerobic sessions than during anaerobic ones. Besides, the predictions made during aerobic sessions were worse than during rest periods. Bertachi et al. [3] reported that more than 50% of severe hypoglycemia, especially for patients using MDI, occurs during the night. A comparative study carried out by [4] on data collected over 14 days demonstrated that exercise did not impact the Time Above Target (TAR) metrics while exercising, but induced large Time Below Target (TBR) values (hypoglycemia) later during the night. This result was even more conclusive for the group of patients doing high-intensity workouts. Parcerisas et al. [48] experimented on virtual patients under MDI to analyze the importance of considering CHO intakes and PA data when predicting nocturnal hypoglycemia. Their prediction approach, composed of both population-based and personalized models, was able to reduce the number of nocturnal hypoglycemia. Bergford et al. [7] designed a Repeated Measures Random Forest (RMRF) to predict hypoglycemia risk during exercise. They trained their model on 459 patients in free-living conditions for a total of 8827 exercise sessions. Repeated measures enable integrating within-subject (like sequential data) and non-linear correlations into classical ML models like RF and DT. This technique was used as a component for explainability, helping identifying the predictors that most impacted the outcomes. They demonstrated the impact of the BGL trend (direction and growth/decay rate) on future BGLs. Their results emphasize the importance of maintaining high BGL before exercise to avoid hypoglycemia episodes. They highlight a higher risk of hypoglycemia for free-living exercise (hiking, walking, physical labor, etc.) than structured exercise (workout, aerobic, etc.), regardless of their intensity, and a slightly higher risk when exercises are performed during the night.

2.3. Training process and BGL prediction quality

In this section, we focus on regression models as our objective in this paper is to predict BGLs. Such models are evaluated using common statistical metrics like RMSE or Mean Absolute Error (MAE). However, several papers also use clinical tools and metrics like the IndexJ [18,49] or the CEG [13,27,34,49,50]. The CEG [23,24] is divided into five zones of errors (A to E) based on their level of risk. Zones A and B are considered clinically safe. Zones D and E are considered very risky. For example, zone E contains predictions of hypoglycemia episodes that are, in reality, hyperglycemia episodes and predictions of hyperglycemia events when the patient was, in fact, in hypoglycemia. Del Favero et al. [51] proposed a new loss function, called glucose-specific Mean Square Error (gMSE), based on the CEG. The gMSE loss function applies penalties for overestimating the BGL during hypoglycemia (with a factor up to 2.5) and underestimating the BGL during hyperglycemia (with a factor up to 2). It has been used in recent papers like [13] or [35] in its gRMSE form. A similar loss function with penalties on harmful prediction errors can be found

in [44]. Using such a function usually deteriorates the (R)MSE but improves CEG-related metrics (percentage in each zone) [44].

The type and quantity (number of patients and sampling duration) of training and validation data also vary between papers. Some researchers use simulators of virtual patients like UVA/Padova [16,26,46], and in-silico datasets such as DirecNet [30,52] and OhioT1DM [6,19,31,34,42,53]. Others test their models on real patients in free-living conditions [7,18]. Zecchin et al. [32] trained an FF-MLP on seven days of data collected over 15 patients and obtained a mean RMSE of 14.0 ± 4.1 mg/dL. Alfian et al. [52] tested an XGBoost model to predict the BGL of five children using six days of data extracted from the DirecNet dataset. The average RMSE reaches 23.219 mg/dL for a 30-min prediction horizon. The same authors later reported an average RMSE of 6.31 ± 2.43 mg/dL over 12 patients. Annuzzi et al. [30] proposed a population-based Multi-output FF-MLP trained on two 7-day datasets of real patients: a dataset of 12 children and a dataset of 15 adults. Their model resulted in an average RMSE of 8.30 ± 2.37 mg/dL. Zarkogianni et al. [54] trained a self-organizing map on the data of 10 patients and achieved an average RMSE of 11.42 mg/dL. They collected 10.70 ± 4.69 days of data depending on the patients. Some analyses have also been done on larger datasets. For instance, Prendin et al. [12] trained an ARIMA on 10 days of CGM data from 124 patients and reached a mean RMSE of 22.15 mg/dL. The population models employed by [44] were trained on 158 patients and tested on the remaining patients of the REPLACE-BG dataset [55]. Using a bagging method, they reached a mean RMSE of 20.31 mg/dL for predictions made 30 min ahead. They also achieved low percentages of predictions falling in zones D (0.01%) and E (0%) of the Parker error grid (a specific version of the CEG). Cichosz et al. [56] combined the REPLACE-BG dataset with the data of 199 other patients from another study [57] and improved the prediction algorithm by adding a screening process. The screening process enables an early detection (first week) of patients with unstable BGLs, which are hard to predict with data-driven algorithms.

As can be seen, the accuracy achieved by similar approaches varies greatly when trained and applied to different patients or datasets. As phrased by [22]: "*although algorithms that have been designed to predict future glucose exhibit relatively low RMSE during non-exercise periods, recent studies have indicated that the accuracy of these algorithms is frequently far worse during exercise (46.16 mg/dL) [21]*". While several models exploit some PA-related information as input features (heart rate and number of steps), only a few datasets and articles explicitly include patients performing exercise sessions. The following results focus on approaches dedicated to such datasets and studies. Exercise-oriented datasets include information such as exercise intensity, duration, and Energy Expenditure (EE). Using DL models, Tyler et al. [22] reached a mean RMSE of 22.1 ± 2.35 mg/dL globally. The OhioT1DM dataset contains eight weeks of data from 12 patients and includes exercise information [58] – the first version has six patients [59]. Applying an XGBoost model to the OhioT1DM dataset, Midroni et al. [31] achieved a mean RMSE of 16.2 mg/dL for a 30-min prediction horizon. Bertachi et al. [42] obtained a mean RMSE of 19.33 mg/dL with a DL model. The tests conducted by [53] reached a mean RMSE of 19.53 mg/dL using an ARX model. The best-performing LSTM models proposed by [19] achieved a mean RMSE of 19.48 mg/dL for direct predictions made 30 min ahead and 19.52 mg/dL for the recursive ones. Li et al. [60] designed a CNN-LSTM model to predict the BGL of 10 patients, exploiting six months of data for each patient, and reached a mean RMSE of 21.07 ± 2.35 mg/dL. De Paoli et al. [47] experimented on six adult patients with data varying between 6 and 81 days. They designed a personalized DL model, called Jump Neural Network, with residual connections between the inputs and its last layer. They also tested three training processes: offline (training the model only once with the first 24 h of data), online (retraining the model every five minutes with the data of the last 24 h), and using reinforcement learning (applying a custom loss function with penalties). Their best-performing model reached an average RMSE of 24.5 mg/dL.

Despite the number of contributions made in the field, data-driven T1D management is still in its infancy and many gaps still need to be filled, out of which two are of particular interest to us: (1) to date, there is no universal method to take into account the impact of PA and exercising on future blood glucose levels and variability [18,19]; (2) results from recent papers like [7,20], and [4] contradict previous experiments that highlighted only the effects of energy expenditure and intensity on BGL; several authors consider current methods not sufficient to handle the high inter- and intra-patient variability in RWC [21,22]. In what follows, we contribute to addressing the first gap by considering a dataset with explicit PA data. Then, we tackle inter-patient variability (second gap) by developing a model training methodology on a per-patient basis (including pre-training on the patient population).

3. Proposed data-driven approaches

This section presents the proposed prediction models and the dataset used to train and test them. As already mentioned, we intend to build a personalized approach to predict BGL of patients living and exercising in RWC. We considered different ML and DL regression models. This offers a certain degree of diversity in the type of data-driven models tested and increases the chance of finding a good prediction model for each patient. All proposed models predict the BGL over six different time horizons simultaneously: 5, 10, 15, 20, 25, and 30 min. We also dynamically adapt the technical architecture and configuration of the models, such as the layers, neurons, loss function, and use of memory (sequential data inputs). As will be described in the following, our methodology can be divided into two stages. The first stage consists of a data pre-processing and feature engineering pipeline (Section 3.2). The second stage, called the machine and deep learning pipeline, consists of designing and training our prediction models and tuning their hyper-parameters (Section 3.3). Before explaining the two stages in detail, we start by describing the dataset used.

3.1. Dataset description

Our dataset is obtained from T1DEXI [25], a real-world study of at-home exercise in which 497 adults diagnosed with type 1 diabetes (for at least two years) have participated. Participants were randomly assigned to at least six structured exercise sessions over four weeks. They were asked to provide different information related to their socio-demographic status, health, and lifestyle habits, as well as more specific data on the daily management of their diabetes and their PA. The information collected allowed us to have a rich set of data to train our AI algorithms, namely CGM, pumps, carbohydrate, heart rate, exercise, and step data for each participant. For more details on the T1DEXI study, we refer the reader to the full protocol available on the official website of the Jaeb Center for Health Research [25].

For our research, we only considered the participants using an insulin pump (either standard or closed loop), who achieved a minimum of 150 min of exercise per week and who reported their carbohydrate intake. These inclusion criteria were established to access detailed and accurate data regarding the amount of insulin injected and carbohydrate intake. Pump reports provide basal insulin amounts every five minutes, while bolus insulin amounts and carbohydrates consumed are typically recorded automatically in the pump reports. Additionally, we aimed to test data-driven approaches on active patients engaged in a minimum level of PA, as PA is one of the most influential factors affecting variations in blood glucose levels, making BGL predictions more challenging. A total of 79 patients, consisting of 59 females and 20 males, met these criteria and were thus used to train and test our AI models. Table 1 summarizes the characteristics of the 79 participants (measured in number of patients): sex, exercise type, age group, number of years since T1D diagnosis, Body Mass Index (BMI), HbA1C range, and the number of severe hypoglycemia requiring

Table 1
Characteristics of the 79 selected participants.

Sex					
Male			Female		
20			59		
Exercise					
Interval exercise		Aerobic		Resistance training	
31		25		23	
Age group					
18–25 years		26–44 years		≥45 years	
23		33		23	
Years since T1D diagnosis					
≤4 years		5–9 years		≥10 years	
10		13		56	
Body Mass Index					
<25 kg/m^2		25–30 kg/m^2		>30 kg/m^2	
39		27		13	
HbA1c range					
<6%	6-6.5%	6.5%–7%	7–7.5%	7.5%–8%	>8%
13	28	13	15	5	5
Severe hypoglycemia					
0	1	2	3–4	5–10	>10
49	9	5	10	4	2

Table 2
BGL variation of the 79 selected participants.

	Min	Q1	Median	Q3	Max
TIR	7.55%	66.35%	77.29%	85.24%	99.05%
TBR	0.0%	1.15%	2.24%	4.28%	20.31%
TAR	0.0%	10.71%	19.51%	30.05%	92.45%
CV	16.53%	29.12%	32.92%	36.77%	57.16%

assistance experienced during the study. Table 2 provides additional information related to BGL variations during the four weeks of the study. As data are generally not normally distributed, we report the minimum, maximum, median, first and third quartile values for the Time In Target (TIR), Time Above Target (TAR), Time Below Target (TBR), and the coefficient of variation (CV) of the CGM values. To allow for a better evaluation of the generalizability of our findings, Tables 11 and 12 (Annex E) compare the demographics, medication (insulin administration), diabetes duration, HbA1c levels, TIR, TAR, TBR, and CV between the 79 patients who met our inclusion criteria and all the 497 patients of the T1DEXI study. As can be noticed, the 79 patients who met our inclusion criteria have characteristics similar to all T1DEXI patients except for PA duration and insulin mode of administration: the selected patients are more active and use insulin pumps.

Detailed characteristics related to hypoglycemia duration, PA, and CHO intake are displayed in Figs. 1 to 4, which provide aggregate information over the four weeks for each patient. Fig. 1 illustrates the total duration (in minutes) of hypoglycemia, both severe and non-severe, recorded for each patient (the x-axis) over the four weeks. To draw Fig. 1, we computed the total time each patient's CGM value was under 70 mg/dL. Our analysis reveals that the average duration in hypoglycemia (over the 79 selected participants) is 1261 min (21 h), with a standard deviation of 1017 min (17 h), indicating significant variability in the amount of time spent in hypoglycemia between patients. Specifically, the time spent in hypoglycemia ranges from less than 2 h (80 min) for Patient 1100 to over 80 h for Patient 1427.

Regarding exercise sessions, 31 participants engaged in interval training, 25 in aerobic activities, and 23 in resistance exercises. Fig. 2 displays the total duration (in minutes) of exercise sessions (across all types) reported for each of the 79 participants. The average exercise duration is 1319 min (22 h), with a standard deviation of 726 min (12 h), indicating a considerable variation in exercise participation among the 79 patients. The total exercise duration ranges from 520 min (almost 9 h) for Patient 1261 to 5245 min (nearly 87 h) for Patient 1155.

Fig. 3 shows the total number of steps (over the four weeks) for the 31 participants (among the 79 included) for whom step data was available. On average, participants took 235,881 steps, with a standard deviation of 93,237 steps. The total number of steps ranged from 83,569 to 492,523, highlighting a significant variability in PA habits between patients.

Fig. 4 illustrates the total carbohydrate intake (in grams) over the four weeks, as reported in T1DEXI for each of the 79 participants of our dataset. The average carbohydrate intake is 3756 g, with a standard deviation of 1535 g. Reported carbohydrate intake ranges from 680 g to 9950 g.

3.2. Data processing pipeline and feature engineering

3.2.1. Data pre-processing

To train our AI algorithms, it is essential to build a comprehensive dataset associated with the different preselected features considered relevant for BGL prediction. In our case, the feature values were available in T1DEXI in different formats (different datasets). They were collected at different frequencies and used distinct units of measure or magnitude. In the following, we describe the different steps that were required to combine and integrate these features.

CGM sensors generally give CGM values every five minutes. In the T1DEXI datasets, a CGM value was available for each patient at least every five minutes. In some cases, different CGM values were reported within the same five-minute interval, at different times. Resampling the signal to five-minute intervals was thus needed. A total of 8064 CGM observations (associated with the four weeks of the study) were thus considered for each patient.

Exercise data included event entries for each exercise session of each participant. As already mentioned, the 79 patients we considered exercised a minimum of 150 min a week. Three types of information were considered for each exercise event: its intensity, its duration, and whether it was competitive or not. To consolidate the CGM and the exercise datasets, we had to align each exercise event timestamp to the next available CGM timestamp. This was required to handle the potential discordance in timestamps between the two datasets, such as having an exercise event starting at 03:03[3] whereas CGM observations are every five minutes starting at 00:00. Maintaining temporal alignment is crucial to avoid data leakage and inaccuracies. For example, if a patient initiated an exercise session at 12:07:25, we recorded this event at 12:10:00.

Integrating CGM and exercise data resulted in missing values for the periods of time when a patient was not exercising. To fill in these missing values, we proceeded as follows. A value of −1 was considered for the timestamps without exercise. For the timestamps covering an exercise session, we populated the intensity and competitiveness columns uniformly with the same values that initiated the exercise event. For the exercise duration, we indicated for each timestamp t the total duration of exercise that had been performed from the beginning of the exercise session until timestamp t. This value progressively increases as we move forward until the exercise session is terminated. For example, if a patient started a 30-min exercise session at 12:00:00, the exercise duration value is five minutes at 12:05:00, 10 min at 12:10:00, . . . , and 30 min at 12:30:00.

Regarding the step data, we computed the sum of the number of steps associated with each five-minute time interval and reported the value for each timestamp in the combined exercise and CGM dataset.

[3] We employ the 24-h clock time notation, HH:MM:SS where HH represents the hour, MM represents the minutes, and SS represents the seconds. We drop the hour when it is not necessary, e.g., 03:03 means 3 min and 3 s.

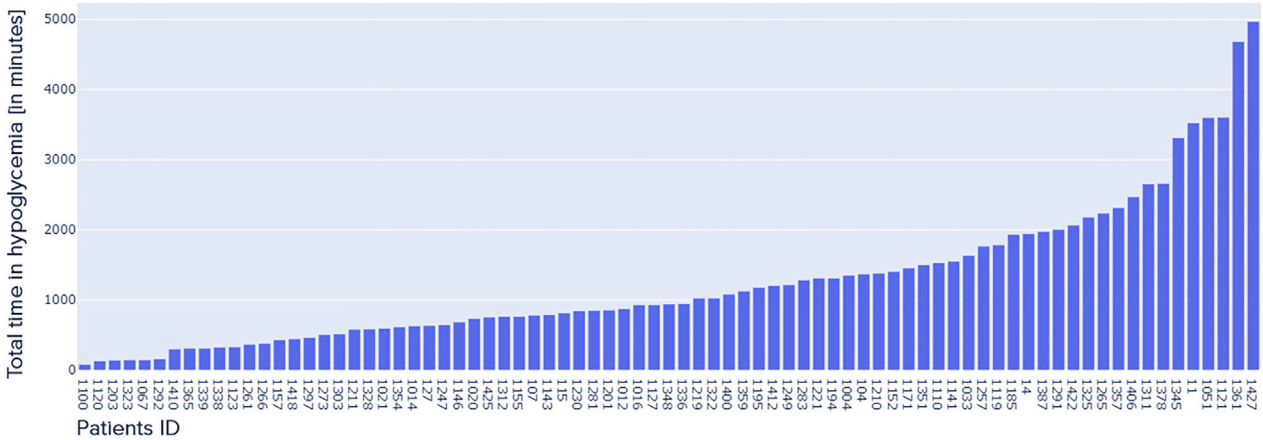


Fig. 1. Total time in hypoglycemia (in minutes).

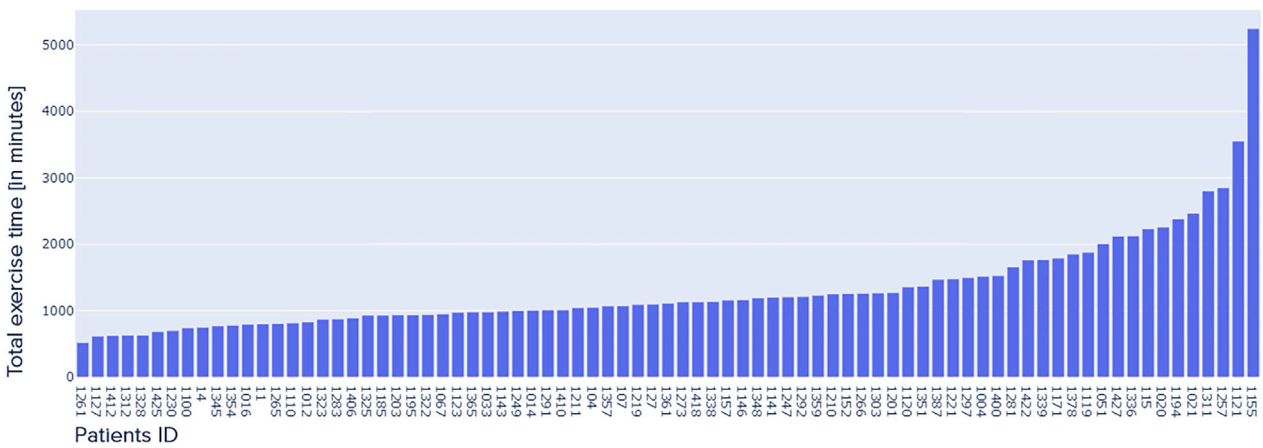


Fig. 2. Total exercise time (in minutes).

