Lightweight Sequential Transformers for Blood Glucose Level Prediction in Type-1 Diabetes

Mirko Paolo Barbato, Giorgia Rigamonti, Davide Marelli, Paolo Napoletano, *Senior Member, IEEE*

***Abstract*— Type 1 Diabetes (T1D) affects millions worldwide, requiring continuous monitoring to prevent severe hypo- and hyperglycemic events. While continuous glucose monitoring has improved blood glucose management, deploying predictive models on wearable devices remains challenging due to computational and memory constraints. To address this, we propose a novel Lightweight Sequential Transformer model designed for blood glucose prediction in T1D. By integrating the strengths of Transformers' attention mechanisms and the sequential processing of recurrent neural networks, our architecture captures long-term dependencies while maintaining computational efficiency. The model is optimized for deployment on resource-constrained edge devices and incorporates a balanced loss function to handle the inherent data imbalance in hypo- and hyperglycemic events. Experiments on two benchmark datasets, OhioT1DM and DiaTrend, demonstrate that the proposed model outperforms state-of-the-art methods in predicting glucose levels and detecting adverse events. This work fills the gap between high-performance modeling and practical deployment, providing a reliable and efficient T1D management solution.**

***Index Terms*— Blood glucose prediction, continuous glucose monitoring, sequential deep transformers, lightweight deep learning.**

I. Introduction

Type 1 Diabetes (T1D) [1] is a chronic autoimmune condition requiring lifelong blood glucose concentration (BGC) monitoring to prevent life-threatening complications such as hypoglycemia (BGC below 70 mg/dL [2]) and hyperglycemia (BGC above 180 mg/dL [3]). Continuous glucose monitoring (CGM) [4] technology has transformed T1D management by providing real-time tracking of glucose levels, enabling timely interventions and improving patient outcomes. However, predicting future glucose levels accurately remains a critical challenge, as it directly impacts insulin administration and overall disease management.

State-of-the-art methods for BGC prediction leverage complex deep-learning architectures, such as Transformers and Recurrent Neural Networks (RNNs) [5], to capture temporal dependencies and achieve high predictive accuracy. Despite their effectiveness, these methods are computationally intensive and have large memory footprints, limiting their deployment on resource-constrained wearable devices [6], [7]. Cloud-based solutions mitigate computational demands but introduce privacy concerns, latency issues, and reliance on network connectivity [8], [9], reducing their reliability for real-world applications. Furthermore, existing models often struggle with the inherent imbalance in glucose datasets, where adverse events (e.g., hypoglycemia) occur less frequently than normal glucose levels, leading to suboptimal performance in detecting critical events.

To address these limitations, this paper introduces a novel *Lightweight Sequential Transformer* model for blood glucose prediction in individuals with T1D. The proposed architecture combines the attention mechanism of Transformers with the sequential processing capability of RNNs, enabling it to effectively capture both long-term and short-term temporal dependencies in BGC data. By employing a compact neural network design, the model is optimized for deployment on wearable devices with limited computational resources. Additionally, a novel balanced loss function is introduced to improve the detection of hypo- and hyperglycemic events, addressing the issue of dataset imbalance. Experiments on OhioT1DM [10] and DiaTrend [11] show that the proposed model outperforms state-of-the-art baselines in predictive accuracy and adverse event detection. Beyond supporting single- and multimodal inputs, its robustness is validated through ablation studies on different training configurations and feature combinations, as well as preliminary investigations on personalization, complementing the main focus on a generalized setting. Code of the proposed method is available at: `https://github.com/unimib-islab/Diabetes_Sequential_transformer`.

This work proposes a methodology to fill the gap between high-performance blood glucose prediction and practical deployment, aiming to provide reliable and efficient solutions in edge computing systems. The proposed approach is a step forward in improving T1D management, thus enabling real-world adoption of advanced predictive models, thanks to its performance and adaptability.

II. Related work

Over time, machine and deep learning techniques have proven to be accurate and reliable tools for estimating BGC [5]. Various approaches for BGC prediction have been explored, ranging from traditional statistical models, such as the Autoregressive Integrated Moving Average (ARIMA) [12] and the Unobserved Components Model (UCM) [13], to more advanced machine learning algorithms, including Random Forest and eXtreme Gradient Boosting (XGBoost) [14]. While

The authors are with the Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano 20126, Italy (e-mail: mirko.barbato@unimib.it; giorgia.rigamonti@unimib.it; davide.marelli@unimib.it; paolo.napoletano@unimib.it).

these methods laid the groundwork for BGC prediction, they cannot often model complex temporal dependencies and physiological interactions, limiting their performance in dynamic, real-world scenarios.

More recently, deep learning approaches have significantly advanced the robustness and accuracy of BGC prediction [15], [16]. Architectures such as Recurrent Neural Networks (RNNs) [17], [18], Convolutional RNNs (CRNNs) [19], and Long Short-Term Memory networks (LSTMs) [20] have been widely applied, but their ability to capture long-term dependencies is often limited by vanishing gradients and the sequential nature of their computation.

To address these limitations, Transformer-based models have gained attention for their self-attention mechanisms, which facilitate parallel processing and improve the modeling of long-range dependencies. Several Transformer variants have been explored for glucose prediction and related time-series tasks [21]–[25]. Among them, general-purpose architectures such as Informer [26] and TimesNet [27] have demonstrated strong performance when adapted to glucose forecasting [21]. Domain-specific designs, such as Gluformer [24], incorporate additional features like uncertainty quantification to enhance clinical usability. Hybrid approaches (e.g., Transformer-LSTM [22]) have also been proposed, combining the strengths of sequential models and attention mechanisms to improve prediction accuracy.

Existing state-of-the-art methods can broadly be divided into single-modality and multimodal frameworks. Single-modality approaches rely solely on continuous glucose monitoring (CGM) data, where Transformer-based solutions such as Informer and TimesNet highlight the scalability of attention mechanisms in time-series modeling. Multimodal methods, by contrast, integrate additional inputs such as insulin delivery, carbohydrate intake, and physiological parameters. Recent examples include BG-BERT [28], which employs self-supervised pretraining with a BERT-like masked auto-encoder, and Bi-GRU [29], which provides a parameter-efficient alternative for multimodal modeling.

Despite these advances, efficiency remains an underexplored dimension in BGC prediction. Most Transformer-based methods emphasize accuracy but rely on large parameter counts and high computational overhead, which limits their deployment in real-world scenarios such as edge devices or resource-constrained clinical settings.

In summary, while Transformer architectures have already shown promise for both single- and multimodal BGC prediction, existing solutions often emphasize performance at the expense of efficiency. The proposed Lightweight Sequential Transformer is designed to fill this gap: it combines the efficiency of sequential processing with the representational power of self-attention, enabling accurate predictions with substantially lower computational requirements. This design makes it particularly suitable for real-world applications that demand lightweight, resource-aware models without sacrificing predictive accuracy.

III. Methodology

In this section, we outline the problem addressed and the solutions proposed in this work. Specifically, we detail both the designed architecture and the training strategy employed to enable the model to effectively predict BGC and recognize adverse events.

A. Problem Formulation

A glucose forecaster is a computational model $\mathcal{M}$ that predicts future glucose levels given a sequence of observed data. The input of the model $\mathcal{M}$ is a sequence of T glucose observations $\mathbf{g} = \{g_1, \cdots, g_T\}$, where the length of T depends on the forecasting horizon and usually ranges from 2 to 4 hours [28]. The output is a sequence of L glucose predictions $\hat{\mathbf{g}} = \{\hat{g}_1, \cdots, \hat{g}_L\}$, where L stands for forecasting horizon that usually is 30 or 60 minutes [30]. When available, it is possible to use a multimodal approach by exploiting extra features, such as quantities of basal, bolus, carbs, and other physiological parameters, as input of the model $\mathcal{M}$ to get more accurate predictions. Extra features are usually pre-processed to obtain a sequence of the same length of $\mathbf{g}$ [30].