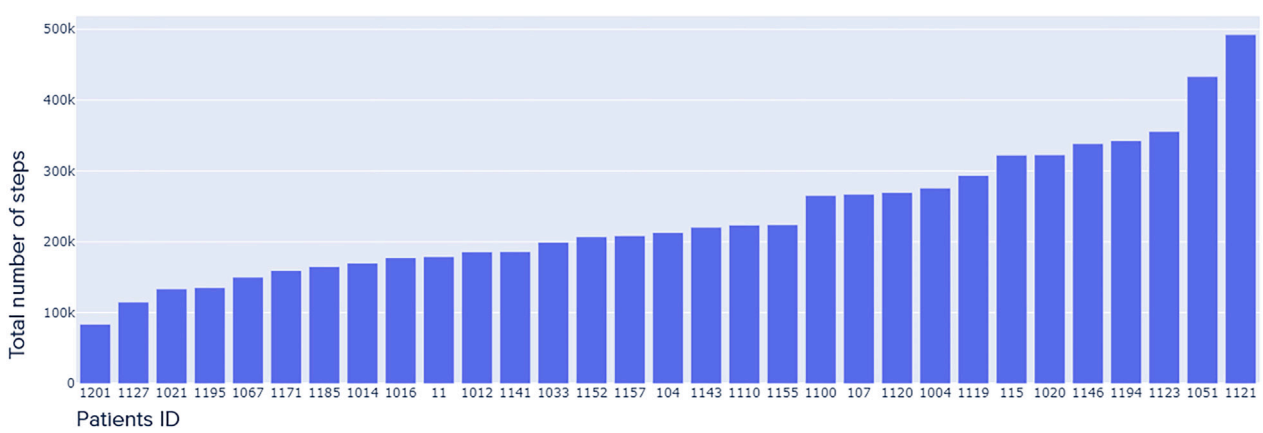


Fig. 3. Total number of steps.

The heart rate data was collected every 10 s and had to be re-sampled to a five-minute interval. To mitigate the information loss, we represented each five-minute interval of heart rate data using the average, standard deviation, and skewness statistics. After that, we merged this data with the combined exercise-CGM dataset.

Our dataset included only patients using insulin pumps. Two types of insulin injections are available in a pump: the basal and the bolus. A basal insulin is continuously injected by the pump in an automated way following a pre-specified basal profile already entered into the pump by the user. A bolus of insulin is injected punctually and manually by the patient when CHOs are consumed, or a hyperglycemia episode needs to be managed. In the datasets made available by T1DEXI, total basal insulin doses were reported for each patient over different time periods that generally did not coincide with the CGM timestamps. A data processing step was thus necessary to compute the basal injection rate per second and then determine the total injected basal insulin over the five minutes associated with each timestamp. For bolus insulin data, total injected doses were reported, as well as the time at which they were administered. To integrate this data into our combined dataset, we proceeded as follows. If no bolus insulin was injected at any

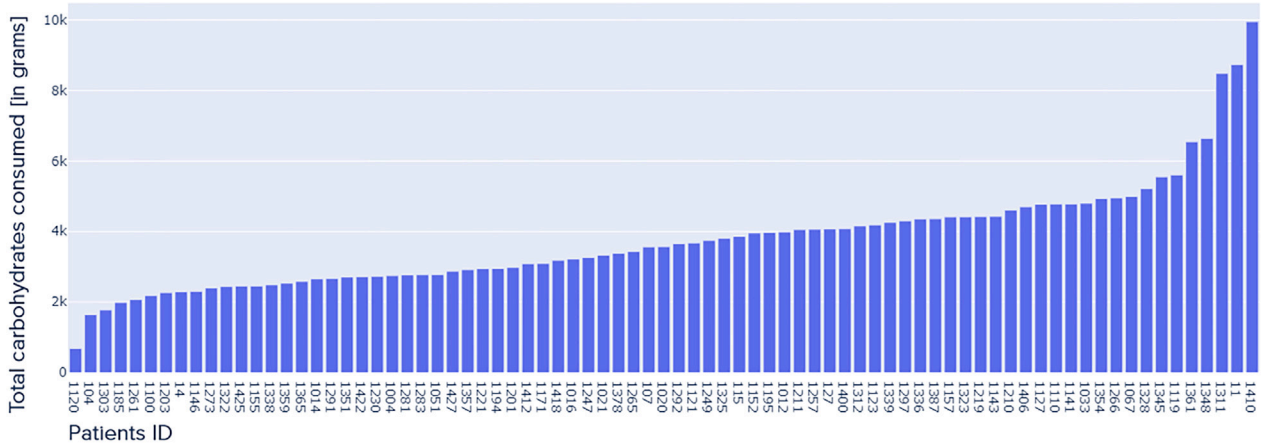


Fig. 4. Total carbohydrate intake (in grams).

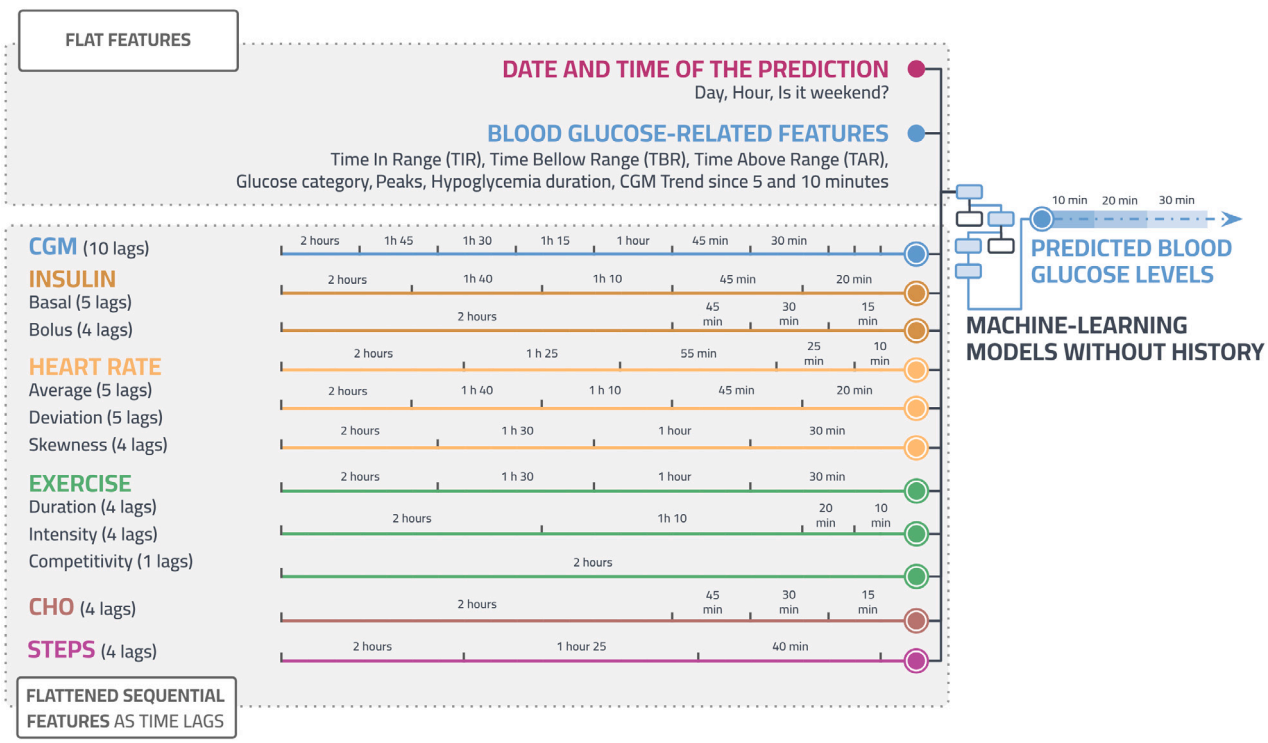


Fig. 5. Input features of learning models without history.

period covered by a CGM timestamp, a value of zero was considered for this timestamp. Otherwise, we considered the bolus insulin dose or the sum of bolus insulin doses that had been injected during the timestamp.

Carbohydrate intake information was provided by two data sources. First, a patient could enter the total amount of carbohydrates consumed (in grams) through an application provided in the T1DEXI study. Second, the patient could take a picture of the meal/snack she/he consumed. These pictures were analyzed by an AI algorithm to estimate the carbohydrate amount consumed. To integrate this information into our combined dataset, we needed a unique CHO value for each timestamp. We proceeded as follows. If, within a given timestamp, both data sources provided carbohydrate values, we kept only the value entered by the patient. If, for a given timestamp, a unique value was reported, we considered it regardless of its origin. For the remaining timestamps where no carbohydrate values were present, we set the value to zero.

3.2.2. *Feature engineering*

In our data-driven predictive modeling approach, the emphasis on feature engineering played a critical role in bridging the gap between the inherent complexity of raw data and the effective utilization of machine learning and deep learning algorithms. To circumvent the difficulties that may be potentially caused by the limited dataset size, our focus was on meticulously defining features to strengthen the predictive capabilities of our models and to model the time series profile of our data. More precisely, we considered time-based features related to the date and time of a CGM value observation and whether the observation was made on a weekend day or not. This enabled us to capture the variability in CGM profiles that may be observed between weekdays and weekends. For CGM data, in addition to the CGM values, we considered the following features:

- TIR: the percentage of time, on a 24-h rolling horizon, the CGM values were within the recommended target;

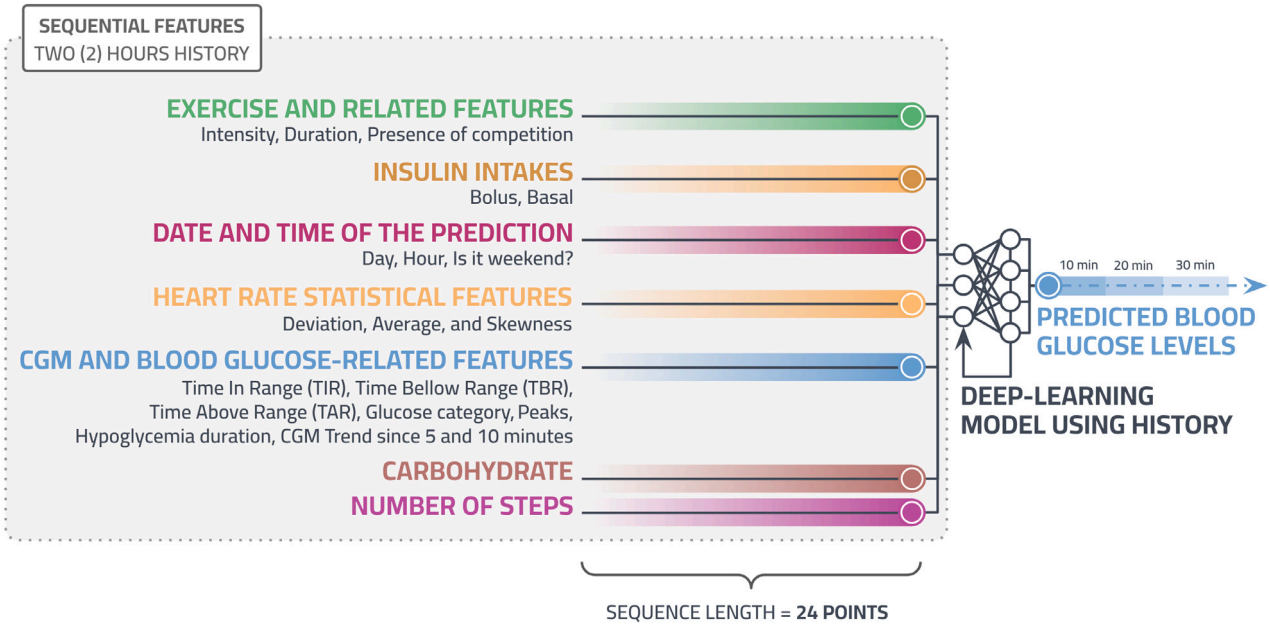


Fig. 6. Sequential input features of deep recurrent models using history.

- TAR: the percentage of time, on a 24-h rolling horizon, CGM values exceeded 180 mg/dL;
- TBR: the percentage of time, on a 24-h rolling horizon, CGM values fell below 70 mg/dL;
- The rate of change in CGM values for the previous five and 10 min;
- The time spent in hypoglycemia since its occurrence. More precisely, if at a timestamp t, the CGM value was below 70mg/dL (the patient was experiencing hyperglycemia at timestamp t), the feature value was set to the time duration that had elapsed since the start of this hypoglycemia. Otherwise (no hypoglycemia at timestamp t), the feature had value 0;
- CGM category: CGM values were categorized into distinct glucose ranges based on the risk thresholds used in the gMSE measure. Five categories associated with the following risk threshold intervals were considered: <70 mg/dL; $[70, 85[$; $[85, 155[$; $[155, 180[$; and ≥180mg/dL;
- Peak/Hollow: peaks and hollows, if any, observed at the previous timestamp $(t-1)$;
- Relative standard deviation (RSD): computed over the last 30 min.

As will be described in Section 3.3, we considered AI models with and without history. For the models with no history, we additionally considered lag features to integrate some historical information. Figs. 5 and 6 illustrate the feature engineering process for the two types of models, respectively. Deep learning models (LSTM with history, Dual-encoder, and CNN-LSTM) use sequential data as input information. In particular, an input matrix of 24×24 data points enables considering a history of two hours (based on five-minute data points) for 24 distinct features. Machine learning models, and LSTM without history, use a one-dimensional vector as input information. Eleven features, for which past information could be important, were hence flattened and modeled as time lags (see Fig. 5). Given that modeling 24 separate lags (two hours of history) for all 11 types of inputs would produce a very large number of distinct features (264), the total number of lags was limited to 50, based on the Gain measures. The Gain computes, during training, the relative importance or contribution of each timestamp of each type of input for each patient [61]. It measures the improvement in accuracy brought by a feature to the branches it is on. The number of lags considered for each type of input (10, five, four, or one) is fixed according to each time lag relative contribution. Then, time intervals of different widths are generated for each type of input (over 24 timestamps) ensuring a relatively similar cumulative contribution of each interval.

3.3. Machine and deep-learning pipeline: models design, tuning, and training

We implemented three tree-based machine learning models: a Random Forest model and two versions of XGBoost – one version using the classical RMSE as a loss function, and a second version with gMSE as the loss function. The actual CGM values were used when computing errors both for loss computation and quality assessment. We used Python Scikit-Learn[4] and XGBoost[5] libraries to implement those models. Each of the three ML models was trained three times, using different subsets of training data. The final prediction corresponds to an algebraic average of the outputs of the three models.

Using TensorFlow,[6] we also implemented four types of deep learning models: two variants of Long Short-Term Memory (LSTM) neural networks, a hybrid model combining the Convolutional Neural Network (CNN) with the LSTM architecture, and a dual-encoder architecture enhanced with attention layers. The first variant of LSTM (referred to as LSTM without history), does not consider historical data but rather uses the lag features described in Fig. 5. The second LSTM version (referred to as LSTM with history) considers historical information on a fixed time window of two hours. We implemented each of the four DL models twice: firstly by training the model only on the data of the target patient (three weeks of training and one week for testing) and secondly by pre-training the model on the four-week data of 78 patients and then fine-tuning it to the target patient (three weeks of training on the target patient data and the fourth week for testing). During the fine-tuning stage, a 10x smaller learning rate and fewer epochs (35 instead of 180) were applied to avoid damaging already good weights and overfitting the target patient. In the following sections, we prefix the fine-tuned models by "FT" when referring to them. Hence, we finally obtained a diverse combination of 11 distinct models and configurations (loss

[4] Scikit-Learn: https://scikit-learn.org/.
[5] XGBoost: https://xgboost.readthedocs.io/.
[6] TensorFlow: https://www.tensorflow.org/.

function, use of history, and method of training). In what follows, we refer to each combination as a "model" to alleviate the presentation.

We used Optuna,[7] an open-source hyperparameter optimization framework, to tune the hyperparameters of the three different ML models. Hyperparameters space was explored with the tree-structured Parzen estimator. For the Random Forest model, we optimized the number of trees in the forest, the maximum depth of the trees, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node. For the XGBoost model, we fine-tuned hyperparameters, including the learning rate, the subsample fraction, the feature fraction per tree, the minimum loss reduction required to perform additional partitioning, and the L1 and L2 regularization terms.

KerasTuner,[8] an open-source hyperparameter tuning library for neural networks, was used to tune the hyperparameters of each DL model. A random search was chosen for hyperparameters space exploration. The main hyperparameters tuned were the number of cells at each layer and the learning rate of the Adam optimizer.

Our ML and DL models aim to predict the BGL values for different prediction horizons from five to 30 min. Hence, we are in the context of multi-step forecasting. A Multiple-Input Multiple-Output (MIMO) strategy was considered to learn a single model that produces the entire prediction horizon as a direct output. Figs. 7(a), 7(b), and 8 display the abstract architectures of our DL models before hyperparameter tuning. The architecture of both our LSTM models consists of a stack of two layers of LSTM cells (used to process the input sequences together, highlighting the relationships between information at each timestamp) and two layers of perceptrons to make the final prediction. Then, both the Dual-Encoder and the CNN-LSTM networks are extensions of the LSTM model designed to enhance its feature extraction and representation learning capability. The CNN-LSTM uses additional convolutional and pooling layers to extract local features before processing the sequences. The Dual-Encoder uses an attention layer to highlight not only the relationships between timestamps on the same sequence but also with information from other sequences (e.g., between the combined CGM sequence and a meal event at time $t - 10$). Its final MLP layers could not make a complete prediction only using the attention scores. Hence, those scores are concatenated with the final output of the first LSTM encoder to retrieve a plainer version of the input information.

We implemented a complete Machine Learning Operations (MLOps) pipeline to automatically execute all three phases (data cleaning and pre-processing; feature building; and models' training and testing) on supercomputers provided by the Digital Research Alliance of Canada.[9] These supercomputers offer high-performance GPUs (like Nvidia Tensor Core A100) and enough working memory (up to 100 Giga of RAM) to concurrently train several models on significantly large datasets.

4. Results

The purpose of this section is threefold. First, we evaluate the performance of the implemented models in general (Section 4.1). Second, we assess the quality of the predictions made for the 79 studied patients globally over the one-week test (Section 4.2). Third, we specifically analyze the impact of exercising on the quality of our predictions during PA and over two intervals of time following a PA: two hours and four hours (Section 4.3). This is done for each category of PA, namely aerobic, resistance, and interval training. Although the proposed models predict the BGL over six different time horizons simultaneously (5, 10, 15, 20, 25, and 30 min), our analyses focus on three prediction horizons (10, 20, and 30 min) to alleviate the presentation. They mainly rely on two metrics extensively used in T1D studies and for regression models: the statistical metric RMSE (measured in mg/dL) and the clinical metric Clarke Error Grid Analysis (CEGA). More precisely, we measure the PDE in the CEG. In the following, the main results of each section are presented at the end of the section. Each new result is referred to by "Rx" where "x" indicates the result's number. Main findings and limitations are presented in Section 5.

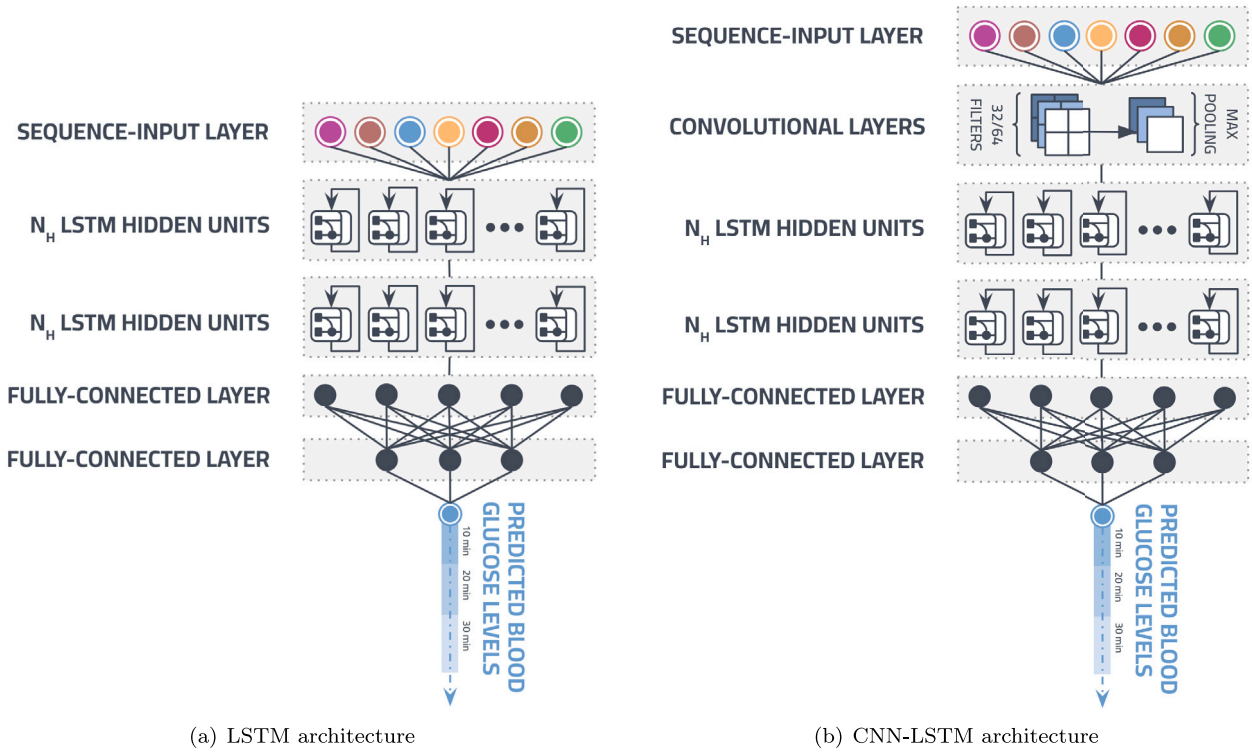
4.1. Models performance

Table 3 summarizes the RMSE results obtained with each model for each prediction horizon. The column "Best" displays, for each model, the number of patients (over the 79 considered) for which the model shows the best performance (the smallest RMSE) in comparison to the other models. The column "Worst" indicates the number of patients for which the model shows the worst performance. Table 3 additionally reports for each model, under columns "Top 3" and "Worst 3", the number of patients for which the model results in an RMSE among the three smallest and the three largest ones, respectively. For example, when compared to the other 10 models, the model "XGBoost with gMSE" (first line): (i) resulted in the smallest RMSE values for 4 patients, (ii) was, for 20 patients, among the models that gave the three smallest RMSE values, (iii) never gave the largest RMSE value for any patient, and (iv) was among the three worst models (with regard to RMSE) for only 3 patients. Columns "Min", "Max", "Mean", "Median", "Q1", and "Q3" in Table 3 give respectively the smallest, the largest, the average, the median, the first, and third quartile values obtained for each model. These values were computed over the RMSE values obtained with this model for the 79 patients. We display the median, Q1, and Q3 values because, for both metrics and all models, the patients' performances are not normally distributed (according to the Shapiro–Wilk test). Table 4 reports the same information as Table 3, but for the PDE metric.

Figs. 9 and 10 illustrate the box plots associated with the RMSE and PDE metrics, respectively, for each model and each prediction horizon. These box plots represent the distribution of RMSE and PDE values, providing a summary of central tendency and spread.

Twelve main results can be deduced from Table 3 and Fig. 9. First, FT LSTM with history, FT Dual-encoder, and FT LSTM without history emerge as being the best-performing models in terms of RMSE (based on columns "Best", "Top 3", "Min", and "Max", as well as the mean, the median, Q1, and Q3 values) for the majority of the patients and for the three prediction horizons (**R1**). This result is even clearer for 30-min predictions. Both versions of XGBoost appeared as the best ("Min" or "Top 3") models for several patients, especially for predictions made 10 and 20 min ahead (**R2**). We can also notice that fine-tuned deep-learning models (including FT CNN-LSTM) perform better than their single patient variants (**R3**). Single patient DL models performed worse, for some patients, than classical ML models (especially XGBoost) because they need more data to generalize and learn rare patterns (**R4**). This may also be one of the reasons explaining why fine-tuned models, pre-trained with the data of the whole population, perform better. When comparing, for each patient, the predictions obtained with its best-performing model, often an FT model, and the corresponding single-patient variant, we first observed that the predictions generated by FT models are often closer to the true BGL values (**R5**). One should notice, however, that the curve shape of the predictions output by the single-patient models is closer to that of the true BGL values, even if they are more shifted upwards or downwards, inducing more significant prediction errors. This is especially the case for patients with relatively stable BGLs (the easiest patients to predict). Predictions generated by FT models sometimes repeat existing patterns from the curves of other patients yielding a curve shape that may deviate from the real BGL curve shape but still with smaller prediction errors than the single-patient models (**R6**). Figs. 11 and 12 illustrate those results by comparing the behaviors of different models (LSTM with history and Dual-encoder) in diverse contexts: while exercising, during

[7] Optuna: https://optuna.org/.
[8] KerasTuner: https://keras.io/keras_tuner/.
[9] Digital Research Alliance of Canada: https://alliancecan.ca/.



(a) LSTM architecture (b) CNN-LSTM architecture

Fig. 7. LSTM and CNN-LSTM architectures.

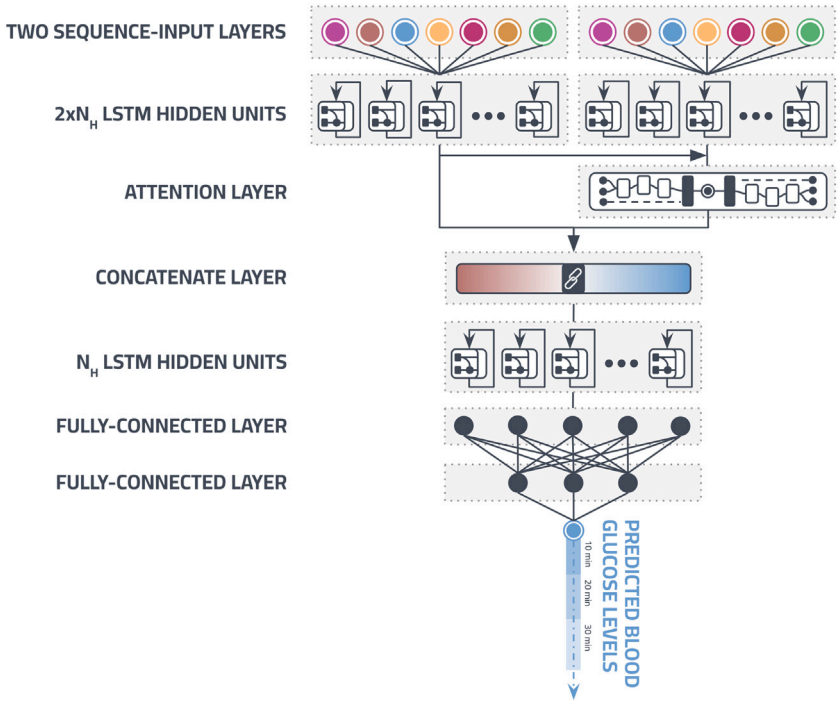


Fig. 8. Dual-Encoder architecture.

hypoglycemia and hyperglycemia episodes, or after a meal or a bolus injection for two patients. In Fig. 12, vertical green bars correspond to exercise sessions, red bars to meals, and yellow bars to bolus injections.

We can also infer that the configuration choice in terms of loss function – for XGBoost – and use of history strongly depends on the patient (**R7**). According to columns "Max", "Worst", and "Worst 3", the worst-performing models are single patient CNN-LSTM followed by its fine-tuned version (**R8**). However, other single-patient DL models can reach, for several patients, the worst "Max" value. According to median and Q1 values, the RMSEs obtained by a model for most patients are close to the best one it achieved (**R9**). This is especially true for 10-min predictions. For all the models and all the prediction horizons, the median value is lower than the average. This means that large values from a few challenging patients increase the mean value (**R10**). We can

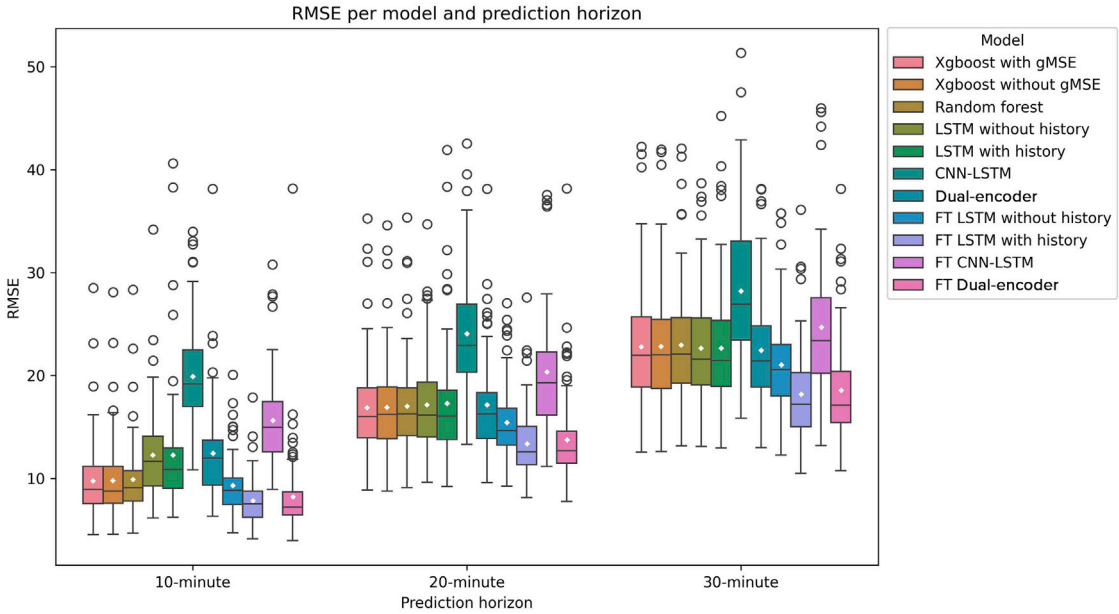


Fig. 9. Comparison of the models based on the RMSE metric for 10-, 20- and 30-min prediction horizons.

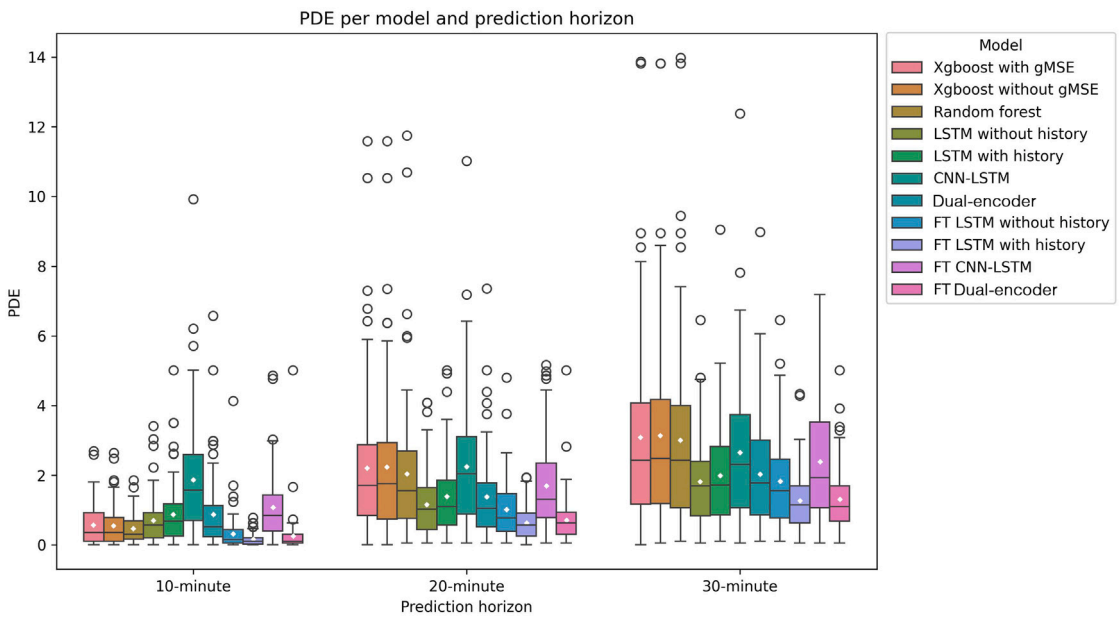


Fig. 10. Comparison of the models based on the PDE metric for 10-, 20- and 30-min prediction horizons.

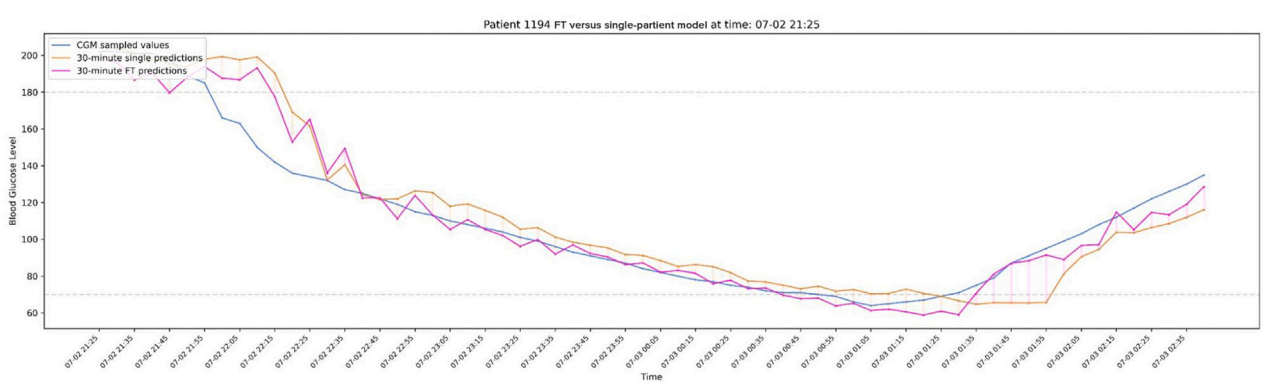


Fig. 11. Comparison between a fine-tuned model (FT Dual-encoder) and its single-patient counterpart (Dual-encoder) around hyperglycemia and hypoglycemia episodes for patient 1194.

Table 3
Models comparison based on the RMSE metric.

Model	Best	Top 3	Worst	Worst 3	Min	Max	Mean	Median	Q1	Q3
10-min predictions										
XGBoost with gMSE	4	20	0	3	4.55	28.50	9.76	8.93	7.56	11.18
XGBoost without gMSE	4	19	0	3	4.57	28.08	9.81	8.77	7.60	11.19
Random Forest	0	12	0	2	4.69	28.34	9.90	9.10	7.82	10.76
LSTM without history	0	2	1	26	6.15	34.17	12.27	11.67	9.30	14.10
LSTM with history	0	0	2	20	6.23	40.60	12.29	10.88	9.06	12.98
CNN-LSTM	0	0	69	77	10.83	33.99	19.92	19.19	17.00	22.48
Dual-encoder	0	1	1	36	6.32	38.13	12.45	11.97	9.36	13.73
FT LSTM without history	2	41	0	0	4.72	20.07	9.31	8.83	7.47	10.06
FT LSTM with history	35	70	0	0	4.13	17.89	7.86	7.54	6.22	8.76
FT CNN-LSTM	0	0	6	69	8.94	30.78	15.65	14.97	12.58	17.45
FT Dual-encoder	34	72	0	1	3.95	38.17	8.24	7.22	6.46	8.70
20-min predictions										
XGBoost with gMSE	1	4	0	10	8.87	35.23	16.89	16.04	13.96	18.81
XGBoost without gMSE	0	9	1	11	8.77	34.60	16.92	16.23	13.86	18.86
Random Forest	0	1	0	21	9.10	35.34	17.02	16.27	14.16	18.82
LSTM without history	0	4	1	15	9.63	34.68	17.15	16.17	14.02	19.34
LSTM with history	0	9	2	14	9.21	41.93	17.30	16.06	13.79	18.57
CNN-LSTM	0	0	68	77	13.31	42.52	24.06	22.95	20.33	26.93
Dual-encoder	0	6	0	19	9.61	38.13	17.17	16.27	13.91	18.34
FT LSTM without history	0	53	0	2	9.24	27.03	15.45	14.65	13.26	16.82
FT LSTM with history	45	76	0	0	8.14	27.58	13.36	12.60	11.35	15.08
FT CNN-LSTM	0	0	7	66	11.17	37.56	20.34	19.31	16.16	22.28
FT Dual-encoder	33	75	0	2	7.79	38.16	13.77	12.68	11.49	14.59
30-min predictions										
XGBoost with gMSE	0	7	0	17	12.56	42.22	22.81	21.96	18.86	25.69
XGBoost without gMSE	0	8	1	21	12.63	41.94	22.84	22.00	18.76	25.44
Random Forest	1	2	0	28	13.19	42.06	22.97	22.08	19.27	25.60
LSTM without history	0	4	1	18	13.09	38.70	22.65	21.61	19.07	25.60
LSTM with history	0	8	2	9	12.96	45.22	22.66	21.46	18.96	25.33
CNN-LSTM	0	0	64	76	15.87	51.34	28.20	26.92	23.46	33.05
Dual-encoder	0	8	0	14	12.99	38.14	22.47	21.43	18.88	24.84
FT LSTM without history	0	45	0	1	12.29	35.75	21.04	20.57	18.04	23.00
FT LSTM with history	44	76	0	0	10.48	36.12	18.21	17.21	15.04	20.28
FT CNN-LSTM	1	3	11	52	13.21	45.96	24.68	23.40	20.23	27.56
FT Dual-encoder	33	76	0	1	10.75	38.14	18.58	17.13	15.44	20.39

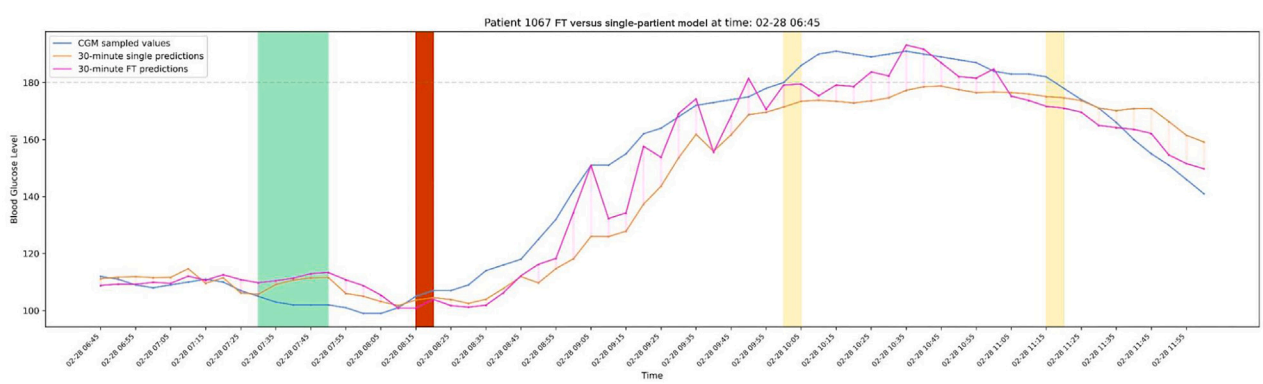


Fig. 12. Comparison between a fine-tuned model (FT LSTM with history) and its single-patient counterpart (LSTM with history) around exercise (in green), meal (in red), and bolus events (in yellow) for patient 1067.

deduce from the metrics "Min", "Max", "Mean", "Q1", and "Q3" that the accuracy of the predictions is strongly affected by the prediction horizon (**R11**). Even if some models perform better than others in general, they can be the worst ones for some patients. Likewise, some models are often among the worst-performing models but can be the best ones for some patients (especially for a 30-min prediction horizon). Hence, we can conclude that each patient has one or a few well-performing prediction models. However, we cannot establish that one model is always better than the others for all patients (**R12**).

Table 4 and Fig. 10 highlight five additional observations (R13 to R17). First, results R3, R7, R8, R9, R10, and R11 are also valid for the PDE metric. At all prediction horizons, the Random Forest model, both XGBoost, and the two versions of CNN-LSTM are often among the worst-performing models with regard to PDE (**R13**). FT LSTM with history, FT Dual-encoder, and FT LSTM without history are once again the best-performing models for most patients (**R14**). Both XGBoost models perform well for 10-min predictions but are often among the worst ones when the prediction horizon increases (**R15**). However, we notice that several models appear many times in columns "Best"/"Top 3" but also in columns "Worst"/"Worst 3". Hence, the dependency between the model performance and the patient (noted in result R10) is even stronger for the PDE metric (**R16**) confirming once again that personalization is important and that there is no single model that fits to all patients. We notice that for all prediction horizons, some patients who did not experience many hypoglycemia and hyperglycemia episodes obtained very low (<1%) PDE values (**R17**).