B. Proposed forecaster model

The proposed Sequential Transformer draws inspiration from the architecture of the classical Transfomers [31] and RNNs [32]. The main idea is to create an architecture that exploits the concept of attention typical of Transformers without losing the notion of sequentiality that characterizes time series, which we consider fundamental for accurately predicting future information. A standard Transformer handles the entire observed sequence all at once, giving the same importance to all the samples over time. However, this approach may not fully capture the temporal dynamics necessary for tasks like glucose level prediction. By contrast, similarly to RNNs, the Sequential Transformer processes the elements of the observed sequence incrementally, one at a time. This strategy emphasizes the most recent observations in the sequence, which are temporally closer to the glucose levels to predict. This approach aligns with the glucose behavior over time, which typically fluctuates at a low frequency due to factors such as meal intake, insulin response, and circadian rhythms [33]. At the same time, the Sequential Transformer retains the attention mechanism (via query, key, and value), which is well-known for outperforming traditional architectures because it allows the model to capture long-range dependencies, handle variable-length sequences effectively, and prioritize relevant information dynamically [31]. Furthermore, the sequential nature of our approach, thanks to the parameter-sharing policy, inherently reduces the number of parameters compared to a standard Transformer, making it more efficient.

The proposed model $\mathcal{M}$, depicted in Figure 1, is designed to handle both single-modality inputs, such as glucose sequences, and multi-modality inputs, including feature sequences in addition to the glucose one. Additionally, the model accepts the daytime information corresponding to the observed input sequences. The model architecture $\mathcal{M}$ consists of the

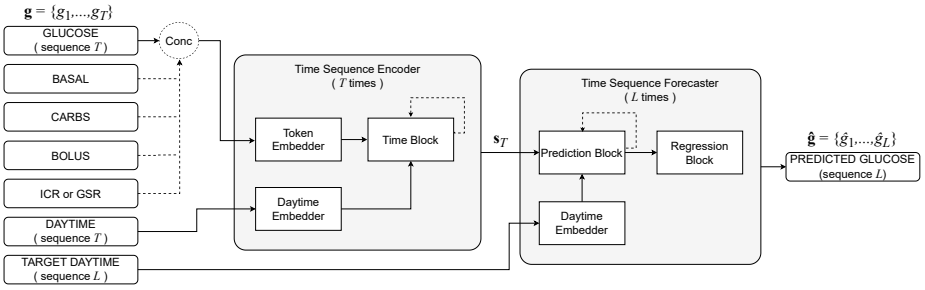

Fig. 1. The architecture of the proposed Sequential Transformer $\mathcal{M}$.

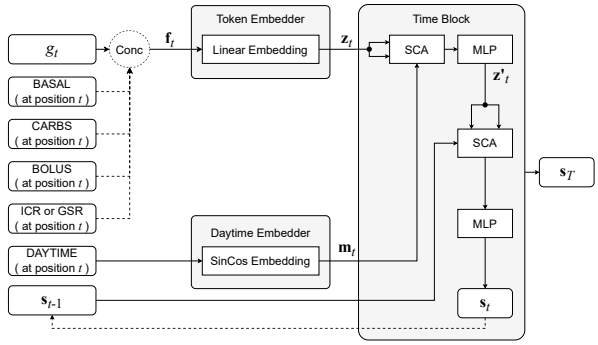

Fig. 2. Time Sequence Encoder $\mathcal{E}$.

concatenation of a Time Sequence Encoder $\mathcal{E}$ and a Time Sequence Forecaster $\mathcal{F}$. Each of these modules is based on the Multi-Layer Perceptron (MLP) operation and a slightly modified version of the Multi-head Cross-attention operation [34], which we refer to as Sigmoid Cross-attention (SCA). These are defined as follows:

a) Multi-Layer Perceptron: an MLP is defined as a sequence of two blocks, each composed of a linear projection, a GELU activation function [35], and a dropout layer. Specifically, given a vector $\mathbf{x}$, each block is defined as $\mathrm{dropout}(\mathrm{GELU}(\mathrm{Linear}(\mathbf{x})))$.

b) Sigmoid Cross-Attention: a single-head cross-attention operation is defined as: $\mathrm{c}_t^{att}(\mathbf{q}_t, \mathbf{k}_t, \mathbf{v}_t) = \sigma\left(\frac{\mathbf{q}_t \times \mathbf{k}_t^T}{\sqrt{d}}\right) \times \mathbf{v}_t$, where $\mathbf{q}_t$, $\mathbf{k}_t$ and $\mathbf{v}_t$ are the query, key and value at time t, respectively, d is a scale factor and σ is the sigmoid activation function. The query, key, and value share the same size N, which defines the embedding dimension of the entire model. Cross-attention computes attention between different sequences: queries are derived from one sequence, while keys and values come from another sequence. A Multi-head Cross-attention module, denoted as $\mathrm{c}_{t,D}^{att}$, consists of D heads that are processed in parallel and concatenated before being linearly projected through a linear layer.

C. Time Sequence Encoder

The model $\mathcal{E}$, shown in Figure 2, encodes the inputs sequentially using a Token Embedder, a DayTime Embedder, and a Time Block module. The inputs consist of the features of the observed sequence at time t defined as $\mathbf{f}_t$, the corresponding daytime at t, and a vector $\mathbf{s}_{t-1}$ of dimension N.

In the single-modality case, $\mathbf{f}_t$ corresponds to the glucose level $\mathbf{g}_t$; in the multi-modality case, instead, $\mathbf{g}_t$ at time t is concatenated with the other additional features. Instead, the vector $\mathbf{s}_{t-1}$ encodes the information and history of all observed sequences from time 1 to time $t-1$. Initially, $\mathbf{s}_0$ is randomly initialized and is the same for all sequences.

a) Token Embedder: it takes $\mathbf{f}_t$ as input and performs a Linear Embedding operation to (1) optionally fuse the multimodal input, and (2) project the input into an embedding of dimension N. The output is a vector denoted as $\mathbf{z}_t$.

b) DayTime Embedder: it uses the sine and cosine functions to project the daytime at time t into an embedding of the same size N, as in [31]. The resulting embedding is defined as $\mathbf{m}_t$.

c) Time Block module: it takes as inputs the embedding $\mathbf{z}_t$ representing the features at time t, the embedding $\mathbf{m}_t$ representing the daytime, and the encoding $\mathbf{s}_{t-1}$ representing the entire observed sequence up to $t-1$. In the first step, $\mathbf{z}_t$ and $\mathbf{m}_t$ are combined into a new vector $\mathbf{z'}_t$ using an MLP and a SCA mechanism. Specifically, $\mathbf{z'}_t = \mathrm{MLP}(\mathrm{c}_{t,D}^{att}(\mathbf{m}_t, \mathbf{z}_t, \mathbf{z}_t))$, where $\mathbf{m}_t$ serves as the query, and $\mathbf{z}_t$ acts as the key and value. This operation weights $\mathbf{z}_t$ based on the daytime. In the second step, $\mathbf{z'}_t$ is combined with the sequence encoding $\mathbf{s}_{t-1}$ observed till that moment, resulting in a new vector $\mathbf{s}_t$ that encapsulates the entire sequence information, including observations at time t. This step also employs an MLP and SCA mechanism; however, in this case, $\mathbf{s}_{t-1}$ is considered as the query, while $\mathbf{z'}_t$ as the key and the value, emphasizing the most recent element of the observed sequence. Formally, $\mathbf{s}_t = \mathrm{MLP}(\mathrm{c}_t^{att}(\mathbf{s}_{t-1}, \mathbf{z'}_t, \mathbf{z'}_t))$.