Table 4
Models comparison based on the PDE metric.

Model	Best	Top 3	Worst	Worst 3	Min	Max	Mean	Median	Q1	Q3
10-min predictions										
XGBoost with gMSE	3	9	1	13	0.0%	2.69%	0.57%	0.36%	0.11%	0.93%
XGBoost without gMSE	2	16	1	10	0.0%	2.64%	0.55%	0.36%	0.10%	0.78%
Random Forest	2	13	2	9	0.0%	1.86%	0.47%	0.31%	0.16%	0.70%
LSTM without history	1	5	1	18	0.0%	3.41%	0.71%	0.57%	0.21%	0.93%
LSTM with history	0	3	4	33	0.0%	5.01%	0.87%	0.68%	0.26%	1.17%
CNN-LSTM	0	0	58	74	0.0%	9.92%	1.87%	1.57%	0.7%	2.58%
Dual-encoder	1	6	1	29	0.0%	6.58%	0.87%	0.52%	0.24%	1.12%
FT LSTM without history	10	58	0	2	0.0%	4.13%	0.32%	0.15%	0.05%	0.44%
FT LSTM with history	31	64	0	0	0.0%	0.78%	0.17%	0.10%	0.02%	0.21%
FT CNN-LSTM	0	2	11	47	0.0%	4.86%	1.08%	0.84%	0.4%	1.44%
FT Dual-encoder	29	61	0	2	0.0%	5.01%	0.25%	0.10%	0.05%	0.31%
20-min predictions										
XGBoost with gMSE	1	3	9	40	0.0%	11.59%	2.21%	1.71%	0.84%	2.88%
XGBoost without gMSE	1	2	7	47	0.0%	11.59%	2.23%	1.76%	0.74%	2.93%
Random Forest	0	3	9	38	0.05%	11.75%	2.04%	1.55%	0.76%	2.69%
LSTM without history	5	19	2	3	0.05%	4.08%	1.16%	1.03%	0.44%	1.65%
LSTM with history	0	10	2	13	0.05%	5.01%	1.39%	1.10%	0.57%	1.86%
CNN-LSTM	1	1	37	56	0.05%	11.02%	2.25%	2.04%	0.89%	3.11%
Dual-encoder	0	11	0	7	0.05%	7.36%	1.38%	1.05%	0.52%	1.78%
FT LSTM without history	9	39	0	2	0.05%	4.80%	1.02%	0.77%	0.38%	1.47%
FT LSTM with history	35	74	0	0	0.0%	1.94%	0.63%	0.57%	0.26%	0.92%
FT CNN-LSTM	0	5	13	30	0.05%	5.17%	1.70%	1.31%	0.78%	2.35%
FT Dual-encoder	27	70	0	1	0.05%	5.01%	0.71%	0.63%	0.31%	0.94%
30-min predictions										
XGBoost with gMSE	1	2	8	54	0.0%	13.87%	3.09%	2.43%	1.16%	4.06%
XGBoost without gMSE	0	0	13	52	0.05%	13.82%	3.14%	2.48%	1.19%	4.17%
Random Forest	0	2	18	46	0.10%	13.98%	3.0%	2.43%	1.06%	4.00%
LSTM without history	3	26	2	2	0.05%	6.45%	1.82%	1.7%	0.83%	2.40%
LSTM with history	0	13	3	6	0.10%	9.05%	2.00%	1.72%	0.86%	2.82%
CNN-LSTM	1	1	20	33	0.05%	12.38%	2.65%	2.3%	1.07%	3.74%
Dual-encoder	2	13	0	5	0.10%	8.98%	2.03%	1.78%	0.86%	3.00%
FT LSTM without history	7	31	0	5	0.10%	6.45%	1.82%	1.55%	0.77%	2.46%
FT LSTM with history	33	72	0	0	0.05%	4.33%	1.27%	1.15%	0.63%	1.7%
FT CNN-LSTM	0	7	15	32	0.05%	7.18%	2.38%	1.93%	1.07%	3.53%
FT Dual-encoder	32	70	0	2	0.05%	5.01%	1.31%	1.10%	0.68%	1.69%

4.2. Predictions quality per patient

To obtain a personalized approach in practice, one first needs to collect enough data from each new patient to train (or fine-tune) the AI models. In this research, we considered three weeks of data for training and one week for testing. The testing is required to evaluate the models' performance for each patient. To identify the best performing model for each patient, one can consider either the model that achieves the lowest RMSE value (RMSE criterion) or the one that produces the lowest PDE value (PDE criterion). In the following, to alleviate the presentation, the best performing model (for each patient) determined based on the RMSE criterion will be referred to as: "the best RMSE model". The one (for each patient) determined based on the PDE criterion will be referred to as: "the best PDE model". It is common for two (or more) models to achieve the same PDE value for a patient. In case of a tie, the RMSE metric is secondly used to determine the best PDE model. In this section, we study the best model on the test set based on either the RMSE or the PDE criterion. Our main goal is to evaluate the quality of the prediction of the best model (with regard to a given criterion) found for each patient on the test set. Table 5 gives the RMSE and PDE values for each patient (on the test set), achieved by their best RMSE and PDE models. Displaying these results allows assessing if there exists a model (not necessarily the same model for all patients) that can accurately and safely predict blood glucose levels for each patient. Table 6 summarizes the statistics for all patients, namely the lowest, highest, mean, average, Q1, and Q3 values for both the PDE and RMSE metrics. These results provide an overview of the prediction quality across all patients considering their best RMSE and PDE models.

Figs. 13–16 (Annex A) display the RMSE and PDE values reached by all the models for each patient for 10- and 30-min prediction horizons. To go further in our analysis, we additionally draw in Figs. 17–21 (Annex B) the predictions, the actual BGLs, and the corresponding errors for particular patients for whom we either obtained the best (lowest) or worst (largest) RMSE values. In these figures, the dark green zones correspond to exercise sessions, and the light green zones to the period of two hours following an exercise session. The two horizontal lines delimit the target zone (BGL $\geq$ 70 mg/dL and BGL $\leq$ 180 mg/dL). Figs. 17, 18, and 19 display this information for the three patients with the three best (lowest) RMSE values. Figs. 20 and 21 give the information for the two patients with the largest RMSE values (obtained with their best RMSE models). We also displayed in Figs. 22, 23, and 24 (Annex C) the CEGs of the three patients with the best (lowest) PDE values. Figs. 25 and 26 (Annex C) depict the CEGs of the two patients with the worst (highest) PDE values (obtained with their best PDE models).

Eight main results can be highlighted (R18 to R25). First, the RMSE values obtained for all patients with their best RMSE models are relatively small (**R18**). This is particularly true for predictions made 10 min ahead where RMSE values range from 3.95 mg/dL (Patient 1322) to 15.31 mg/dL (Patient 1351) with an average of 7.47 mg/dL over the 79 patients. When compared to existing literature, an average RMSE of 17.74 mg/dL (<1 mmol/L) for 30-min predictions (as achieved in this study) is considered as relatively low, especially in the context of the T1DEXI study from which our data set is derived, which involved a large number of patients exercising in free-living conditions and models tested over a week of data (**R19**). Best PDE models show good PDE performance with an average value of 0.11% for 10-min predictions and 1.12% for 30-min predictions (**R20**). The gap in average PDE values between 10-min and 30-min predictions is not large, ≈1% (**R21**). The maximal value reached is 0.63% for 10-min predictions (Patient

Table 5
RMSE and PDE values for each patient and each prediction horizon obtained with their best RMSE and PDE models.

Patient	Best RMSE models						Best PDE models					
PH →	10 min		20 min		30 min		10 min		20 min		30 min	
	RMSE	PDE	RMSE	PDE	RMSE	PDE	RMSE	PDE	RMSE	PDE	RMSE	PDE
11	5.35	0.31%	9.29	0.94%	12.65	2.19%	5.74	0.16%	9.54	0.89%	12.73	2.14%
14	11.46	0.63%	18.63	1.41%	24.84	2.19%	11.67	0.47%	19.07	1.10%	25.32	2.14%
104	6.53	0.00%	11.17	0.16%	15.07	0.26%	6.53	0.00%	11.29	0.05%	15.07	0.26%
107	9.48	0.05%	15.85	0.47%	21.22	0.89%	9.48	0.05%	15.85	0.47%	21.22	0.89%
115	6.71	0.00%	13.86	0.63%	20.40	1.04%	6.71	0.00%	14.30	0.57%	21.43	0.77%
127	7.33	0.00%	12.15	0.47%	16.33	0.78%	7.33	0.00%	12.15	0.47%	16.33	0.78%
1004	8.70	0.10%	15.60	0.10%	20.39	0.16%	8.73	0.00%	15.60	0.1%	20.39	0.16%
1012	7.66	0.37%	12.28	0.78%	16.15	1.10%	7.66	0.37%	12.28	0.78%	16.15	1.10%
1014	8.06	0.63%	12.05	1.10%	14.81	1.62%	8.86	0.42%	12.74	0.94%	19.73	1.31%
1016	7.07	0.00%	12.68	0.05%	16.67	0.42%	7.07	0.00%	12.68	0.05%	16.67	0.42%
1020	6.45	0.10%	9.97	0.57%	12.92	0.99%	6.45	0.10%	11.09	0.46%	12.92	0.99%
1021	5.90	0.21%	9.82	0.78%	13.20	1.46%	6.35	0.10%	9.82	0.78%	13.20	1.46%
1033	6.97	0.00%	13.88	0.68%	19.95	1.36%	6.97	0.00%	14.09	0.47%	19.96	1.10%
1051	8.28	0.31%	14.78	1.41%	21.33	2.82%	8.28	0.31%	15.31	0.68%	21.52	1.62%
1067	5.11	0.05%	8.14	0.16%	10.48	0.31%	5.54	0.00%	9.24	0.15%	10.48	0.31%
1100	8.08	0.1%	12.68	0.16%	16.98	0.16%	8.36	0.00%	19.45	0.05%	18.65	0.15%
1110	8.75	0.21%	13.89	0.68%	18.54	1.62%	9.07	0.10%	14.20	0.63%	18.73	1.57%
1119	6.43	0.37%	11.54	1.31%	15.98	2.47%	9.71	0.31%	13.95	1.14%	19.09	1.97%
1120	5.74	0.10%	10.40	0.05%	14.53	0.05%	5.74	0.10%	10.40	0.05%	14.53	0.05%
1121	8.76	0.31%	14.48	0.99%	20.01	2.61%	8.76	0.31%	14.48	0.99%	21.12	2.04%
1123	11.61	0.1%	18.58	0.73%	24.93	0.89%	11.61	0.10%	25.30	0.57%	24.93	0.89%
1127	5.27	0.05%	10.83	0.31%	15.71	0.73%	5.95	0.00%	10.83	0.31%	15.71	0.73%
1141	7.16	0.26%	12.97	2.54%	17.84	2.54%	7.39	0.21%	19.61	1.2%	23.52	1.88%
1143	7.29	0.16%	12.85	0.42%	16.97	0.68%	8.61	0.15%	12.85	0.42%	16.97	0.68%
1146	7.78	0.05%	13.16	0.52%	18.30	1.04%	7.78	0.05%	20.73	0.41%	18.53	0.73%
1152	11.72	0.37%	21.47	0.94%	30.37	1.88%	12.24	0.31%	21.47	0.94%	30.37	1.88%
1155	6.96	0.62%	13.13	0.37%	18.71	0.89%	7.12	0.00%	13.23	0.26%	18.71	0.89%
1157	6.84	0.00%	11.51	0.05%	14.38	0.05%	6.84	0.00%	11.51	0.05%	14.38	0.05%
1171	8.51	0.16%	13.62	0.89%	17.72	1.10%	8.51	0.16%	14.46	0.68%	17.72	1.10%
1185	5.32	0.1%	10.32	0.63%	14.56	1.15%	5.32	0.10%	12.00	0.52%	14.94	1.04%
1194	4.55	0.00%	7.79	0.05%	10.75	0.10%	4.55	0.00%	8.77	0.00%	12.56	0.00%
1195	7.23	0.00%	12.35	0.1%	17.21	0.84%	7.23	0.00%	12.35	0.10%	17.21	0.84%
1201	7.78	0.16%	12.44	1.04%	15.81	1.51%	7.78	0.16%	12.63	0.84%	15.81	1.51%
1202	4.72	0.00%	9.36	0.52%	14.59	1.73%	4.72	0.00%	9.36	0.52%	14.59	1.73%
1203	7.54	0.11%	11.33	0.26%	14.49	0.37%	8.92	0.10%	11.33	0.26%	16.59	0.36%
1210	7.05	0.73%	12.11	1.88%	16.43	3.08%	7.20	0.63%	12.19	1.83%	16.46	3.03%
1211	6.82	0.10%	13.56	0.68%	19.61	1.10%	6.82	0.10%	14.96	0.63%	19.61	1.10%
1219	5.39	0.10%	10.79	0.47%	15.04	1.10%	5.39	0.10%	11.90	0.46%	15.04	1.10%
1221	5.81	0.05%	12.30	0.57%	17.92	1.31%	5.81	0.05%	12.30	0.57%	18.44	1.10%
1230	7.31	0.26%	11.75	1.41%	15.01	2.04%	7.31	0.26%	12.05	0.94%	15.59	1.67%
1247	11.08	0.41%	18.58	0.73%	24.41	2.09%	11.82	0.21%	18.58	0.73%	24.87	1.19%
1249	6.89	0.05%	12.33	0.16%	17.29	0.47%	6.89	0.05%	17.40	0.10%	17.29	0.47%
1257	8.91	0.37%	15.30	0.99%	21.47	1.98%	20.07	0.15%	24.34	0.83%	21.47	1.98%
1261	8.64	0.00%	13.83	0.10%	18.12	0.21%	8.64	0.00%	13.83	0.10%	21.04	0.15%
1265	4.66	0.00%	10.07	0.26%	15.31	0.84%	4.66	0.00%	10.07	0.26%	15.31	0.84%
1266	8.63	0.00%	13.66	0.05%	17.56	0.10%	8.63	0.00%	13.66	0.05%	17.56	0.10%
1273	6.93	0.16%	11.18	0.57%	14.51	0.94%	6.93	0.16%	11.18	0.57%	14.51	0.94%
1281	4.23	0.00%	8.37	0.00%	11.82	0.05%	4.23	0.00%	8.37	0.00%	11.82	0.05%
1283	5.81	0.00%	13.12	1.10%	20.66	4.28%	5.81	0.00%	13.12	1.10%	22.02	3.92%
1291	7.48	0.05%	12.54	0.37%	16.68	0.68%	8.55	0.00%	14.66	0.31%	16.68	0.68%
1292	7.45	0.16%	13.67	0.31%	18.85	0.63%	10.06	0.05%	13.67	0.31%	18.85	0.63%
1297	6.83	0.16%	11.29	0.63%	14.62	0.89%	8.01	0.15%	13.46	0.36%	17.75	0.52%
1303	6.53	0.00%	12.78	0.16%	17.25	0.47%	6.53	0.00%	13.19	0.10%	17.68	0.31%
1311	6.15	0.00%	11.51	0.21%	16.65	1.10%	6.15	0.00%	11.51	0.21%	16.65	1.10%
1312	10.17	0.21%	15.06	0.47%	18.44	1.52%	10.17	0.21%	15.06	0.47%	18.44	1.52%
1322	3.95	0.00%	8.19	0.47%	11.77	0.84%	3.95	0.00%	9.59	0.26%	11.77	0.84%
1325	6.66	0.47%	12.51	1.67%	17.08	2.82%	9.03	0.21%	12.51	1.67%	17.08	2.82%
1328	10.07	0.00%	17.18	0.26%	23.51	0.63%	10.07	0.00%	20.84	0.15%	26.74	0.52%
1336	9.55	0.42%	15.85	0.99%	21.43	1.36%	9.55	0.42%	15.85	0.99%	21.43	1.36%
1338	7.03	0.16%	12.24	0.31%	16.82	0.52%	7.70	0.10%	12.24	0.31%	16.82	0.52%
1339	8.91	0.21%	15.53	0.73%	21.54	1.31%	9.15	0.05%	15.53	0.73%	24.87	1.14%
1345	5.66	0.05%	11.54	0.37%	16.89	0.63%	5.66	0.05%	11.75	0.31%	16.89	0.63%
1348	13.07	0.21%	22.17	0.89%	30.58	1.52%	13.97	0.16%	22.84	0.78%	31.07	1.36%
1351	15.31	0.63%	24.65	1.36%	32.31	3.03%	17.89	0.52%	24.65	1.36%	32.31	3.03%
1354	5.34	0.05%	10.94	0.16%	15.58	0.37%	6.45	0.00%	17.42	0.10%	21.47	0.21%
1357	4.79	0.00%	10.73	0.63%	16.85	1.83%	4.79	0.00%	10.73	0.63%	17.13	1.51%
1359	8.58	0.42%	14.41	0.94%	19.34	1.67%	9.12	0.21%	15.17	0.78%	19.34	1.67%
1361	6.21	0.10%	12.05	1.10%	16.66	2.94%	6.46	0.00%	12.12	1.00%	16.76	2.78%
1365	6.86	0.10%	11.35	0.57%	14.99	0.78%	6.86	0.10%	11.71	0.47%	14.99	0.78%

(*continued on next page*)

Table 5 (*continued*).

Patient	Best RMSE models						Best PDE models					
PH →	10 min		20 min		30 min		10 min		20 min		30 min	
	RMSE	PDE	RMSE	PDE	RMSE	PDE	RMSE	PDE	RMSE	PDE	RMSE	PDE
1378	6.25	0.00%	10.45	0.47%	13.67	0.94%	6.25	0.00%	10.90	0.21%	13.67	0.94%
1387	7.83	0.10%	13.57	1.15%	19.35	1.57%	7.83	0.10%	13.57	1.15%	19.35	1.57%
1400	5.02	0.10%	9.06	0.21%	12.27	0.16%	5.31	0.05%	9.10	0.16%	12.27	0.16%
1406	6.74	0.05%	11.72	0.52%	16.19	1.78%	6.74	0.05%	12.30	0.42%	16.19	1.78%
1410	6.99	0.21%	11.70	0.37%	15.29	0.52%	8.17	0.15%	11.70	0.37%	15.29	0.52%
1412	6.91	0.10%	11.73	0.63%	15.72	1.20%	6.91	0.10%	12.02	0.58%	21.61	1.09%
1418	9.48	0.21%	16.37	0.68%	22.14	1.25%	9.48	0.21%	16.37	0.68%	22.14	1.25%
1422	13.48	0.37%	21.91	0.78%	29.12	1.83%	14.72	0.21%	21.91	0.78%	29.12	1.83%
1425	6.23	0.05%	9.90	0.05%	12.93	0.16%	6.23	0.05%	9.90	0.05%	12.93	0.16%
1427	6.72	0.10%	12.29	0.78%	16.83	2.19%	6.72	0.10%	15.83	0.77%	16.83	2.19%

Table 6
Summary of RMSE and PDE values for all patients obtained with their best RMSE and PDE models.

Metric	10 min	20 min	30 min
Best RMSE models			
Min RMSE	3.95 mg/dL	7.79 mg/dL	10.48 mg/dL
Max RMSE	15.31 mg/dL	24.65 mg/dL	32.31 mg/dL
Mean RMSE	7.46 mg/dL	13.00 mg/dL	17.74 mg/dL
Median RMSE	6.99 mg/dL	12.33 mg/dL	16.85 mg/dL
Q1 RMSE	6.22 mg/dL	11.23 mg/dL	15.02 mg/dL
Q3 RMSE	8.54 mg/dL	13.87 mg/dL	19.48 mg/dL
Min PDE	0.00%	0.00%	0.05%
Max PDE	0.73%	2.54%	4.28%
Mean PDE	0.16%	0.63%	1.22%
Median PDE	0.10%	0.57%	1.10%
Q1 PDE	0.02%	0.26%	0.63%
Q3 PDE	0.21%	0.89%	1.70%
Best PDE models			
Min RMSE	3.95 mg/dL	8.37 mg/dL	10.48 mg/dL
Max RMSE	20.07 mg/dL	25.30 mg/dL	32.31 mg/dL
Mean RMSE	7.95 mg/dL	14.00 mg/dL	18.39 mg/dL
Median RMSE	7.31 mg/dL	12.85 mg/dL	17.29 mg/dL
Q1 RMSE	6.40 mg/dL	11.60 mg/dL	15.45 mg/dL
Q3 RMSE	8.89 mg/dL	15.42 mg/dL	21.08 mg/dL
Min PDE	0.00%	0.00%	0.00%
Max PDE	0.63%	1.83%	3.92%
Mean PDE	0.11%	0.54%	1.12%
Median PDE	0.10%	0.47%	1.04%
Q1 PDE	0.00%	0.23%	0.52%
Q3 PDE	0.16%	0.78%	1.57%

1210) and 3.92% for 30-min predictions (Patient 1283) (**R22**). The results summarized in Table 6 show that the best models chosen using either the RMSE or the PDE criterion yield good and comparable results. Indeed, the mean and maximal RMSE values obtained with the best PDE models are approximately 1 mg/dL larger than those obtained with the best RMSE models for predictions made 20 min ahead. The maximal RMSE value is the same for predictions made 30 min ahead (**R23**). The maximal and mean PDE values obtained with the best RMSE models for a 30-min prediction horizon are increased by 0.36% and 0.1%, respectively, when compared to the best PDE models (**R24**). Hence, the best RMSE models also ensure safe clinical results (**R25**).

As depicted in Figs. 13–16, the BGL of some patients are more challenging to predict than others. This is true for both metrics, different prediction horizons (10 and 30 min), and all models (**R26**). Indeed, while some patients obtained good predictions with all models, relatively large RMSE and PDE values were observed for a few ones. This confirms our result (R8) that the average RMSE and PDE values increased due to a few challenging patients. We observe that for most patients, two to five models achieve almost the same performance with regard to both the RMSE and PDE metrics, and one to three models perform relatively badly (**R27**). This is true even for the patient with the best results (regarding either RMSE or PDE). The worst performing models are often the two CNN-LSTM models (with and without history). For the RMSE metric, the best-performing models for the easiest patients to predict (the patients for whom we obtained the three lowest RMSE values) are FT LSTM with history and FT Dual-encoder. This is also the case for the two most challenging patients to predict (the two patients for whom we obtained the two largest RMSE values with their best RMSE models). In contrast, the models that performed best in terms of PDE (the lowest PDE values) are different between patients (**R26**). Finally, we can also notice in Figs. 13–16 that the difference in performance between the models is larger for the PDE metric than for the RMSE metric (**R29**).

We also studied the data of the best and worst-predicted patients, with regard to RMSE, to better understand the possible reasons behind our results. The worst-predicted patients were not among the ones who experienced the longest or the shortest hypoglycemia events (see Fig. 1), nor the ones who consumed the highest or the lowest quantities of carbohydrates (see Fig. 4) (**R30**). However, in comparison with the patients with the best predictions, the most challenging patients used a relatively large quantity of bolus insulin during the data collection period (**R31**). The most challenging patients did also a lot of exercise sessions, not particularly long ones, but mainly of low intensity (**R32**). Apart from these few characteristics, it is difficult to differentiate what features make a patient hard to predict (**R33**). It would be rather explained by a specific combination of several features and patterns.

Studying the prediction, the actual BGL, and the error curves of some patients with the best and worst RMSE values (Figs. 17–21) allow us to highlight five additional observations (R34 to R38). The following observations are especially valid for those patients, but may also help understand the aggregated results. First, for the same patient, the error curves of the three prediction horizons share their most prominent peaks (**R34**). Prediction errors can be either positive (overestimation) or negative (underestimation), and rather depend on the sudden change of BGL direction (**R35**). Their amplitude also depends on the amplitude of a sudden change in BGL (**R36**). One can also notice that the patients for whom the worst predictions are obtained did almost twice as much exercise as the ones with the best predictions (**R37**). Likewise, patients with the best predictions had several hyperglycemia episodes but fewer severe hypoglycemia episodes in comparison with the most challenging ones (**R38**). For example, Patient 1351 had more than 10 hypoglycemia events in the last (fourth) week.

Analyzing the CEGs of the patients for which we obtained the best and worst PDE values (Figs. 22–26), allows us to emphasize six additional observations (R39 to R43). For all patients, the vast majority of risky predictions actually fell into zone D but not into zone E (**R39**). This means that the best PDE model sometimes fails in predicting hypoglycemia or hyperglycemia, but the predicted values were almost all the time within the target interval. Patients with the lowest PDE values did not experience many hypoglycemia episodes but experienced several hyperglycemia episodes (**R40**). The two patients with the largest PDE values experienced more than 10 hypoglycemia episodes during the fourth (test) week (**R41**). Predictions falling in Zone D were mainly obtained during hypoglycemia events. Hence, hypoglycemia episodes

Table 7
Prediction quality during and after aerobic exercises.

Metric	During exercise		2 h after		4 h after	
Best RMSE models						
	10 min	30 min	10 min	30 min	10 min	30 min
Min RMSE	2.52 mg/dL	5.65 mg/dL	3.61 mg/dL	8.97 mg/dL	3.54 mg/dL	8.25 mg/dL
Max RMSE	27.17 mg/dL	46.97 mg/dL	20.61 mg/dL	40.09 mg/dL	20.81 mg/dL	38.13 mg/dL
Mean RMSE	9.03 mg/dL	21.30 mg/dL	8.24 mg/dL	20.13 mg/dL	8.31 mg/dL	19.41 mg/dL
Median RMSE	8.08 mg/dL	20.38 mg/dL	7.65 mg/dL	18.85 mg/dL	7.66 mg/dL	18.99 mg/dL
Min PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max PDE	0.14%	0.21%	0.02%	0.08%	0.02%	0.14%
Mean PDE	0.00%	0.03%	0.00%	0.01%	0.00%	0.01%
Median PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Best PDE models						
	10 min	30 min	10 min	30 min	10 min	30 min
Min RMSE	2.18 mg/dL	5.86 mg/dL	3.52 mg/dL	10.31 mg/dL	3.84 mg/dL	8.25 mg/dL
Max RMSE	37.52 mg/dL	46.02 mg/dL	20.61 mg/dL	40.30 mg/dL	20.81 mg/dL	38.13 mg/dL
Mean RMSE	9.30 mg/dL	21.65 mg/dL	8.66 mg/dL	20.91 mg/dL	8.76 mg/dL	20.12 mg/dL
Median RMSE	8.24 mg/dL	21.01 mg/dL	7.96 mg/dL	19.32 mg/dL	8.04 mg/dL	19.45 mg/dL
Min PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max PDE	0.14%	0.21%	0.02%	0.08%	0.04%	0.14%
Mean PDE	0.00%	0.03%	0.00%	0.01%	0.00%	0.01%
Median PDE	0.00%	0.00%	0.00%	0.00%	0.0%	0.01%

seem, in our context, harder to predict than hyperglycemia ones (**R42**). The vast majority of predictions belong to clinically safe zones A and B (**R43**).

4.3. Predictions during and after physical activity

This section aims to analyze the impact of exercising on prediction quality. Physical activities are grouped into three categories: aerobic, resistance, and interval training. The results are computed based on the PA sessions carried out by all patients during the last week. Not all the studied patients exercised in all the PA categories during the test week. They did a total of 940 aerobic sessions (631 h), 136 resistance activities (95 h), and 96 interval training (88 h) during the test week. We observed that PA sessions yielded more hyperglycemia than hypoglycemia, both during and after exercising, regardless of the type of activity. The average TBRs in the test week were 4.60%, 4.67%, and 5.63% during aerobic, resistance, and interval training sessions, respectively. The average TARs were higher: 20.57%, 13.24%, and 24.77%. Similar percentages were observed during the four-hour periods following the end of an exercise. While resistance training is the activity during which the patients experienced the less hyperglycemia episodes, interval training caused both the most hyperglycemia and hypoglycemia events (on average across all patients).

Tables 7, 8, and 9 report the results obtained for each category, separately. They display the minimum, maximum, mean, and median RMSE and PDE values during exercise sessions and over the two-hour and four-hour periods following an exercise. These values are computed using the prediction output by the best RMSE and PDE models for each patient according to its performance throughout the entire test week as explained in Section 4.2 and not only during or after PA. Results are reported for predictions made 10 min and 30 min ahead.

Recall that the RMSE results of the best RMSE models for predictions made 30 min ahead over the whole test week were: min RMSE = 10.48 mg/dL, max RMSE = 32.31 mg/dL, mean RMSE = 17.74 mg/dL, and median RMSE = 16.85 mg/dL (see Table 6). The PDE results of the best PDE models and a prediction horizon of 30 min were: min PDE = 0.00%, max PDE = 3.92%, mean PDE = 1.12%, and median PDE = 1.04%. In comparison to these global results, we observed an increase in the max, mean, and median RMSE values during PA sessions, especially for predictions made 30 min ahead. Indeed, the max, mean, and median RMSE values obtained during aerobic sessions were respectively equal to 46.97 mg/dL, 21.30 mg/dL, and 20.38 mg/dL. During resistance activities, these values were equal to 32.75 mg/dL, 16.45 mg/dL, and 15.97 mg/dL. For interval training, the maximum, mean and median RMSE values were equal to 39.48 mg/dL, 18.46 mg/dL, and 18.99 mg/dL, respectively (**R44**). Hence, in terms of RMSE, BGLs during aerobic sessions appear the hardest to predict when compared to resistance and interval sessions. The variation rate and magnitude of the BGL curve during aerobic sessions could be responsible (**R45**). Recall, however, that the total duration of aerobic sessions are much longer than the other two PA categories (almost seven times more) which makes the comparison with the other categories inconclusive. The BGLs were slightly easier to predict after than during the aerobic sessions. The max, mean, and median RMSE during aerobic sessions for predictions made 30 min ahead were: 46.97 mg/dL, 21.30 mg/dL, and 20.38 mg/dL, respectively. For the interval of two hours following aerobic sessions, these values were: 40.09 mg/dL, 20.13 mg/dL, and 18.85 mg/dL, respectively. For the four hours following aerobic sessions, these values were: 38.13 mg/dL, 19.41 mg/dL, and 18.99 mg/dL, respectively (**R46**). On the contrary, BGLs were harder to predict after resistance and interval training sessions than during them. For both types of exercise, the BGLs were even harder to predict during the two hours following an exercise (**R47**). For all best models (either RMSE or PDE based) and all PA categories, PDE values were very small: below 0.5% for all patients. This is true during and over the four hours following exercises. The average and median PDE for predictions made 30 min ahead were lower during exercise sessions than globally over the whole test week (**R48**). These results show that even if the prediction errors observed during and after exercise are larger than those obtained globally over the test week, these errors occurred most of the time in the safe zones A and B of the CEG. Finally, the best models with respect to RMSE and PDE both show similar performances in terms of RMSE and PDE metrics for all the PA categories. Only the max and average RMSEs for predictions made both 10 and 30 min ahead during the two hours following interval training sessions were notably different in the two cases. In particular, the max RMSE reached 30 min ahead was 60.27 mg/dL for the best PDE models (see Table 9) and 41.21 mg/dL for the best RMSE models (**R49**).

5. Discussion

5.1. Main findings

Four major results emerge from our analyses. First, there is no single algorithm capable of correctly predicting the blood glucose levels of all patients exercising in free-living conditions. Hence a personalized

Table 8
Prediction quality during and after resistance exercises.

Metric	During exercise		2 h after		4 h after	
Best RMSE models						
	10 min	30 min	10 min	30 min	10 min	30 min
Min RMSE	0.32 mg/dL	3.40 mg/dL	3.34 mg/dL	9.11 mg/dL	3.10 mg/dL	7.42 mg/dL
Max RMSE	22.18 mg/dL	32.75 mg/dL	12.38 mg/dL	35.33 mg/dL	11.43 mg/dL	30.31 mg/dL
Mean RMSE	6.74 mg/dL	16.45 mg/dL	7.34 mg/dL	19.49 mg/dL	7.23 mg/dL	18.88 mg/dL
Median RMSE	6.00 mg/dL	15.97 mg/dL	7.10 mg/dL	18.47 mg/dL	7.00 mg/dL	17.11 mg/dL
Min PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max PDE	0.02%	0.50%	0.00%	0.24%	0.02%	0.12%
Mean PDE	0.00%	0.05%	0.00%	0.01%	0.00%	0.02%
Median PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Best PDE models						
	10 min	30 min	10 min	30 min	10 min	30 min
Min RMSE	1.74 mg/dL	3.40 mg/dL	3.34 mg/dL	9.11 mg/dL	3.10 mg/dL	7.42 mg/dL
Max RMSE	22.18 mg/dL	32.75 mg/dL	15.11 mg/dL	39.21 mg/dL	13.21 mg/dL	30.31 mg/dL
Mean RMSE	6.75 mg/dL	16.86 mg/dL	7.63 mg/dL	20.19 mg/dL	7.69 mg/dL	19.84 mg/dL
Median RMSE	6.17 mg/dL	16.79 mg/dL	7.23 mg/dL	18.39 mg/dL	7.00 mg/dL	18.21 mg/dL
Min PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max PDE	0.02%	0.50%	0.01%	0.24%	0.02%	0.12%
Mean PDE	0.00%	0.05%	0.00%	0.01%	0.00%	0.02%
Median PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 9
Prediction quality during and after interval exercises.

Metric	During exercise		2 h after		4 h after	
Best RMSE models						
	10 min	30 min	10 min	30 min	10 min	30 min
Min RMSE	2.67 mg/dL	7.82 mg/dL	3.26 mg/dL	4.4 mg/dL	3.89 mg/dL	7.81 mg/dL
Max RMSE	37.54 mg/dL	39.48 mg/dL	17.18 mg/dL	41.21 mg/dL	26.56 mg/dL	39.08 mg/dL
Mean RMSE	9.48 mg/dL	18.46 mg/dL	7.58 mg/dL	20.56 mg/dL	8.44 mg/dL	20.01 mg/dL
Median RMSE	6.84 mg/dL	18.99 mg/dL	7.00 mg/dL	20.86 mg/dL	7.05 mg/dL	18.70 mg/dL
Min PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max PDE	0.05%	0.05%	0.01%	0.07%	0.01%	0.04%
Mean PDE	0.00%	0.00%	0.00%	0.01%	0.00%	0.01%
Median PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Best PDE models						
	10 min	30 min	10 min	30 min	10 min	30 min
Min RMSE	3.47 mg/dL	7.82 mg/dL	3.26 mg/dL	4.40 mg/dL	3.89 mg/dL	7.81 mg/dL
Max RMSE	19.37 mg/dL	32.91 mg/dL	34.93 mg/dL	60.27 mg/dL	25.35 mg/dL	44.26 mg/dL
Mean RMSE	7.65 mg/dL	19.13 mg/dL	9.48 mg/dL	22.63 mg/dL	9.03 mg/dL	20.99 mg/dL
Median RMSE	6.63 mg/dL	19.19 mg/dL	7.70 mg/dL	20.68 mg/dL	7.68 mg/dL	18.53 mg/dL
Min PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Max PDE	0.07%	0.22%	0.10%	0.08%	0.07%	0.04%
Mean PDE	0.01%	0.04%	0.01%	0.01%	0.00%	0.01%
Median PDE	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

approach, adapted to the specificities of each patient, is required. The results we obtained confirm that it is generally better to train a model on many patients and then fine-tune it to each patient individually. Such an approach circumvents the difficulties related to the reduced data size and the lack of information for some patients (mainly CHO intake) while still considering each patient's particularities at the fine-tuning stage.

Second, periods of physical activities are (at least partially) responsible for the most significant statistical errors (RMSE values) yielded by our models. Indeed, the average RMSE during and after exercise periods is larger than the average value over the whole test week. This is especially true for aerobic exercises with a mean value of 21.30 mg/dL during the sessions. Except for aerobic sessions, the BGL was more challenging to predict during the two hours following resistance and interval training exercises than during the activity itself. However, exercise periods do not cause an increase in PDE values, implying that prediction errors observed during and after exercise sessions do not occur, most of the time, in the unsafe clinical zones D et E of the CEG. Yet, as mentioned in Section 4.3, in our test data patients did experience hypoglycemia and hyperglycemia episodes both during and after the three types of exercises.

Third, we were always able to produce a model for each patient that performs relatively well in terms of both the RMSE and the PDE metrics. A good-performing model was always obtained for 10-min, 20-min, and 30-min prediction horizons. For predictions made 30 min ahead, the average RMSE value (for the 79 patients) was 17.74 mg/dL (<1 mmol/L), and did not exceed 7.47 mg/dL for predictions made 10 min ahead. PDE values of many patients did not exceed 1%, and the mean value was 1.12%. The majority of the predictions were in the clinically safe zones A and B, and almost no predictions fell in zone E. Such results are considered excellent when compared to the existing literature and given the context we tackled: T1D patients exercising in free-living conditions and data collected over four weeks.

Finally, although the patients we considered in our research included many more females than males, this did not bias the precision of our models in favor of one sex in particular (a binary sex categorization (male/female) is considered and refers to the set of biological attributes at birth). Indeed, when we compared the minimal, maximal, and average RMSE values of the best RMSE models for male patients to the values obtained for female patients, the difference was relatively small for the three prediction horizons. This was also the case when we compared the minimal, maximal, and average PDE values of the

best PDE models (see Table 10 in Annex D). We also observed that the four most challenging patients with regard to RMSE and PDE metrics were all females (patients 1351, 1348 in Annex B and 1283 and 1351 in Annex C) and that two over the six less challenging ones were males (Patient 1194 in Annex B and 1120 in Annex C). It is worth recalling that our selection criteria for the patients from the T1DEXI study were not based on their sex but rather on the availability of a certain volume and type of data reported for these patients. Indeed, we required precise insulin information that was only available to pump users. Addressing exercising in free-living conditions required a minimum volume of data for exercise sessions. We also required that patients report their CHO intake. Finally, the T1DEXI study only reports the sex of patients and not their gender. Hence, an analysis of our results with regard to gender was not possible.

5.2. Limitations

Despite these excellent results, some issues need to be discussed. Indeed, even if some fine-tuned deep learning models, especially FT Dual-encoder and FT LSTM (with and without history), performed in general better than the others, there is no straightforward explanation for the matching between a patient and her/his best-performing models. This would suggest that all 11 models have to be trained for each new patient to decide which one would perform the best for her/him. In practice, it could be only the fine-tuning stage that has to be executed to alleviate the training.

After analyzing the data of the least and most challenging patients, there are no straightforward explanations of why a patient is more challenging to predict than another. We can, however, notice that the BGL curve of patients who are challenging to predict varies more quickly and with a greater magnitude. The sole other discernible differences were a relatively slightly larger number of exercise sessions (6 to 12 sessions for the worst predicted patients against five for the three best-predicted ones), a lack of CHO data, more severe hypoglycemia episodes, and a relatively large quantity of bolus insulin injected during the data collection period. These observations have to be confirmed with additional tests on new data.

Our prediction approach is firmly based on the availability and reliability of the data. To deliver accurate BGL predictions, our data-driven models require the T1D patient to use an insulin pump (either a closed-loop or a regular one), to wear a CGM sensor and a smartwatch, and to provide precise CHO intake. In practice, this last part is often done manually, which can be time consuming and represents a source of error. Recent research focus more on how to lighten the burden on patients by preventing them from calculating carbohydrates each time, especially for closed-loop pump users. The results we obtained with our fine-tuned models can help move in this direction by training our models on a population of patients (with all the required CHO information) and fine-tuning on the target patient without requiring that the CHO data is entered each time CHO are consumed. We are currently studying the impact, on the quality of the predictions, of CHO intake values and their frequencies in the training and test datasets.