As previously mentioned, the entire Time Sequence Encoder is applied iteratively to each element of the observed sequence over t, resulting in T computations. After T iterations, the final output is a vector $\mathbf{s}_T$, which encodes the entire observed sequence input.

D. Time Sequence Forecaster

The Time Sequence Forecaster $\mathcal{F}$, illustrated in Figure 3, predicts future glucose levels using a DayTime Embedder, a Prediction Block, and a Regression Block module. Even in this case, the inputs are processed sequentially with operations repeated L times, where L is the number of glucose levels to predict. The forecaster $\mathcal{F}$ takes as input a vector $\mathbf{r}_i$, where i ranges from 0 to $L-1$, and the daytime corresponding to the next glucose level to be predicted, at time $i+1$. Initially, $\mathbf{r}_0$

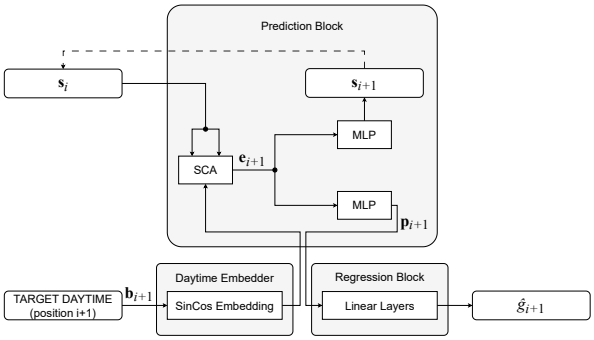

Fig. 3. Time Sequence Forecaster $\mathcal{F}$.

is set to $\mathbf{s}_T$, the output of the Time Sequence Encoder $\mathcal{E}$. For subsequent stages, when $i \geq 1$, $\mathbf{r}_i$ encodes the information of the sequence from time 1 to $T+i$, thus including all the observed sequence, from 1 to T, and the estimated sequence, up to time to $T+i$. The output of $\mathcal{F}$ is $\hat{g}_{i+1}$, the predicted glucose level at time $i+1_{th}$.

a) DayTime Embedder: it functions identically to the Day-Time Embedder described in III-C.0.b. However, instead of processing the daytime of the observed sequence, it handles the daytime corresponding to the element to predict. Specifically, it processes the daytime at time $i+1$ and outputs an embedding $\mathbf{b}_{i+1}$ of dimension N.

b) Prediction Block: it takes the embedding $\mathbf{b}_{i+1}$ of the daytime for the glucose level to be predicted at time $i+1$, and the embedding of the sequence $\mathbf{r}_i$ at time i. The output are two vectors of dimension N: the first, $\mathbf{r}_{i+1}$, represents the estimated sequence at time $i+1$, and the second, $\mathbf{p}_{i+1}$, represents the predicted glucose level at time $i+1$.

The process begins by combining $\mathbf{b}_{i+1}$ with $\mathbf{r}_i$ to obtain a vector $\mathbf{e}_{i+1}$. This combination is performed using a SCA mechanism. Formally, $\mathbf{e}_{i+1} = \mathrm{c}^{\mathrm{att}}_{\mathrm{t,D}}(\mathbf{b}_{i+1}, \mathbf{r}_i, \mathbf{r}_i)$, where $\mathbf{b}_{i+1}$ serves as the query, and $\mathbf{r}_i$ is considered as key and value. Through this operation, $\mathbf{e}_{i+1}$ represents the estimated information at time $i+1$, computed by weighting the prior information using the daytime of the element to predict. The next step involves extracting $\mathbf{r}_{i+1}$ and $\mathbf{p}_{i+1}$ from $\mathbf{e}_{i+1}$. This is achieved using two separate MLP s: $\mathbf{r}_{i+1} = \mathrm{MLP}(\mathbf{e}_{i+1})$ and $\mathbf{p}_{i+1} = \mathrm{MLP}(\mathbf{e}_{i+1})$.

c) Regression Block: it is the final module of the proposed model. It takes as input $\mathbf{p}_{i+1}$ and predicts the corresponding glucose level $\hat{g}_{i+1}$ at time $i+1$. The module consists of a sequence of linear layers.

As described earlier, the entire Time Sequence Forecaster operates sequentially, iterating over i, for a total of L times where L corresponds to the number of glucose levels to predict. At the end of the L iterations, the resulting output is the vector $\hat{\mathbf{g}} = \{\hat{g}_1, \cdots, \hat{g}_L\}$, where each element represents a predicted glucose level.

E. Loss function

The problem of BGC prediction is characterized by highly imbalanced datasets, particularly for normal, hypo-, and hyperglycemic events. Addressing this imbalance during training is crucial to ensure accurate predictions across all BGCs. Standard loss functions, such as Mean Squared Error (MSE), tend to be dominated by the majority class (i.e. normal glucose levels), which can lead to suboptimal performance in detecting rare but clinically critical events like hypoglycemia. To mitigate this, we use a *balanced MSE*, defined as:

$$\mathrm{BalancedMSE}(\mathbf{g}, \hat{\mathbf{g}}) = \sum_{i=1}^{L} w_i \cdot (g_i - \hat{g}_i)^2 \qquad (1)$$

where w_i represents the weight applied to the difference between the predicted value and the target value, and it is selected from the weights w_{event} based on the type of event associated with the target value $\hat{g}_i$.

The weights w_{event} are computed by first subtracting the frequency of each event type (hypoglycemic, normal, and hyperglycemic) in the training set from one. These values are then scaled by a $relevance_{event}$ parameter, according to the decreasing frequency of the events in the training set: hypoglycemic events, being the least frequent, are multiplied by 3, hyperglycemic events by 2, and normal events by 1: $w_{event} = relevance_{event} \cdot \left(1 - \frac{num\_events}{total\_events}\right)$.

The balanced MSE ensures that the model pays proportionally more attention to less frequent but critical events during training, improving its sensitivity to hypoglycemia, even if it slightly sacrifices precision in the more frequent normal class.

IV. Experiments

In this section, we present the experimental setup and metrics used to evaluate the proposed architecture.

A. Benchmark datasets

We conducted our experiments on two benchmark datasets: OhioT1DM [10] and DiaTrend [11]. The OhioT1DM dataset, widely used for BGC prediction, contains eight weeks of data from 12 individuals with T1D collected during the 2018 and 2020 challenges. It includes 5-minute CGM readings, insulin doses, blood glucose levels, carbohydrate intake, self-reported events (e.g., exercise, stress), and optional wearable fitness data (e.g., heart rate, activity levels). In contrast, the DiaTrend dataset, one of the largest open-source resources for diabetes research, features longitudinal data from 54 individuals with T1D. It comprises 27,561 days of CGM data at 5-minute intervals, 8,220 days of insulin pump data (including basal insulin for 17 subjects), bolus doses, carbohydrate intake, pump settings, and detailed demographic and clinical profiles. For our analysis, we focused on the 17 subjects with detailed basal insulin data.

In line with prior studies [28], [29], we employed a temporal-based data split, partitioning the dataset into 64% for training, 16% for validation, and 20% for testing. This approach trains the model on earlier data, validates it on data closer to the training period, and evaluates it on future data, offering a realistic assessment of predictive performance.