When analyzing the impact of certain features on glycemic variations, we concluded that tested models are good for predicting BGLs but not as good to be used for recommendations of therapy treatments. This is because our models are fully data-driven and cannot learn biologically possible patterns absent from past data (due to patient habits). For example, we can assume that patients never tried harmful actions such as increasing insulin dosage during severe hypoglycemia or eating carbs while being in severe hyperglycemia. A few recent papers, such as Deng et al. [62] or Zou et al. [63], tried to inform data-driven models with biological laws by biasing the learning process toward known dynamics using System-Biology Informed Neural Network (SBINN) loss functions [64] and ODE-based physiological models [65]. Such a hybrid approach could predict the impact of an action or recommendation never experienced before.

Another limitation of this paper is the absence of external validation. Indeed, all tested patients are drawn from the same T1DEXI study.

6. Conclusion and avenues of research

Our main objective in this paper was to assess whether a data-driven approach can accurately and safely predict blood glucose levels in patients with type 1 diabetes exercising in free-living conditions. We also aimed to determine whether the approach that performs best varies from one patient to another. To this end, we proposed and compared different data-driven approaches to predict, up to 30 min ahead, blood glucose levels of 79 type-1 diabetes adult patients. The data used for training and testing our models was extracted from the T1DEXI study for all the adult patients who used insulin pumps (either regular or closed-loop ones), exercised at least 150 min a week and reported their CHO intake over the four weeks of the study. We adapted and tested machine learning models (XGBoost and Random Forest) and deep learning architectures (LSTM, CNN-LSTM, Dual-encoder with attention layer). We also implemented each DL model twice: firstly, by training the model only on the data of the target patient and secondly by training the model on the data of all patients and then fine-tuning it to the target patient.

Our results demonstrate the relevance of using a personalized approach for blood glucose level prediction in free-living conditions. Indeed, the best-performing model differs from one patient to another. Approaches in which a population-based model is first trained and then fine-tuned to the target patient stand out as the best for most patients. The BGL was slightly harder to predict during (and after) exercise sessions. Indeed, our models, regardless of their architecture, generally yielded a higher RMSE on average during and over the four hours following a PA. However, for all PA categories, PDE values were very small (below 0.5% for all patients) implying that prediction errors during and up to four hours after exercise sessions occurred most of the time in the safe clinical zones A and B of the CEG.

Our results also confirm that artificial intelligence models are worth exploiting to be used by both the research and healthcare industry communities to predict blood-glucose levels of type 1 diabetes patients exercising in free-living conditions, provided that all the required data are available and reliable. Indeed, the lack of data remains one of the most challenging problems. In this context, we are currently working on patient clustering methods, data augmentation, and Meta-Learning approaches which could be good alternatives.

Our analysis of the different patients' curves suggests that our models still need improvement to tackle the sudden variations in CGM values. We are working on defining additional features to improve the predictions for the most challenging patients and lengthen the prediction horizon, especially to handle nocturnal hypoglycemia.

Finally, to be adopted by T1D patients and clinical communities, AI models need to offer some information to explain their predictions to patients and physicians [6,30]. eXplainable Artificial Intelligence (XAI) and evidential learning appear to be promising avenues of research. Evidential learning techniques enable associating the outcome (prediction or recommendation) with a statistical level of confidence. Several modern XAI techniques, like the "SHapley Additive exPlanation (SHAP)" algorithm used in [10], enable measuring the positive and negative effects of each feature (and combination of features) on the prediction. Other approaches, like "Rules Extraction", enable describing the trained model. Even if a few contributions in XAI have been made with classical machine learning models [10], no concrete proposition for complex deep learning and recurrent neural networks is reported in the literature.

Acronyms

ANN Artificial Neural Network

ARIMAX ARIMA with eXogenous input

BGL Blood Glucose Level

CEG Clarke Error Grid

CEGA Clarke Error Grid Analysis

CGM Continuous Glucose Monitor

CHO Carbohydrates

CNN Convolutional Neural Network

COB Carbohydrates-On-Board

CSII Continuous Subcutaneous Insulin Infusion

DL Deep Learning

DT Decision Tree

EE Energy Expenditure

EOB Exercise-On-Board (sometimes Activity-On-Board)

FF-MLP Feed-Forward Multi-Layered Perceptron

gMSE glucose-specific Mean Square Error

gRMSE glucose-specific Root Mean Square Error

GRU Gated Recurrent Unit

HIIT High-Intensity Interval Training

IOB Insulin-On-Board

KNN K-Nearest Neighbors

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MDI Multiple Daily Injections

MIMO Multiple-Input Multiple-Output

ML Machine Learning

MSE Mean Square Error

NB Naive Bayes

NH Nocturnal Hypoglycemia

PA Physical Activity

PDE Percentage of Predictions in clinically critical zones D and E

PH Prediction Horizon

PWT1D People With Type-1 Diabetes

RF Random Forest

RMRF Repeated Measures Random Forest

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

RWC Real World (free-living) Conditions

SVM Support Vector Machine

T1D Type-1 Diabetes

TAR Time Above Target

TBR Time Below Target

TCN Temporal Convolutional Network

TIR Time In Target

XAI eXplainable Artificial Intelligence

XGBoost eXtrem Gradient Boosting

CRediT authorship contribution statement

Anas Neumann: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Yessine Zghal:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Marzia Angela Cremona:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Adnene Hajji:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Michael Morin:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Monia Rekik:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Ethics approval

The research was approved by the Research Ethics Board of Université Laval no 2024–076 on March 6, 2024.

Declaration of competing interest

The authors declare no financial and no personal interests.

Acknowledgments

This paper is based on research using data from the Type 1 Diabetes EXercise Initiative (T1DEXI) Study that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

This project was funded by the Faculty of Business Administration of Université Laval and Fondation des étoiles du CHU de Québec, Canada. This support is greatly acknowledged. The sources of funding were not involved in study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

Marzia A. Cremona acknowledges the support of the Fonds de recherche du Québec Health (FRQS), Canada.

Annex A. RMSE and PDE by patient

See Figs. 13–16.

Annex B. Prediction curves

See Figs. 17–21.

Annex C. Clarke error grids

See Figs. 22–26.

Annex D. Results for female and male patients

See Table 10.

Annex E. Characteristics comparison of the 79 selected patients and all the T1DEXI patients

See Tables 11 and 12.

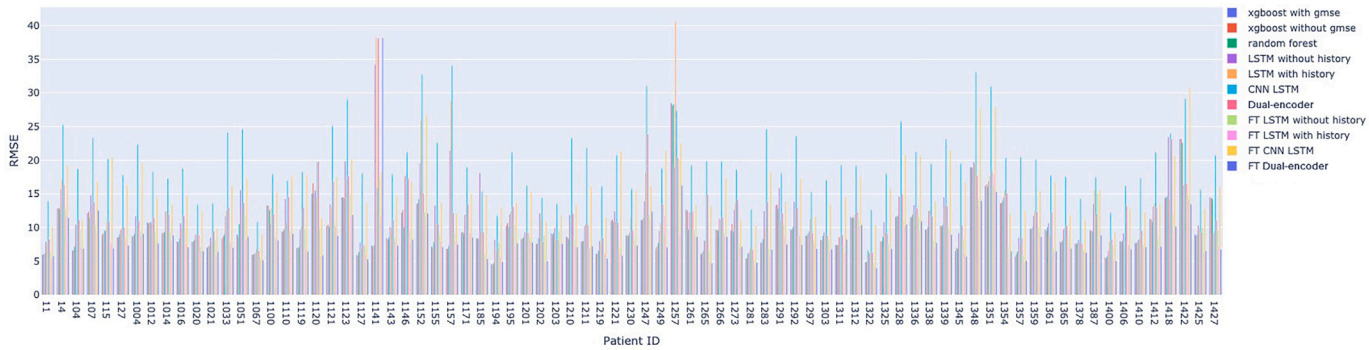


Fig. 13. RMSE (*in mg/dL*) per model and patient with a 10-min prediction horizon.

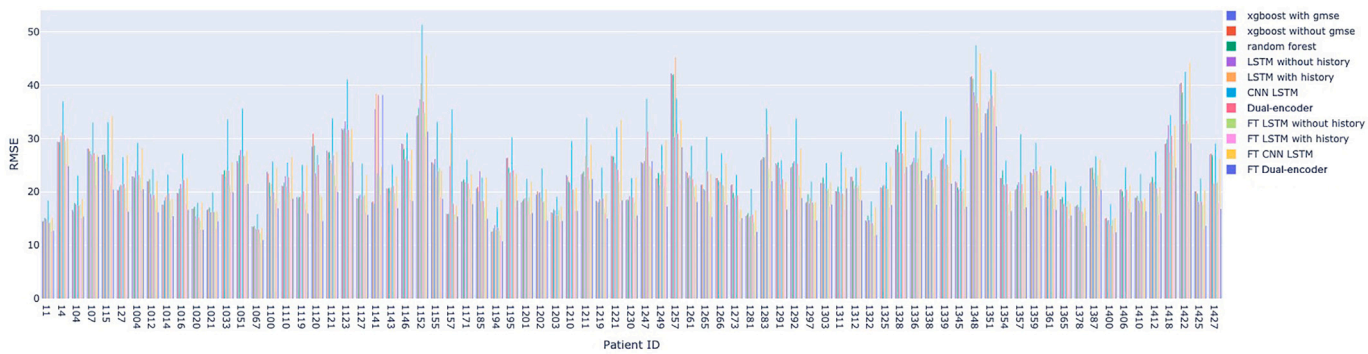


Fig. 14. RMSE (*in mg/dL*) per model and patient with a 30-min prediction horizon.

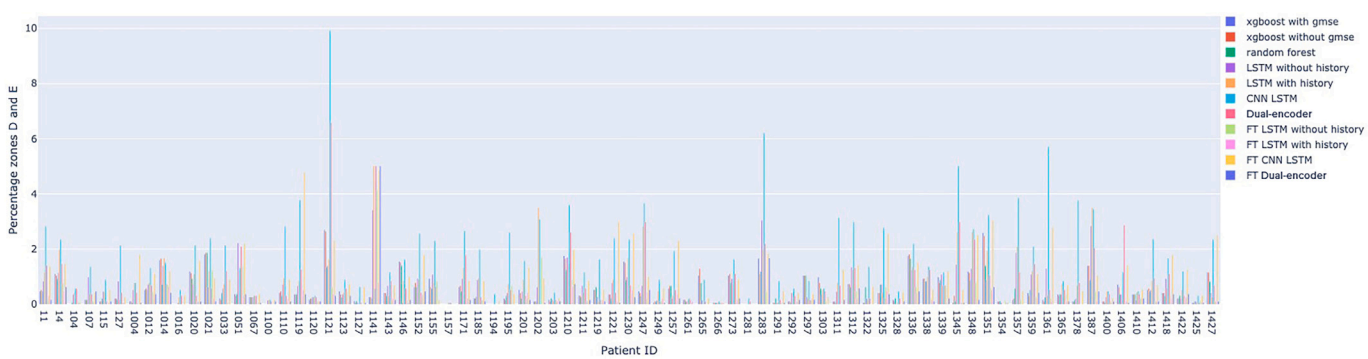


Fig. 15. PDE (*in percentage*) per model and patient with a 10-min prediction horizon.

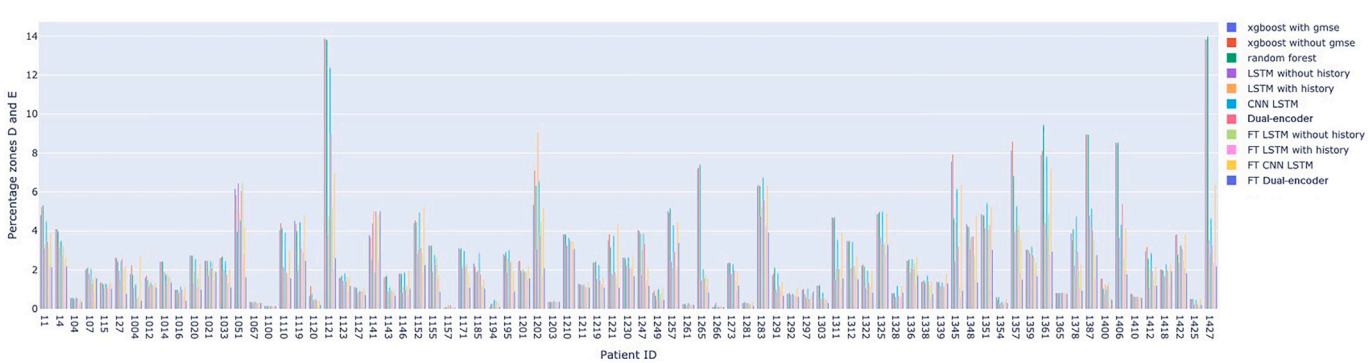


Fig. 16. PDE (*in percentage*) per model and patient with a 30-min prediction horizon.

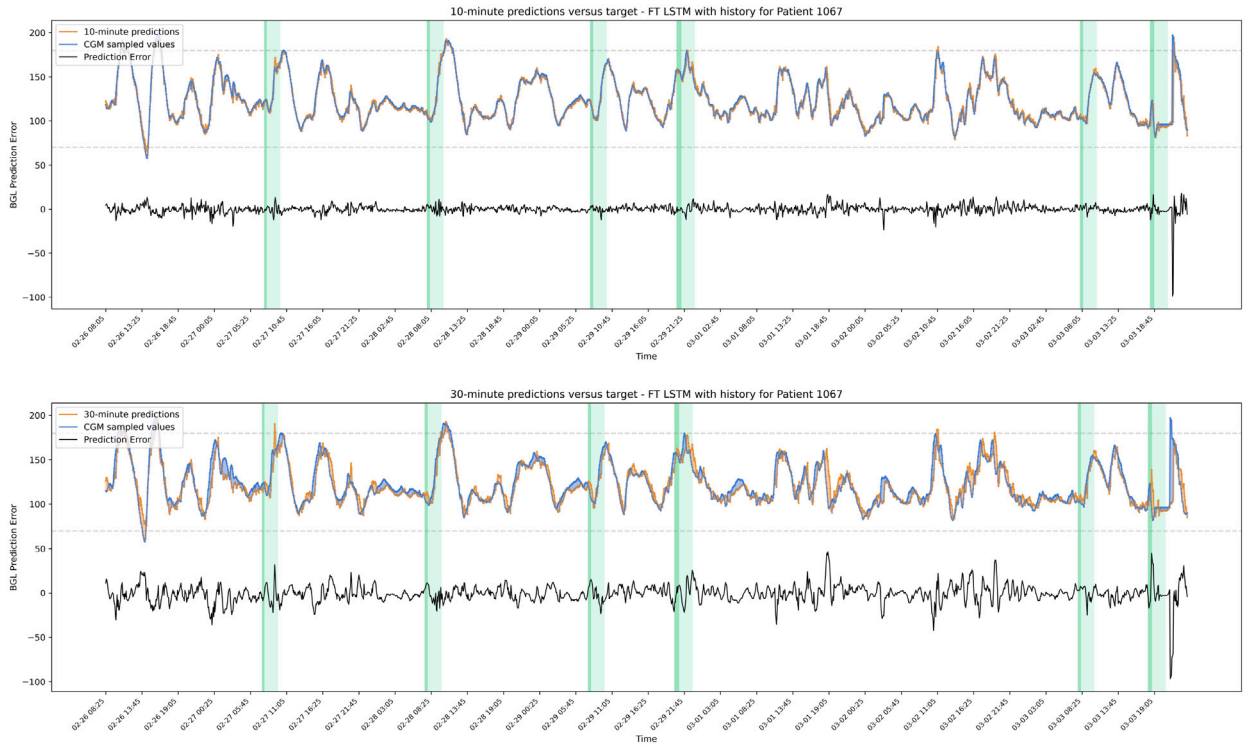


Fig. 17. Predictions and errors of Patient 1067 (*Lowest RMSE*).

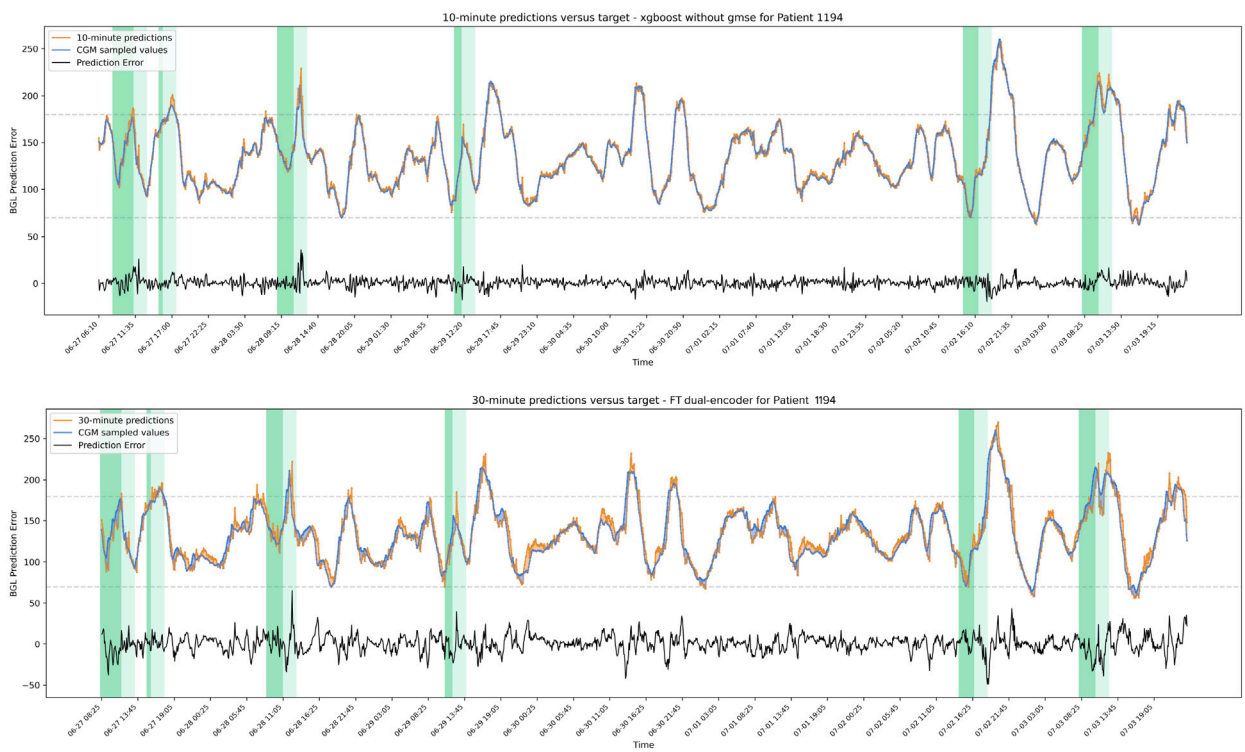


Fig. 18. Predictions and errors of Patient 1194 (*2nd lowest RMSE*).