See Supplementary Materials[1] for more details on the pre-processing steps and the data split.

B. Baseline

To demonstrate the advantages of the Sequential Transformer, we compare its effectiveness against state-of-the-art methods, including a statistical baseline (ARIMA) [36] and a machine learning method (XGBoost) [37], as well as deep learning models. Among the latter, we consider transformer-based architectures such as Informer [26], TimesNet [27], Gluformer [24], and BG-BERT [28], together with a RNN baseline, Bi-GRU [29], with a particular focus on performance and model parameter count. Specifically, we benchmark it about two key settings: single-modal, using only the CGM feature, and multimodal, adopting the same features as BG-BERT and Bi-GRU, such as CGM, carbohydrate intake, bolus insulin dose, basal insulin rate, and Insulin-to-Carb Ratio (ICR) or Galvanic Skin Response (GSR), depending on the dataset being utilized (DiaTrend or OhioT1DM, respectively). Informer and TimesNet implementations are taken from `https://github.com/thuml/Time-Series-Library`.

To support our hypothesis, we also compared our method with an architecture based on the classical transformer encoder. The Standard Transformer architecture consists of a standard transformer encoder [31] followed by the same regression layers as our method (see Supplementary Materials[1] for details).

C. Experimental Setup

All experiments are carried out on the two datasets described in the previous section, with two configurations (different observation and prediction windows) for each dataset. Specifically, the first configuration features a 30-timestamp window (150 mins) divided into 24 timestamps (120 mins) for observed data and six timestamps (30 mins) for predictions. The second configuration uses a 60-timestamp window (300 mins) divided into 48 timestamps (240 mins) for observed data and 12 timestamps (60 mins) for predictions. These configurations are defined as OhioT1DM with PH = 30 mins, OhioT1DM with PH = 60 mins, DiaTrend with PH = 30 mins, and DiaTrend with PH = 60 mins. Since all configurations exhibit a significant imbalance between normal and adverse glucose levels, we applied the Synthetic Minority Over-sampling Technique (SMOTE) for data augmentation to the training set of all configurations [38]. This technique helps improve the representation of adverse glucose levels by generating synthetic samples based on k-nearest neighbors.

Besides the benchmark evaluation, we included an ablation study comparing the performance of the Sequential Transformer with different training settings. In detail, a balanced vs unbalanced training, with or without data augmentation, and with a single or multi-modality approach. The experiments are carried out in the same way as the benchmark evaluation. In addition, we carried out two further ablation studies: the first examined different feature combinations, while the second assessed the impact of including or excluding temporal information. These analyses were designed to isolate the contribution of each module in the proposed architecture and to investigate the role of specific input features, thereby providing more detailed evidence to support the validity of the findings. We also performed preliminary personalization experiments, where a model pre-trained on one dataset was directly evaluated on individual patients from a second dataset. Subsequently, patient-specific fine-tuning was conducted by retraining the model individually for each patient to assess the potential benefits of a personalized approach.

The training setup for each experiment, in both benchmark and ablation studies, has been characterized by 2000 epochs and early stopping with a patience of 150. We adopted the Adam optimizer with a learning rate of $1e^{-4}$ and a Reduce Learning Rate on the Plateau scheduler with 15 patience and 0.5 factor. The training strategies used for the state-of-the-art models, such as the choice of the loss function, are the same as proposed as default by the corresponding original works.

D. Evaluation Metrics

To assess the performance of the proposed model, we adopt three complementary types of evaluation.

1) Analytical Evaluation Metrics: The analytical assessment of each experiment—covering both benchmark and ablation studies—relies on four widely used metrics in the context of blood glucose prediction [28], [29]: Root Mean Square Error (RMSE), Sensitivity to Hyperglycemic Events (Hyper Sen), Sensitivity to Hypoglycemic Events (Hypo Sen), and Time Gain (TG).

The RMSE, defined in Equation 2, quantifies the average discrepancy between the predicted glucose levels $\hat{g}_i$ and the corresponding reference values g_i:

$$\mathrm{RMSE}(\mathbf{g}, \hat{\mathbf{g}}) = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(g_i - \hat{g}_i)^2} \tag{2}$$

The Hyper Sen and Hypo Sen metrics (Equations 3 and 4) evaluate the model's ability to correctly detect hyperglycemic and hypoglycemic events, respectively:

$$\text{Hyper Sen} = \frac{\text{True hyper events}}{\text{True hyper events} + \text{Missed hyper events}} \tag{3}$$

$$\text{Hypo Sen} = \frac{\text{True hypo events}}{\text{True hypo events} + \text{Missed hypo events}} \tag{4}$$

Here, *True Hyper Events* and *True Hypo Events* denote hyperglycemic and hypoglycemic episodes correctly identified by the model, whereas *Missed Hyper Events* and *Missed Hypo Events* refer to episodes that the model failed to recognize.

Finally, the TG metric, reported in Equation 5, measures the average anticipation time achieved by the model in detecting adverse events:

$$\mathrm{TG}(\mathbf{g}, \hat{\mathbf{g}}) = PH - \mathrm{delay}(\mathbf{g}, \hat{\mathbf{g}}) \tag{5}$$

[1]Supplementary materials here: `https://github.com/unimib-islab/Diabetes_Sequential_transformer/blob/23127d1e43e5c6bcec63d05a83ab800f5c31600d/Supplementary%20material%20-%20IEEE_JBHI___Sequential_Transformers_for_Glucose_Prediction.pdf`

TABLE I
OVERALL PERFORMANCE ON OHIOT1DM DATASET*

Model	PH = 30 mins				PH = 60 mins				Features °	Params	Ranking
	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)			
ARIMA	15.85	7.51	86.86	75.75	25.48	8.21	78.35	62.44	-	-	-
XGBoost	10.77	13.81	84.55	46.31	17.80	26.84	72.64	24.04	-	-	-
Informer	17.13	14.89	82.15	74.64	25.76	23.16	78.13	29.37	-	181K	5
TimesNet	13.63	14.71	90.22	**85.72**	23.48	23.40	80.21	67.27	-	18749K	9
Gluformer	17.27	**19.12**	74.33	0.33	27.14	29.46	43.07	0.48	-	11247K	8
Standard-T	15.37	14.90	69.11	8.96	24.43	30.58	50.09	0.89	-	107K	7
Sequential-T	14.96	17.56	**96.26**	75.82	26.42	33.11	**90.57**	68.95	-	123K	1
BG-BERT†	14.02	16.56	82.54	73.24	23.67	31.16	69.24	54.12	✓	2091K	4
Bi-GRU‡	**13.04**	17.57	84.13	80.03	**22.52**	32.99	66.48	61.03	✓	633K	3
Standard-T	17.17	15.57	82.98	5.21	24.74	29.81	63.49	0.03	✓	107K	6
Sequential-T	16.00	17.79	94.66	81.45	28.99	**33.59**	89.32	**69.96**	✓	123K	2

* In **bold** the best score and underlined the second best score (not including ARIMA and XGBoost); ° The Features column indicates if the model considers a multimodal approach using all the features available (✓), or uses only the glucose level (-); § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively; † Values reported form [28]; ‡ Values reported form [29].

where PH is the prediction horizon used to generate $\hat{\mathbf{g}}$, and $\text{delay}(\mathbf{g}, \hat{\mathbf{g}})$ represents the optimal time shift k minimizing the distance between the reference and predicted glucose trajectories:

$$\text{delay}(\mathbf{g}, \hat{\mathbf{g}}) = \underset{k}{\text{argmin}} \sum_{i=1}^{L} (g_i - \hat{g}_{i-k})^2 \tag{6}$$

In addition to these metrics, we also evaluate our sequential model (considering both the CGM-only and the multi-feature configurations) in terms of Specificity to Hyperglycemic Events (Hyper Spec), Specificity to Hypoglycemic Events (Hypo Spec) and False Alarm Rate (FAR), to assess their role in mitigating alarm fatigue.

To complement sensitivity, we report Hyper Spec and Hypo Spec (Equations 7 and 8), which correspond to the specificity (true negative rate) for hyperglycemic and hypoglycemic events, respectively:

$$\text{Hyper Spec} = \frac{\text{True non-hyper events}}{\text{True non-hyper events} + \text{False hyper events}} \tag{7}$$

$$\text{Hypo Spec} = \frac{\text{True non-hypo events}}{\text{True non-hypo events} + \text{False hypo events}} \tag{8}$$

The corresponding False Alarm Rates, which quantify the probability of incorrectly forecasting adverse events (lower is better), are defined as:

$$\text{Hyper FAR} = 1 - \text{Hyper Spec} \tag{9}$$

$$\text{Hypo FAR} = 1 - \text{Hypo Spec} \tag{10}$$

2) Clinical Evaluation Metrics: To assess the clinical reliability of BGC predictions, we employ Clarke's Error Grid Analysis (EGA) [39], a widely adopted tool in diabetes management for evaluating the risks associated with prediction errors. This method compares predicted and measured glucose values on a grid divided into five zones (A–E). Predictions in zones A and B are considered clinically safe, while zones C and D indicate decreasing safety, and zone E denotes the highest risk, where misclassification could lead to potentially fatal consequences.

3) Deployment Evaluation Metrics: Finally, we evaluate the feasibility of deploying the proposed models on edge devices by analyzing their computational requirements, memory footprint, and inference time.

V. RESULTS AND DISCUSSION

Table I and Table II respectively show the results achieved on the OhioT1DM and DiaTrend datasets, comparing the proposed method and training strategy with the state-of-the-art models. The reported experiments consider both the single and multimodal approaches. All architectures are evaluated on two dataset horizon configurations (30 and 60 minutes). Each table shows the difference in terms of evaluation metrics and parameters, indicating whether the experiment uses only CGM or also additional features. To facilitate the readability of the results, a Ranking column is also reported. The ranking is determined, considering the two PH configurations separately, by the average of the mean metrics' score, with the score of the parameters. Such scores are computed by normalizing the results of each metric column between 0 and 1. For metrics in which lower values correspond to better performance (e.g., RMSE), the complementary transformation $(1 - \text{normalized})$ was applied after normalization. The parameter score is computed in the same way on the Params column. The final rank is then obtained by averaging the scores across the 30- and 60-minute horizons, independently of the specific horizon considered. It is worth noting that ARIMA and XGBoost are not included in the ranking, as they represent statistical and machine learning models without trainable parameters, and a direct comparison with deep learning architectures would therefore be less meaningful.

a) OhioT1DM: as observable in Table I, the best rank is achieved by the proposed model with a single-modality approach, while the multimodal version achieves the third rank. This means that, on the OhioT1DM dataset, the Sequential Transformer proves to be optimal for real-life situations requiring edge computing, thanks to its extremely low dimensionality in terms of parameters and its performance considering only CGM. The second best performing model is Bi-GRU [29]. However, it is important to notice that even if the model is relatively small in terms of parameters and compared

TABLE II
OVERALL PERFORMANCE ON DIATREND DATASET*

Model	PH = 30 mins				PH = 60 mins				Features °	Params	Ranking
	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)			
ARIMA	17.60	7.48	83.72	57.70	27.74	8.25	73.51	38.61	-	-	-
XGBoost	11.77	14.14	81.60	15.84	19.30	27.54	67.85	8.67	-	-	-
Informer	16.57	13.56	79.83	50.76	25.85	23.41	70.14	14.78	-	181K	5
TimesNet	15.53	13.64	88.17	69.56	25.45	21.98	76.65	**46.62**	-	18749K	9
Gluformer	17.29	**17.74**	79.11	0.03	25.22	29.94	55.64	0.43	-	11247K	8
Standard-T	17.55	16.11	77.78	0.33	25.12	29.60	62.97	0.0	-	107K	6
Sequential-T	16.14	15.46	**95.62**	63.60	27.04	28.28	**88.60**	45.32	-	123K	1
BG-BERT †	14.85	16.47	81.34	62.27	24.95	31.45	64.53	40.10	✓	2091K	4
Bi-GRU ‡	**14.64**	16.66	81.31	63.82	**24.57**	**32.47**	65.72	31.64	✓	633K	3
Standard-T	18.00	16.18	78.22	0.25	37.32	28.93	37.40	0.00	✓	107K	7
Sequential-T	16.04	15.93	94.60	**77.85**	26.40	28.53	86.57	33.79	✓	123K	2

* In **bold** the best score and underlined the second best score (not including ARIMA and XGBoost); ° The Features column indicates if the model considers a multimodal approach using all the features available (✓), or uses only the glucose level (-); § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively; † Values reported form [28]; ‡ Values reported form [29].

TABLE III
SEQUENTIAL TRANSFORMER PERFORMANCE ON OHIOT1DM AND DIATREND DATASETS IN DIFFERENT TRAINING SETTINGS*

Dataset	Bal +	Aug x	Feat °	PH = 30 mins				PH = 60 mins				Ranking
				RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	
OhioT1DM	-	-	-	13.56	15.75	77.54	20.14	22.65	29.33	52.73	8.93	8
	-	-	✓	14.20	16.05	80.83	28.17	22.88	28.56	69.34	13.48	7
	-	✓	-	**13.06**	16.00	85.72	29.54	**21.67**	30.55	64.94	12.67	5
	-	✓	✓	13.83	16.32	82.06	38.11	23.16	29.38	71.70	11.21	6
	✓	-	-	13.93	15.65	95.42	68.07	24.47	30.63	87.79	65.88	3
	✓	-	✓	14.29	16.08	92.57	74.37	26.18	31.56	89.71	59.40	4
	✓	✓	-	14.96	17.56	**96.26**	75.82	26.42	33.11	**90.57**	68.95	1
	✓	✓	✓	16.00	**17.79**	94.66	**81.45**	28.99	**33.59**	89.32	**69.96**	2
DiaTrend	-	-	-	15.13	15.04	75.84	1.13	23.90	29.90	59.88	5.25	8
	-	-	✓	**14.75**	14.75	81.95	28.13	23.48	**30.37**	66.01	7.59	5
	-	✓	-	14.95	15.57	80.29	4.88	23.95	28.62	59.98	2.71	7
	-	✓	✓	14.81	14.90	83.88	22.27	**23.43**	29.63	67.11	8.36	6
	✓	-	-	15.54	14.88	92.55	61.00	26.41	26.61	88.31	36.48	3
	✓	-	✓	15.61	14.37	93.00	57.81	25.98	28.59	86.37	29.83	4
	✓	✓	-	16.14	15.46	**95.62**	**63.60**	27.04	28.28	**88.60**	**45.32**	1
	✓	✓	✓	16.04	**15.93**	94.60	77.85	26.40	28.53	86.57	33.79	2

* In **bold** the best score and underlined the second best score; + The Balance column indicates if the loss used was balanced (✓) or unbalanced (-); x The Data aug column indicates if the data augmentation has been applied (✓) or not (-); ° The Features column indicates if the model considers a multimodal approach using all the features available (✓), or uses only the glucose level (-); § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively.

with other architectures, its dimension can still be a problem on wearable devices. Another fundamental analysis regards the identification of future adverse events. The proposed strategy with the Sequential Transformer can always achieve the best or among the best results in detecting hyper and hypo events. On this particular issue, it is possible to observe that the multimodal Sequential Transformer outperforms the Bi-GRU, thus being more suitable in a real-life scenario.

b) DiaTrend: Table II shows a trend similar to the OhioT1DM dataset. However, in this case, the Sequential Transformer takes both the best and second-best positions on the ranking. The single-modality version achieves the best ranking, with the multimodal version reaching the second place. As before, the Sequential Transformer in every version achieves the best or among the best performance in detecting adverse events. The only exception is TimesNet [27], which achieves the second-best performance in Hypo Sen. Nonetheless, its number of parameters makes it unsuitable for real-life scenarios where edge computing results in a better solution.