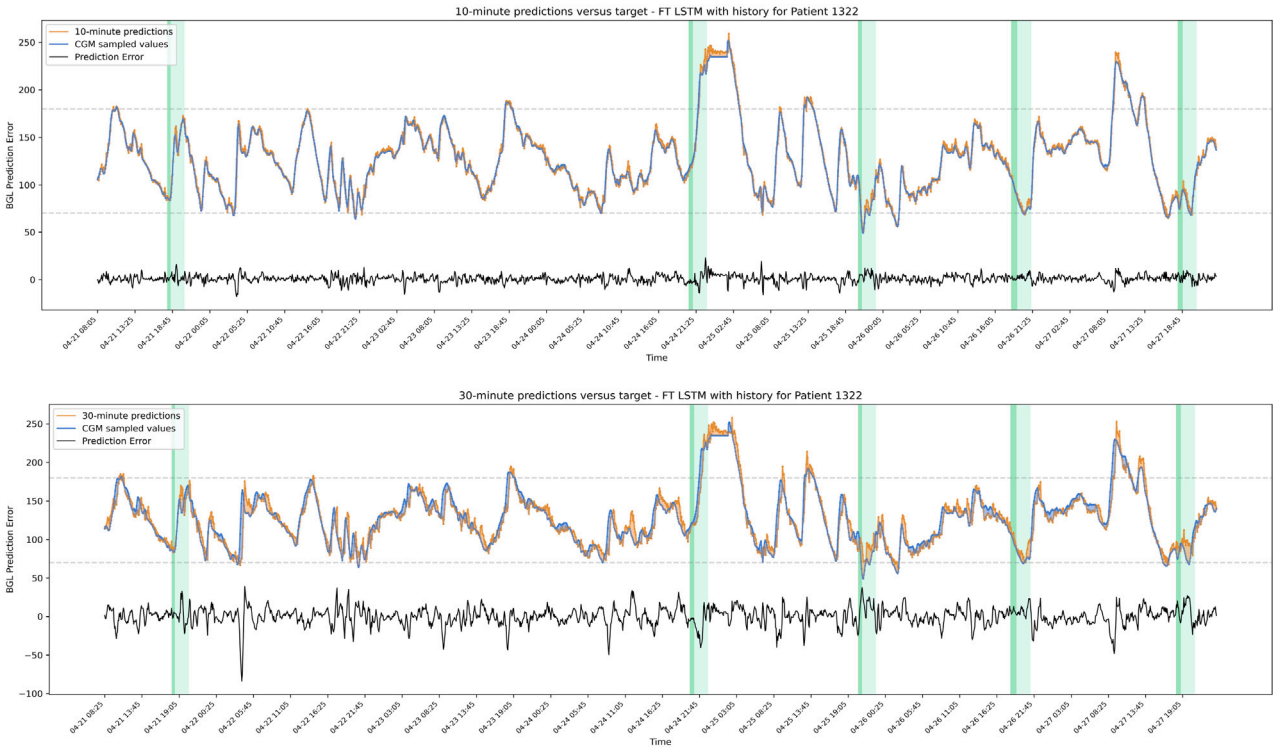


Fig. 19. Predictions and errors of Patient 1322 (*3rd lowest RMSE*).

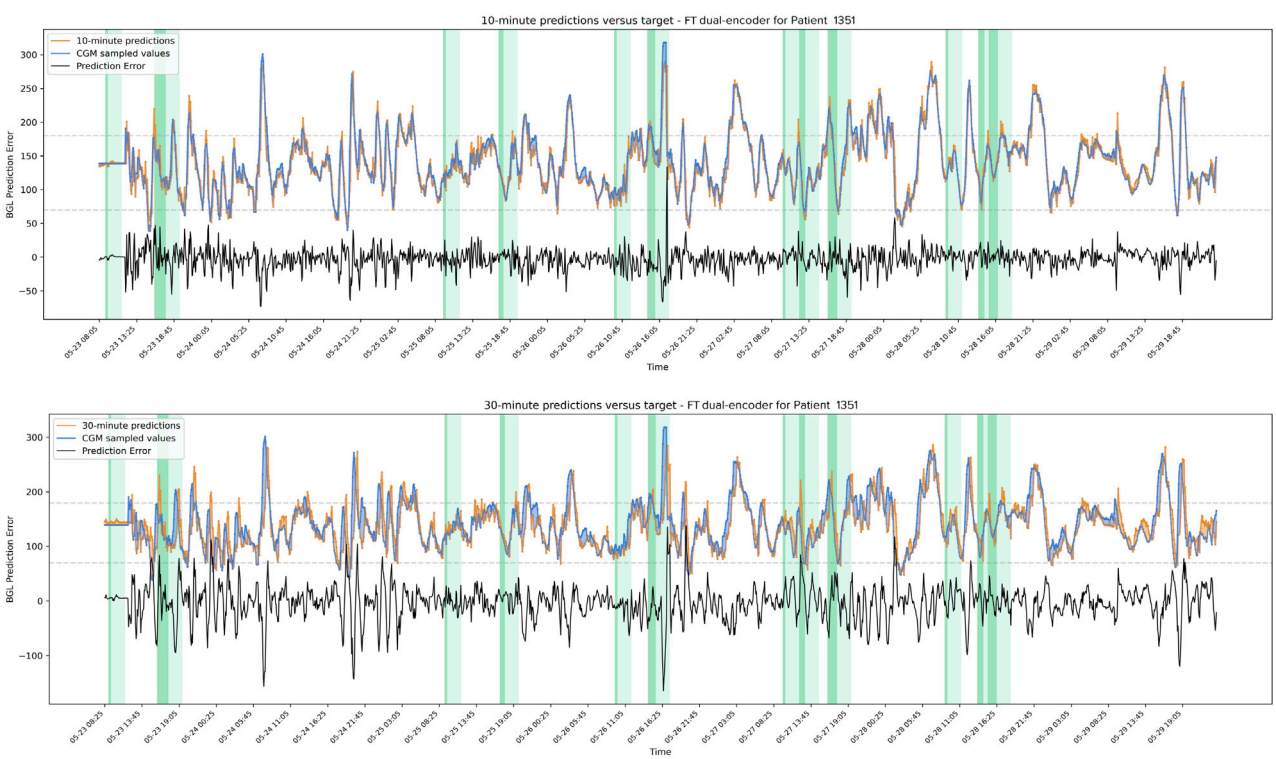


Fig. 20. Predictions and error of Patient 1351 (*Largest RMSE*).

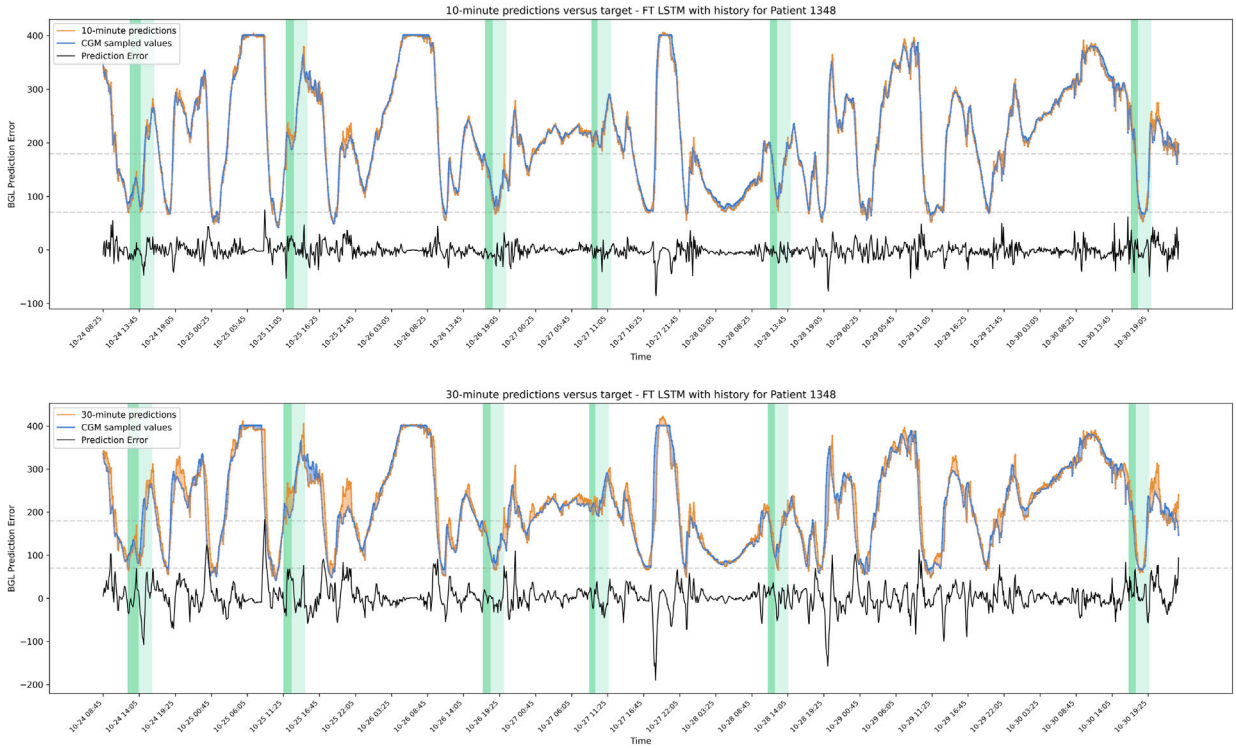


Fig. 21. Predictions and error of Patient 1348, (*2nd largest RMSE*).

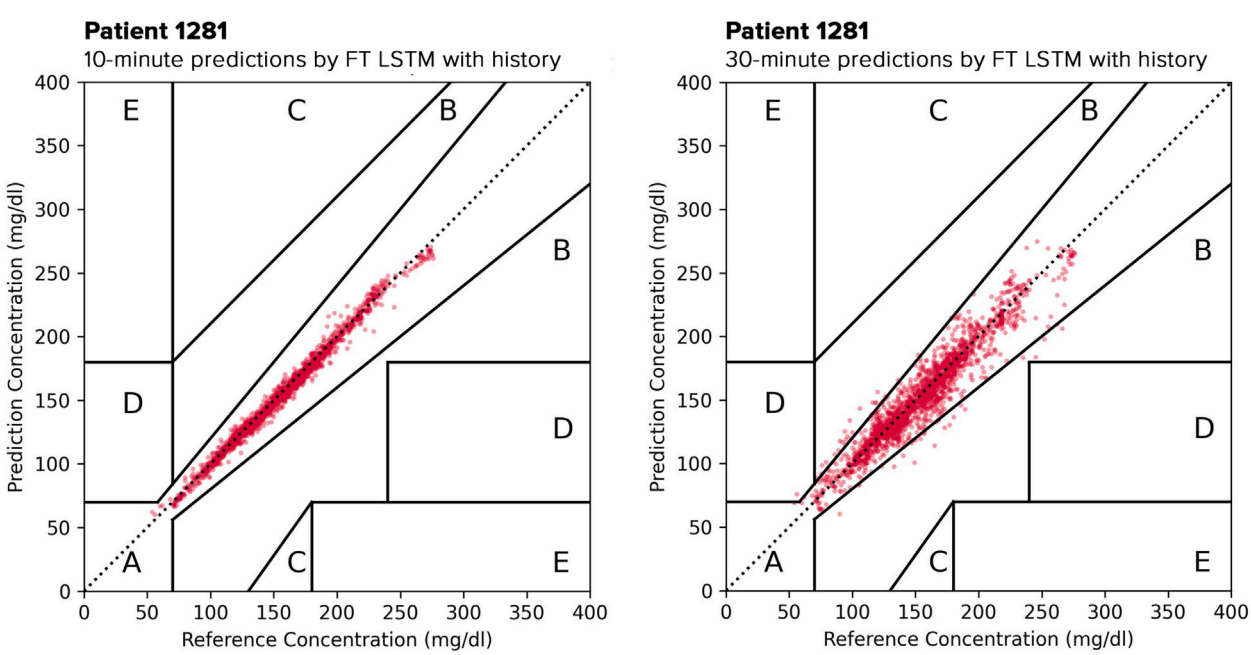


Fig. 22. Clarke Error Grid of Patient 1281 (*Lowest PDE*).

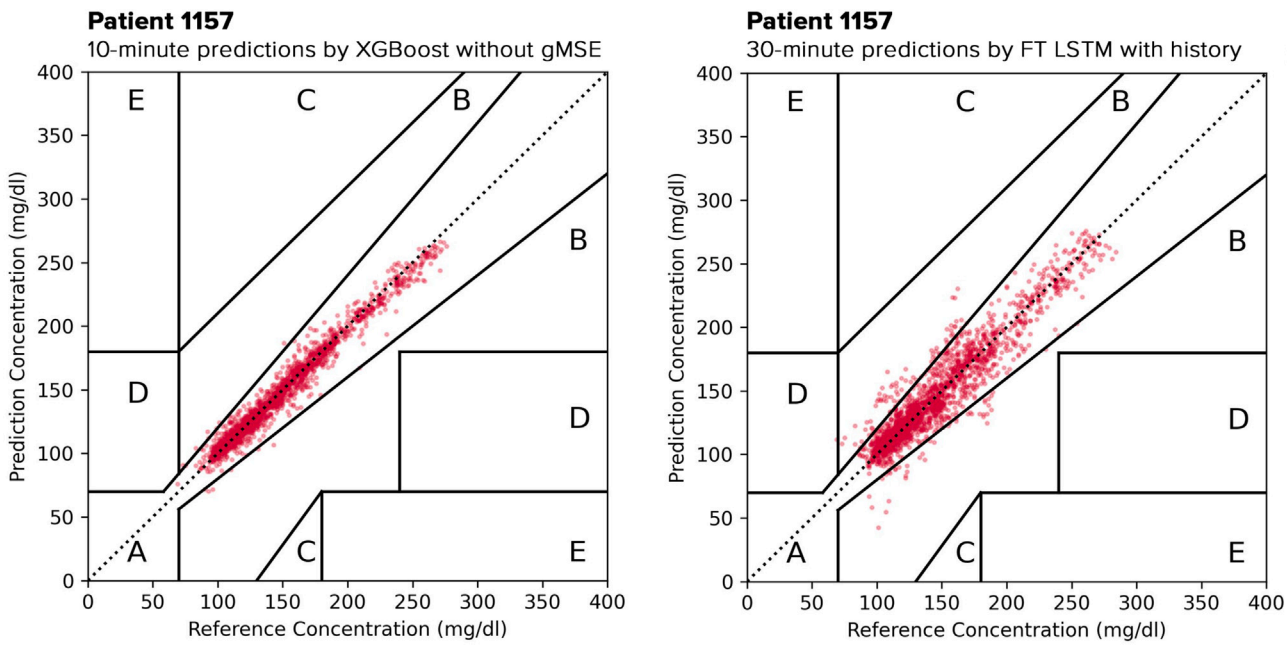


Fig. 23. Clarke Error Grid of Patient 1157 (*2nd lowest PDE*).

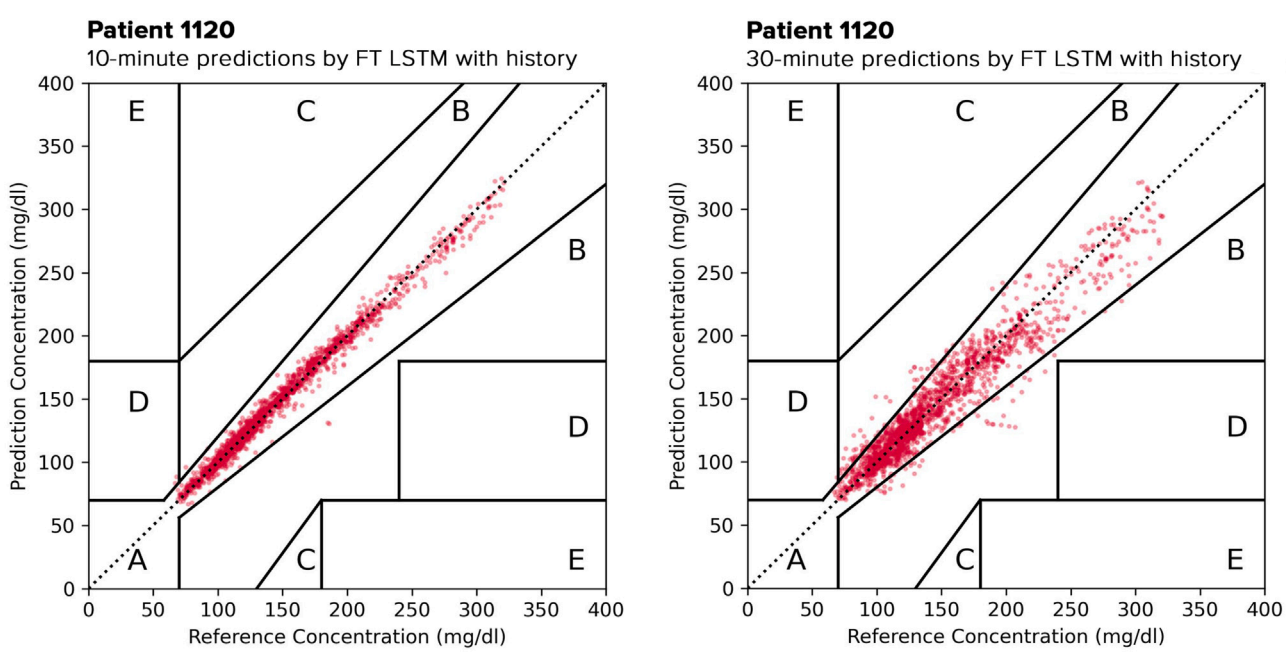


Fig. 24. Clarke Error Grid of Patient 1120 (*3rd lowest PDE*).

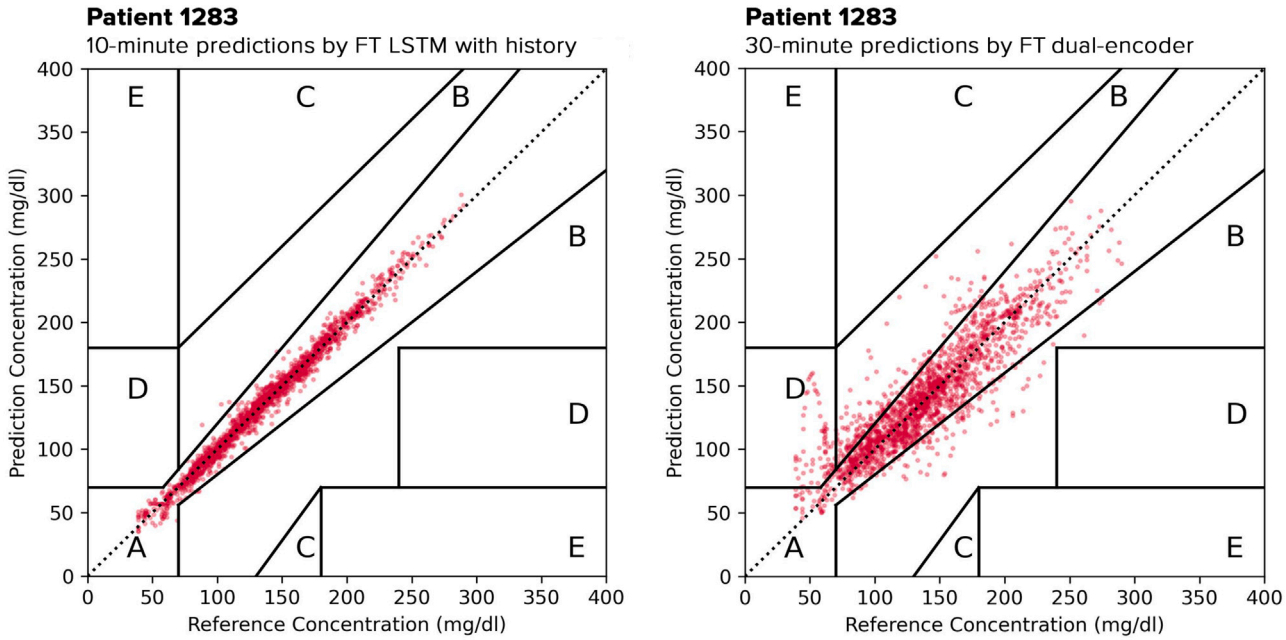


Fig. 25. Clarke Error Grid of Patient 1283 (*Largest PDE*).