A. Ablation studies

In this section, we evaluate different configurations of the proposed method through a set of ablation studies.

1) Training strategies: We investigate the effectiveness of using balance and unbalanced loss for training, the advantages of data augmentation, and the exploitation of multi- and single-modality approaches. As for the afore-described experiments, a Ranking is defined to better identify which method performs best. This ranking only considers the evaluation metrics since all tested configurations use the same architecture.

The first rows of Table III report the scores obtained on the OhioT1DM dataset with different settings. The best-ranked method is represented by our Sequential Transformer using the proposed balancing strategy, implementing data augmentation, and, as determined before, using only CGM. The second-best is the multimodal version, with the same setting in terms of loss and data augmentation. It is interesting to notice that, based on the ranking, the balanced versions represent the first four best choices, demonstrating the effectiveness of the defined training strategy. Similar conclusions can be drawn for the DiaTrend dataset. The first rank is still taken by the single-modality setting with a balanced strategy and data augmentation. The second is the same setting but with a multimodal approach. Even in this case, it is possible to notice that the four best-ranked settings are all based on the balance training strategy, remarking once again on the importance of

such implementation.

In summary, this study demonstrates that the proposed method coupled with the proposed training strategy, is optimal in edge computing systems, thanks to its dimension and the single-modality performance, which remove the problem of communication between multiple and heterogeneous devices [40]. However, the results also indicate that the implementation of the multimodal approach can be refined, leaving room for potential future enhancements.

2) Feature combinations: We evaluated different feature combinations to analyze the contribution of each input signal. Table IV reports the results obtained on the OhioT1DM and DiaTrend datasets when combining CGM (i.e. glucose level) with additional features such as basal and bolus insulin, carbohydrate intake, and other subject-specific information (e.g., GSR or ICR). The results show that CGM alone provides highly competitive performance, often achieving the lowest RMSE and superior sensitivity values across both prediction horizons. For instance, on the OhioT1DM dataset, CGM alone yields the best RMSE at both 30 and 60 minutes, while also achieving the highest hypoglycemia sensitivity. Similarly, on the DiaTrend dataset, CGM alone again leads in terms of RMSE and hypoglycemia sensitivity, confirming its effectiveness as a standalone input modality. Nevertheless, some multimodal configurations bring improvements on specific metrics. For example, on OhioT1DM, combining CGM with carbohydrate intake improves hyperglycemia sensitivity at PH = 60, while on DiaTrend, CGM plus bolus insulin dose achieves the best hyperglycemia sensitivity at the same horizon. These findings indicate that, while CGM is the most informative feature for glucose prediction, complementary signals can provide marginal benefits for selected metrics. Overall, this ablation study highlights that CGM is sufficient to ensure strong performance in most cases, but that carefully selected multimodal settings may further enhance specific aspects of prediction accuracy.

3) Temporal information: The third ablation study focused on assessing the impact of temporal information. To better understand the contribution of each module in the proposed architecture, we selectively disabled the module associated with time (the daytime embedder), in order to evaluate whether the remaining input features could effectively operate on their own. In particular, for this study, only the CGM has been considered as input, while a series of zeros has substituted the time information. Table V reports the results on both OhioT1DM and DiaTrend datasets. The findings show that the inclusion of temporal information consistently enhances the model's ability to detect glucose excursions. In OhioT1DM, the daytime embedder improves both hyperglycemia and hypoglycemia sensitivity, particularly at shorter horizons. Similarly, in DiaTrend, temporal information increases sensitivity — especially for hypoglycemia detection — even if this comes at the cost of a slightly higher RMSE. Overall, these results highlight that temporal context is an essential complementary signal to raw physiological features. While the features alone can achieve competitive performance, the inclusion of temporal information enriches the representation and enables more reliable glucose prediction across different scenarios.

B. Clarke Error Grid Analysis

The Clarke Error Grid is a valuable instrument for assessing the clinical acceptability of BGC predictions at 30- and 60-minute horizons. In this analysis, we report the percentage of samples in each grid zone (see Supplementary Materials[1] for the plots visualization and more details).

For OhioT1DM, the proposed model achieves the following results: 94.15%, 5.67%, 0.01%, 0.17%, and 0.00% in zones A, B, C, D, and E, respectively, for the 30-minute PH, and 82.54%, 16.70%, 0.29%, 0.46%, and 0.02% for the 60-minute PH. On the DiaTrend dataset, the model performs similarly with 93.13%, 6.64%, 0.03%, 0.21%, and 0.00% in zones A–E for the 30-minute PH, and 82.42%, 16.60%, 0.23%, 0.72%, and 0.03% for the 60-minute PH. These results highlight the model's strong performance in predicting BGC at shorter time horizons (30 mins), with only minor deviations, as indicated by the small proportion of samples in zone B. Though there are a few samples in zones D and E, indicating larger errors, these are relatively rare. For longer horizons (60 mins), accuracy decreases as expected, but a significant proportion remains in zone A, showing the model's reliability despite more clinically less accurate predictions. This reflects the usual trade-off between accuracy and time horizon in time-series modeling.

C. Managing Alarm Burden

We further evaluate our sequential model (in both the CGM-only and multi-feature configurations) in terms of Specificity and False Alarm Rate, with a focus on their relevance to alarm fatigue.

The detailed results, reported in Table I of the Supplementary Materials [1], include Time Gain, Hypo/Hyper Sensitivity, Hypo/Hyper Specificity, and the corresponding False Alarm Rates across all training configurations. Overall, the model excels at forecasting hyperglycemic events and maintains good performance on hypoglycemic sensitivity, except OhioT1DM at PH = 60, which yields lower scores. Importantly, in all settings the model reaches specificity values above 95% with False Alarm Rates lower than 5%, with the only exception of OhioT1DM at PH = 60, where specificity slightly falls below the threshold (93.87%, corresponding to a 6.13% FAR).

These findings reinforce the robustness of the proposed method and underline its clinical relevance. In daily life, individuals with diabetes must promptly react to predicted adverse events to avoid potentially life-threatening conditions. However, if a forecasting system generates too many false alarms, users may gradually lose trust in its reliability, ignore alerts, or even disable them altogether, thereby increasing the risk of severe consequences [41].

D. Patient-specific model: a preliminary experiment

All experiments presented in the main manuscript adopt a generalized approach, in which patients are treated as a single pool of time series without considering subject identity. This design ensures that the evaluation reflects a purely subject-independent setting, which represents the primary focus of this work. However, personalization has been widely discussed in the literature as a relevant aspect of glucose prediction,

TABLE IV
SEQUENTIAL TRANSFORMER PERFORMANCE ON OHIOT1DM AND DIATREND DATASETS WITH DIFFERENT FEATURE COUPLES*

Dataset	Features	PH = 30 mins				PH = 60 mins			
		RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)
OhioT1DM	CGM	**14.96**	17.56	**96.26**	**75.82**	**26.42**	33.11	**90.57**	**68.95**
	CGM + Basal Insulin Rate	16.36	**19.05**	93.75	40.55	27.08	35.02	84.71	54.64
	CGM + Bolus Insulin Dose	15.87	18.76	92.88	53.76	27.72	35.02	84.47	40.27
	CGM + Carbohydrate Intake	16.44	18.36	91.54	46.91	28.42	**36.99**	85.07	31.69
	CGM + GSR	16.03	18.57	92.25	48.36	28.66	36.82	83.66	39.70
DiaTrend	CGM	**16.14**	15.46	95.62	**63.60**	**27.04**	28.28	88.60	**45.32**
	CGM + Basal Insulin Rate	18.58	**16.71**	**96.87**	9.35	34.70	37.89	94.39	6.81
	CGM + Bolus Insulin Dose	18.02	16.23	96.27	15.31	35.43	**39.99**	94.48	7.91
	CGM + Carbohydrate Intake	18.28	16.49	96.40	17,74	35.22	39.06	**94.89**	9.43
	CGM + ICR	17.44	15.61	96.62	27.03	37.39	37.04	94.30	7.75

* In **bold** the best score; § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively.

TABLE V
SEQUENTIAL TRANSFORMER PERFORMANCE ON OHIOT1DM AND DIATREND DATASETS CONSIDERING ONLY CGM, WITH AND WITHOUT THE TIME FEATURE*

Dataset	Time+	PH = 30 mins				PH = 60 mins			
		RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)
OhioT1DM	✓	14.96	**17.56**	**96.26**	**75.82**	26.42	**33.11**	**90.57**	**68.95**
	-	**14.23**	16.00	94.42	63.16	**24.96**	31.30	90.31	60.49
DiaTrend	✓	16.14	**15.46**	**95.62**	**63.60**	27.04	**28.28**	**88.60**	**45.32**
	-	**15.96**	15.30	94.01	60.23	**26.39**	26.83	86.48	26.96

* In **bold** the best score; + The Time column indicates if the daytime embedder was enabled (✓) or disabled (-); § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively.

with reported benefits strongly depending on both dataset characteristics and model architecture. To complement our generalized analysis and provide a broader perspective, we conducted preliminary investigations on patient-specific modeling, always considering only CGM as input feature. In particular, we explored two scenarios: (i) directly evaluating a model pre-trained on one dataset on individual patients from a second dataset, and (ii) applying patient-specific fine-tuning by retraining the model for each subject individually.

The Tables VI and VII show the results achieved by pretraining the model with DiaTrend dataset, and testing and fine-tuning it with OhioT1DM, respectively, with 30- and 60-minute horizons. Each row represents a different patient and the corresponding evaluation metrics, while the last row reports the average performance on all patients. The fine-tuning achieve comparable results when the 30-minute horizon is considered, increasing in particular the performance on TG. On the other hand, when 60-minute horizon is considered, all the metrics, apart from RMSE, achieve improvements.

Conversely, Tables VIII and IX show the opposite case, with pretraining on OhioT1DM and fine-tuning on DiaTrend patients. At 30 minutes, performance generally worsens after fine-tuning, while at 60 minutes results remain comparable for RMSE, TG, and hyperglycemia sensitivity. However, patient-specific fine-tuning fails to identify hypoglycemia events, as demonstrated by the drop in Hypo Sen.

Overall, these preliminary findings suggest that patient-specific adaptation can be beneficial in certain conditions, particularly at longer horizons, but its effectiveness is highly dataset-dependent and requires careful design to avoid degrading performance on critical events such as hypoglycemia.

TABLE VI
PERSONALIZATION PERFORMANCE ON OHIOT1DM (PH = 30 MINS) USING A MODEL PRE-TRAINED ON DIATREND (PH = 30 MINS) WITH ONLY CGM*

Patient	Pre-trained on DiaTrend				Fine-tuning			
	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)
540	15.23	15.72	**96.16**	79.71	**14.82**	**17.47**	95.77	**82.63**
544	**12.60**	15.99	**97.78**	**74.44**	12.76	**18.63**	96.54	64.22
552	11.91	15.22	96.49	**74.48**	**11.86**	**15.85**	**98.11**	69.72
559	**13.81**	16.07	95.24	**92.54**	14.31	**18.15**	**95.42**	90.70
563	**11.97**	15.33	70.83	33.33	12.75	**17.41**	**78.40**	**66.66**
567	15.15	15.49	91.93	**87.72**	**14.49**	**19.92**	**93.85**	71.57
570	**12.43**	16.30	**97.55**	**68.52**	12.74	**17.28**	97.02	63.64
575	**11.81**	17.09	**96.86**	**89.17**	11.82	**17.86**	94.70	88.22
584	**16.47**	15.21	**94.11**	29.17	16.97	**16.91**	94.09	**30.56**
588	**12.77**	16.20	**95.89**	**66.66**	13.68	**18.27**	92.09	50.00
591	**17.87**	15.80	**90.83**	**70.42**	18.19	**17.10**	88.71	68.53
596	**13.74**	15.33	**95.12**	**68.60**	14.40	**17.50**	92.99	65.21
Average	**13.81**	15.81	**93.23**	**69.56**	14.07	**17.70**	93.14	67.64

* At each line in **bold** the best score; § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively.

E. Deployment on wearable devices

Our method, used in its best configuration (only CGM), requires 0.49 MB of flash memory to store the model's weights. The RAM usage to temporarily store input data, activations, and output predictions is 320 KB in the 30-minute PH and nearly doubles, reaching 642 KB, in the 60-minute PH. These requirements can be easily satisfied by microcontroller units (MCUs) for the 30-minute PH, making our method suitable for edge computing deployment. More critical remains the 60-minute horizon for which few MCUs can provide enough RAM. In this case, the RAM usage can be lowered thanks to the recurrent structure of the model that allows the processing of the input time-series samples independently. See Supplemental Materials[1] for further discussion on this topic.

TABLE VII
PERSONALIZATION PERFORMANCE ON OHIOT1DM (PH = 60 MINS) USING A MODEL PRE-TRAINED ON DIATREND (PH = 60 MINS) WITH ONLY CGM*

Patient	Pre-trained on DiaTrend				Fine-tuning			
	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)
540	28.44	**29.97**	89.07	47.79	**22.92**	28.75	**91.19**	**61.06**
544	**22.76**	27.46	90.32	46.15	25.28	**33.23**	**91.98**	**56.92**
552	21.75	28.62	88.77	40.12	**21.25**	**31.13**	**92.70**	**51.71**
559	**25.00**	29.28	90.17	69.23	27.50	**32.28**	**91.35**	**82.63**
563	**21.77**	28.30	49.65	5.56	24.3	**33.19**	**61.73**	**41.67**
567	**26.99**	28.52	64.63	57.45	27.23	**28.89**	**72.00**	**69.13**
570	**20.06**	30.85	**95.54**	29.63	22.65	**33.50**	92.18	**30.30**
575	**21.04**	29.54	**82.85**	63.36	23.62	**33.70**	82.76	**83.62**
584	**25.86**	29.58	86.10	19.44	27.10	**30.39**	**86.43**	**26.39**
588	**21.77**	28.61	**87.79**	25.00	25.44	**31.82**	82.37	**31.25**
591	**27.80**	28.09	**76.39**	40.27	29.30	**32.86**	70.83	**68.60**
596	**22.92**	**28.32**	85.77	39.58	25.36	27.36	**87.43**	**52.88**
Average	**23.85**	28.93	82.25	40.30	25.16	**31.43**	**83.58**	**54.68**

* At each line in **bold** the best score; § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively.