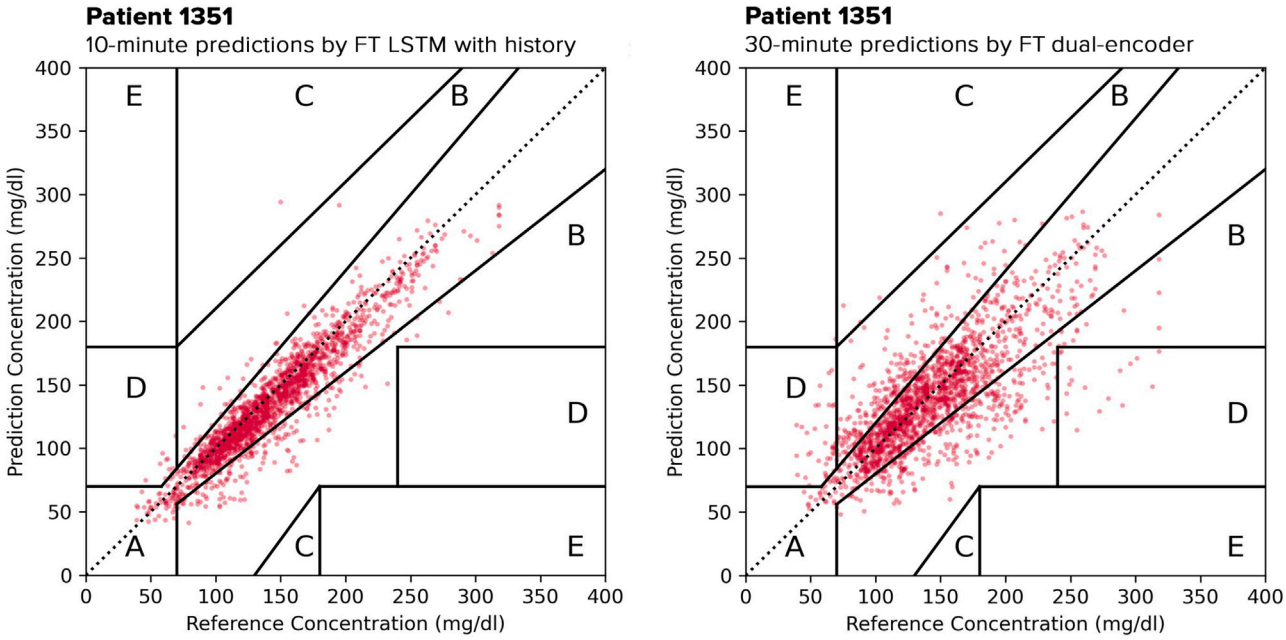


Fig. 26. Clarke Error Grid of Patient 1351 (*2nd largest PDE*).

Table 10
RMSE and PDE metrics for male versus female patients with their best RMSE and PDE models.

Metric	10 min	20 min	30 min
Best RMSE models			
Min RMSE Female	3.95	8.14	10.48
Min RMSE Male	4.81	7.79	10.75
Mean RMSE Female	7.63	13.23	18.04
Mean RMSE Male	6.98	12.31	16.85
Max RMSE Female	15.31	24.65	32.31
Max RMSE Male	11.72	21.47	30.37
Best PDE models			
Min PDE Female	0.00%	0.00%	0.05%
Min PDE Male	0.00%	0.05%	0.05%
Mean PDE Female	0.12%	0.56%	1.16%
Mean PDE Male	0.11%	0.47%	1.00%
Max PDE Female	0.63%	1.83%	3.92%
Max PDE Male	0.42%	1.14%	2.78%

Table 11
Comparison between the 79 selected patients and the complete T1DEXI dataset (497 patients) for numerical characteristics.

Feature	All T1DEXI patients					Selected Patients				
	Min	Q1	Median	Q3	Max	Min	Q1	Median	Q3	Max
Age	18	25	34	46	70	18	25	34	45.25	68
Years since diagnosis	1	11	16	24	66	2	11	16	26	53
BMI (kg/m^2)	18.25	22.71	24.43	26.98	48.47	19.2	23.1	25.06	28.82	35.43
HbA1c range	4.8	6.1	6.5	7.1	10.0	5.3	6.1	6.4	7.1	8.3
Exercise duration	30	726	1034	1480	6329	612	959.75	1171	1530.75	5275
TIR (%)	7.55	66.35	77.29	85.24	99.05	37.8	69.99	80.37	86.86	98.25
TBR (%)	0.0	1.15	2.24	4.28	20.31	0.15	1.32	2.12	3.69	17.41
TAR (%)	0.0	10.71	19.51	30.05	92.45	1.47	9.93	15.86	26.29	60.58
CV (%)	16.53	29.12	32.92	36.77	57.16	19.47	29.19	32.77	36.38	48.11

Table 12
Comparison between the 79 selected patients and the complete T1DEXI dataset (497 patients) for categorical characteristics.

Category	All T1DEXI patients	Selected patients
Female	72.22%	75.0%
Male	27.78%	25.0%
Aerobic	32.69%	31.25%
Interval training	34.19%	40.0%
Resistance training	33.12%	28.75%
Patients using AID	81.2%	100.0%
Patients using MDI	18.8%	0.0%

Data availability

The authors confirm that the data supporting the findings of this review are available on request from Vivli Inc.

References

[1] A. Bertachi, C.M. Ramkissoon, J. Bondia, J. Vehí, Automated blood glucose control in type 1 diabetes: A review of progress and challenges, Endocrinol. Diabetes Nutr. (Engl. Ed.) 65 (3) (2018) 172–181.

[2] V. Felizardo, N.M. Garcia, N. Pombo, I. Megdiche, Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review, Artif. Intell. Med. 118 (2021) 102120.

[3] A. Bertachi, C. Viñals, L. Biagi, I. Contreras, J. Vehí, I. Conget, M. Giménez, Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor, Sensors 20 (6) (2020) 1705.

[4] D. Montt-Blanchard, R. Sánchez, K. Dubois-Camacho, J. Leppe, M.T. Onetto, Hypoglycemia and glycemic variability of people with type 1 diabetes with lower and higher physical activity loads in free-living conditions using continuous subcutaneous insulin infusion with predictive low-glucose suspend system, BMJ Open Diabetes Res. Care 11 (2) (2023) e003082.

[5] J.-F. Yale, B. Paty, P.A. Senior, Diabetes Canada Clinical Practice Guidelines Expert Committee, et al., Hypoglycemia, Can. J. Diabetes 42 (2018) S104–S108.

[6] T. Zhu, K. Li, P. Herrero, P. Georgiou, Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning, IEEE Trans. Biomed. Eng. 70 (1) (2022) 193–204.

[7] S. Bergford, M.C. Riddell, P.G. Jacobs, Z. Li, R.L. Gal, M.A. Clements, F.J. Doyle, C.K. Martin, S.R. Patton, J.R. Castle, et al., The type 1 diabetes and exercise initiative: Predicting hypoglycemia risk during exercise for participants with type 1 diabetes using repeated measures random forest, Diabetes Technol. Ther. 25 (9) (2023) 602–611.

[8] M. Syafrudin, G. Alfian, N.L. Fitriyani, T. Hadibarata, J. Rhee, M. Anshari, Future glycemic events prediction model based on artificial neural network, in: 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT, IEEE, 2022, pp. 151–155.

[9] G. Alfian, M. Syafrudin, M. Anshari, F. Benes, F.T.D. Atmaji, I. Fahrurrozi, A.F. Hidayatullah, J. Rhee, Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features, Biocybern. Biomed. Eng. 40 (4) (2020) 1586–1599.

[10] C. Duckworth, M.J. Guy, A. Kumaran, A.A. O'Kane, A. Ayobi, A. Chapman, P. Marshall, M. Boniface, Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations, J. Diabetes Sci. Technol. 18 (1) (2024) 113–123.

[11] J. Yang, L. Li, Y. Shi, X. Xie, An ARIMA model with adaptive orders for predicting blood glucose concentrations and hypoglycemia, IEEE J. Biomed. Heal. Inform. 23 (3) (2018) 1251–1260.

[12] F. Prendin, S. Del Favero, M. Vettoretti, G. Sparacino, A. Facchinetti, Forecasting of glucose levels and hypoglycemic events: head-to-head comparison of linear and nonlinear data-driven algorithms based on continuous glucose monitoring data only, Sensors 21 (5) (2021) 1647.

[13] S. Faccioli, F. Prendin, A. Facchinetti, G. Sparacino, S. Del Favero, Combined use of glucose-specific model identification and alarm strategy based on prediction-funnel to improve online forecasting of hypoglycemic events, J. Diabetes Sci. Technol. 17 (5) (2023) 1295–1303.

[14] C. Toffanin, E.M. Aiello, C. Cobelli, L. Magni, Hypoglycemia prevention via personalized glucose-insulin models identified in free-living conditions, J. Diabetes Sci. Technol. 13 (6) (2019) 1008–1016.

[15] H. Witte, C. Nakas, L. Bally, A.B. Leichtle, et al., Machine learning prediction of hypoglycemia and hyperglycemia from electronic health records: algorithm development and validation, JMIR Form. Res. 6 (7) (2022) e36176.

[16] F. Iacono, L. Magni, C. Toffanin, Personalized LSTM-based alarm systems for hypoglycemia and hyperglycemia prevention, Biomed. Signal Process. Control. 86 (2023) 105167.

[17] S.-M. Lee, D.-Y. Kim, J. Woo, Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia, IEEE J. Biomed. Heal. Inform. 27 (3) (2023) 1600–1611.

[18] A.Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, G. Hartvigsen, Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes, Artif. Intell. Med. 98 (2019) 109–134.

[19] J. Xie, Q. Wang, Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models, IEEE Trans. Biomed. Eng. 67 (11) (2020) 3101–3124.

[20] M. Sevil, M. Rashid, I. Hajizadeh, M. Park, L. Quinn, A. Cinar, Physical activity and psychological stress detection and assessment of their effects on glucose concentration predictions in diabetes management, IEEE Trans. Biomed. Eng. 68 (7) (2021) 2251–2260.

[21] N. Hobbs, I. Hajizadeh, M. Rashid, K. Turksoy, M. Breton, A. Cinar, Improving glucose prediction accuracy in physically active adolescents with type 1 diabetes, J. Diabetes Sci. Technol. 13 (4) (2019) 718–727.

[22] N.S. Tyler, C. Mosquera-Lopez, G.M. Young, J. El Youssef, J.R. Castle, P.G. Jacobs, Quantifying the impact of physical activity on future glucose trends using machine learning, Iscience 25 (3) (2022).

[23] W.L. Clarke, D. Cox, L.A. Gonder-Frederick, W. Carter, S.L. Pohl, Evaluating clinical accuracy of systems for self-monitoring of blood glucose, Diabetes Care 10 (5) (1987) 622–628.

[24] W.L. Clarke, The original Clarke error grid analysis (CEGA), Diabetes Technol. Ther. 7 (5) (2005) 776–779.

[25] Jaeb Center for Health Research, Type 1 diabetes exercise initiative: The effect of exercise on glycemic control in type 1 diabetes study, 2020, URL: https://www.jaeb.org/projects/. (Accessed 21 March 2024).

[26] D. Kalita, K.B. Mirza, LS-GRUNet: glucose forecasting using deep learning for closed-loop diabetes management, in: 2022 IEEE 7th International Conference for Convergence in Technology, I2CT, IEEE, 2022, pp. 1–6.

[27] J.Q. Toledo-Marín, T. Ali, T. van Rooij, M. Görges, W.W. Wasserman, Prediction of blood risk score in diabetes using deep neural networks, J. Clin. Med. 12 (4) (2023) 1695.

[28] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, J. Wiens, Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1387–1395.

[29] R.A. Karim, I. Vassányi, I. Kósa, After-meal blood glucose level prediction using an absorption model for neural network training, Comput. Biol. Med. 125 (2020) 103956.

[30] G. Annuzzi, A. Apicella, P. Arpaia, L. Bozzetto, S. Criscuolo, E. De Benedetto, M. Pesola, R. Prevete, E. Vallefuoco, Impact of nutritional factors in blood glucose prediction in type 1 diabetes through machine learning, IEEE Access 11 (2023) 17104–17115.

[31] C. Midroni, P.J. Leimbigler, G. Baruah, M. Kolla, A.J. Whitehead, Y. Fossat, Predicting glycemia in type 1 diabetes patients: experiments with XGBoost, Heart 60 (90) (2018) 120.

[32] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, C. Cobelli, Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration, IEEE Trans. Biomed. Eng. 59 (6) (2012) 1550–1560.

[33] M. Muñoz-Organero, P. Queipo-Álvarez, B. García Gutiérrez, Learning carbohydrate digestion and insulin absorption curves using blood glucose level prediction and deep learning models, Sensors 21 (14) (2021) 4926.

[34] T. Yang, X. Yu, N. Ma, R. Wu, H. Li, An autonomous channel deep learning framework for blood glucose prediction, Appl. Soft Comput. 120 (2022) 108636.

[35] T. Zhu, C. Uduku, K. Li, P. Herrero, N. Oliver, P. Georgiou, Enhancing self-management in type 1 diabetes with wearables and deep learning, Npj Digit. Med. 5 (1) (2022) 78.

[36] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018, arXiv preprint arXiv:1803.01271.

[37] M. Armandpour, B. Kidd, Y. Du, J.Z. Huang, Deep personalized glucose level forecasting using attention-based recurrent neural networks, in: 2021 International Joint Conference on Neural Networks, IJCNN, IEEE, 2021, pp. 1–8.

[38] C. Liu, J. Vehí, P. Avari, M. Reddy, N. Oliver, P. Georgiou, P. Herrero, Long-term glucose forecasting using a physiological model and deconvolution of the continuous glucose monitoring signal, Sensors 19 (19) (2019) 4338.

[39] M.D. Breton, S.A. Brown, C.H. Karvetski, L. Kollar, K.A. Topchyan, S.M. Anderson, B.P. Kovatchev, Adding heart rate signal to a control-to-range artificial pancreas system improves the protection against hypoglycemia during exercise in type 1 diabetes, Diabetes Technol. Ther. 16 (8) (2014) 506–511.

[40] Y.C. Kudva, R.E. Carter, C. Cobelli, R. Basu, A. Basu, Closed-loop artificial pancreas systems: physiological input to enhance next-generation devices, Diabetes Care 37 (5) (2014) 1184–1190.

[41] V. Dadlani, J.A. Levine, S.K. McCrady-Spitzer, E. Dassau, Y.C. Kudva, Physical activity capture technology with potential for incorporation into closed-loop control for type 1 diabetes, J. Diabetes Sci. Technol. 9 (6) (2015) 1208–1216.

[42] A. Bertachi, L. Biagi, I. Contreras, N. Luo, J. Vehí, Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks, in: KHD@ IJCAI, 2018, pp. 85–90.

[43] I. Contreras, A. Bertachi, L. Biagi, J. Vehí, S. Oviedo, Using grammatical evolution to generate short-term blood glucose prediction models, in: KHD@ IJCAI, 2018, pp. 91–96.

[44] S.L. Cichosz, T. Kronborg, M.H. Jensen, O. Hejlesen, Penalty weighted glucose prediction models could lead to better clinically usage, Comput. Biol. Med. 138 (2021) 104865.

[45] H.N. Mhaskar, S.V. Pereverzyev, M.D. Van der Walt, A deep learning approach to diabetic blood glucose prediction, Front. Appl. Math. Stat. 3 (2017) 14.

[46] S. Langarica, M. Rodriguez-Fernandez, F. Nunez, F.J. Doyle III, A meta-learning approach to personalized blood glucose prediction in type 1 diabetes, Control Eng. Pract. 135 (2023) 105498.

[47] B. De Paoli, F. D'Antoni, M. Merone, S. Pieralice, V. Piemonte, P. Pozzilli, Blood glucose level forecasting on type-1-diabetes subjects during physical activity: A comparative analysis of different learning techniques, Bioengineering 8 (6) (2021) 72.

[48] A. Parcerisas, I. Contreras, A. Delecourt, A. Bertachi, A. Beneyto, I. Conget, C. Viñals, M. Giménez, J. Vehi, A machine learning approach to minimize nocturnal hypoglycemic events in type 1 diabetic patients under multiple doses of insulin, Sensors 22 (4) (2022) 1665.

[49] S. Sivananthan, V. Naumova, C.D. Man, A. Facchinetti, E. Renard, C. Cobelli, S.V. Pereverzyev, Assessment of blood glucose predictors: the prediction-error grid analysis, Diabetes Technol. Ther. 13 (8) (2011) 787–796.

[50] M. Wadghiri, A. Idri, T. El Idrissi, H. Hakkoum, Ensemble blood glucose prediction in diabetes mellitus: A review, Comput. Biol. Med. 147 (2022) 105674.

[51] S. Del Favero, A. Facchinetti, C. Cobelli, A glucose-specific metric to assess predictors and identify models, IEEE Trans. Biomed. Eng. 59 (5) (2012) 1281–1290.

[52] G. Alfian, M. Syafrudin, J. Rhee, M. Anshari, M. Mustakim, I. Fahrurrozi, Blood glucose prediction model for type 1 diabetes based on extreme gradient boosting, in: IOP Conference Series: Materials Science and Engineering, Vol. 803, IOP Publishing, 2020, 012012.

[53] J. Xie, Q. Wang, Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge, in: KHD@ IJCAI, Vol. 10, 2018.

[54] K. Zarkogianni, K. Mitsis, E. Litsa, M.-T. Arredondo, G. Fico, A. Fioravanti, K.S. Nikita, Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring, Med. Biol. Eng. Comput. 53 (2015) 1333–1343.

[55] G. Aleppo, K.J. Ruedy, T.D. Riddlesworth, D.F. Kruger, A.L. Peters, I. Hirsch, R.M. Bergenstal, E. Toschi, A.J. Ahmann, V.N. Shah, et al., Replace-BG: a randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes, Diabetes Care 40 (4) (2017) 538–545.

[56] S.L. Cichosz, O. Hejlesen, M.H. Jensen, Identification of individuals with diabetes who are eligible for continuous glucose monitoring forecasting, Diabetes Metab. Syndrome: Clin. Res. Rev. 18 (2) (2024) 102972.

[57] R.E. Pratley, L.G. Kanapka, M.R. Rickels, A. Ahmann, G. Aleppo, R. Beck, A. Bhargava, B.W. Bode, A. Carlson, N.S. Chaytor, et al., Effect of continuous glucose monitoring on hypoglycemia in older adults with type 1 diabetes: a randomized clinical trial, Jama 323 (23) (2020) 2397–2406.

[58] C. Marling, R. Bunescu, The OhioT1DM dataset for blood glucose level prediction: Update 2020, in: CEUR Workshop Proceedings, Vol. 2675, NIH Public Access, 2020, p. 71, URL: http://smarthealth.cs.ohio.edu/bglp/OhioT1DM-dataset-paper.pdf.

[59] C. Marling, R. Bunescu, The OhioT1DM dataset for blood glucose level prediction, in: The 3rd International Workshop on Knowledge Discovery in Healthcare Data, Stockholm, Sweden, CEUR Workshop Proceedings, Vol. 2675, NIH Public Access, 2018, p. 71, URL: http://smarthealth.cs.ohio.edu/bglp/OhioT1DM-dataset-paper.pdf.

[60] K. Li, J. Daniels, C. Liu, P. Herrero, P. Georgiou, Convolutional recurrent neural networks for glucose prediction, IEEE J. Biomed. Heal. Inform. 24 (2) (2019) 603–613.

[61] X. Shi, Y.D. Wong, M.Z.-F. Li, C. Palanisamy, C. Chai, A feature learning approach based on XGBoost for driving assessment and risk prediction, Accid. Anal. Prev. 129 (2019) 170–179.

[62] Y. Deng, K. Arao, C.S. Mantzoros, G.E. Karniadakis, Patient-specific deep offline artificial pancreas for blood glucose regulation in type 1 diabetes, 2022, pp. 1–26, BioRxiv, 2022-2010.

[63] B.J. Zou, M.E. Levine, D.P. Zaharieva, R. Johari, E. Fox, Hybrid^2 neural ODE causal modeling and an application to glycemic response, in: Forty-First International Conference on Machine Learning, ICML, 2024, pp. 62934—-62963.

[64] A. Yazdani, L. Lu, M. Raissi, G.E. Karniadakis, Systems biology informed deep learning for inferring parameters and hidden dynamics, PLoS Comput. Biol. 16 (11) (2020) e1007575.

[65] A. Roy, R.S. Parker, Dynamic modeling of exercise effects on plasma glucose and insulin levels, 2007.