TABLE VIII
PERSONALIZATION PERFORMANCE ON DIATREND (PH = 30 MINS) USING A MODEL PRE-TRAINED ON OHIOT1DM (PH = 30 MINS) WITH ONLY CGM*

Patient	Pre-trained on OhioT1DM				Fine-tuning			
	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)
29	15.25	**16.78**	**95.75**	**74.54**	**14.50**	14.81	**95.75**	39.99
30	16.26	**16.00**	87.04	**69.99**	**15.25**	14.32	**87.35**	35.59
31	15.68	**16.98**	95.95	**45.29**	**14.80**	15.73	**96.13**	15.94
36	20.67	**16.03**	95.56	**35.00**	**19.36**	14.68	**96.41**	13.33
37	19.87	**16.64**	89.53	**56.67**	**19.22**	15.11	**90.12**	43.33
38	19.02	**15.55**	**90.60**	**54.86**	**17.65**	14.24	90.25	17.01
39	15.05	**16.08**	**95.40**	**64.48**	**14.31**	14.30	95.17	35.79
42	15.74	**15.92**	96.56	**71.65**	**14.66**	14.44	**97.07**	37.55
45	21.62	**15.55**	84.11	**51.52**	**20.49**	14.26	**85.94**	21.21
46	15.86	**18.30**	96.25	**60.00**	**14.81**	15.12	**97.82**	20.00
47	17.16	**15.76**	83.33	**29.17**	**15.63**	14.68	**87.50**	12.50
49	10.04	**17.28**	-	**52.22**	**9.42**	12.29	-	23.33
50	20.94	**17.93**	95.98	-	**20.22**	16.56	**96.33**	-
51	19.61	**15.32**	96.64	-	**18.91**	13.96	**97.23**	-
52	13.47	**18.78**	98.94	-	**11.94**	17.31	**99.15**	-
53	21.08	**16.54**	**95.12**	**59.42**	**20.46**	14.62	94.29	24.64
54	15.74	**17.80**	**66.67**	**69.23**	**13.94**	15.44	55.56	8.55
Average	17.24	**16.66**	**91.46**	**56.72**	**16.21**	14.82	91.38	24.91

* At each line in **bold** the best score; § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively; - indicates where the sensitivity could not be computed due to a lack of hyper- or hypo-glycemic events.

TABLE IX
PERSONALIZATION PERFORMANCE ON DIATREND (PH = 60 MINS) USING A MODEL PRE-TRAINED ON OHIOT1DM (PH = 60 MINS) WITH ONLY CGM*

Patient	Pre-trained on OhioT1DM				Fine-tuning			
	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)	RMSE (mg/dL)	TG (mins)	Hyper Sen § (%)	Hypo Sen § (%)
29	**28.66**	31.14	**88.31**	**58.40**	29.31	**32.45**	88.00	1.11
30	**27.11**	30.32	**81.25**	**59.51**	29.00	**31.56**	75.62	0.93
31	28.20	29.46	90.62	**27.85**	**27.08**	**30.90**	**91.98**	0.00
36	35.92	28.27	91.23	**21.30**	**31.95**	**30.02**	**92.16**	0.00
37	33.26	**32.53**	78.46	**26.67**	**32.75**	30.15	**78.58**	0.00
38	**31.48**	30.35	**83.89**	**38.53**	31.78	**32.56**	78.43	0.38
39	**27.23**	**32.03**	**91.58**	**50.62**	28.85	30.86	82.97	0.16
42	28.13	31.19	89.12	**58.78**	**27.59**	**32.43**	**92.11**	1.90
45	35.33	28.31	**78.67**	**33.54**	**32.87**	**30.53**	69.77	0.00
46	27.79	**32.58**	92.32	**56.67**	**27.48**	26.11	**95.57**	0.00
47	29.47	32.03	56.25	**16.67**	**29.02**	**34.97**	**66.67**	0.00
49	**19.73**	30.06	-	**32.18**	41.67	**49.29**	-	0.00
50	32.23	**29.99**	92.91	-	**30.98**	25.77	**93.85**	-
51	**31.93**	**31.68**	92.81	-	32.59	29.96	**94.08**	-
52	24.94	**37.97**	96.35	-	**21.82**	37.88	**98.18**	-
53	**39.03**	**30.43**	**84.39**	**39.86**	39.42	30.06	84.30	2.36
54	28.18	32.57	36.11	**54.70**	**26.32**	**36.43**	**72.22**	0.00
Average	**29.92**	31.23	82.77	**41.09**	30.62	**32.47**	**84.66**	0.49

* At each line in **bold** the best score; § Hyper Sen and Hypo Sen indicate the detection sensitivity of hyperglycemia and hypoglycemia respectively; - indicates where the sensitivity could not be computed due to a lack of hyper- or hypo-glycemic events.

VI. CONCLUSION

In this paper, we defined a novel method for BGC prediction in T1D patients, aiming, at the same time, to design an architecture that is both accurate and deployable on low-resource edge computing systems. This task poses significant challenges, requiring lightweight and efficient models capable of operating within the constraints of edge devices. Furthermore, the nature of T1D presents additional challenges, including the inherent imbalance between normal, hypo-, and hyperglycemic events. The lower occurrence rates of these adverse events exacerbate the difficulty of building and training robust deep-learning models. To address these challenges, we proposed the Sequential Transformer, a novel architecture that integrates the strengths of Transformers and RNNs. This design is coupled with a training strategy based on balanced MSE, specifically crafted to address the inherent imbalance in diabetes-related events. Our approach delivers both high performance and a limited dimension model size, meeting the requirements of resource-constrained environments. The ablation studies further validated the effectiveness of the balanced MSE, consistently outperforming conventional strategies. They also highlighted the competitiveness of CGM as a standalone feature, the benefits of temporal information, and the potential—but still limited—contribution of additional modalities. Moreover, preliminary investigations on patient-specific fine-tuning showed selective improvements, particularly at longer horizons, suggesting personalization as a promising yet non-trivial direction. Overall, the proposed model proved effective in both single- and multimodality settings, underscoring its adaptability and practical suitability for edge deployment in real-world applications.

Future studies will refine multimodal strategies and better leverage patient-specific features to improve event prediction, continuing the preliminary investigations conducted in this work. Personalization will be central, along with exploring deployment on edge devices for real-time applications. Finally, since model transparency is essential to provide actionable insights to clinicians, we will investigate dedicated explainability strategies tailored to short CGM sequences.

ACKNOWLEDGMENT

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM) [2]. This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

REFERENCES

[1] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *The lancet*, vol. 383, no. 9911, pp. 69–82, 2014.

[2] P. E. Cryer, S. N. Davis, and H. Shamoon, "Hypoglycemia in diabetes," *Diabetes care*, vol. 26, no. 6, pp. 1902–1912, 2003.

[2] https://fondazioneanthem.it/

[3] M. Kotagal, R. G. Symons, I. B. Hirsch, G. E. Umpierrez, E. P. Dellinger, E. T. Farrokhi, D. R. Flum *et al.*, "Perioperative hyperglycemia and risk of adverse events among patients with and without diabetes," *Annals of surgery*, vol. 261, no. 1, pp. 97–103, 2015.

[4] D. Rodbard, "Continuous glucose monitoring: a review of successes, challenges, and opportunities," *Diabetes technology & therapeutics*, vol. 18, no. S2, pp. S2–3, 2016.

[5] K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, ser. CEUR Workshop Proceedings, vol. 2675. CEUR-WS.org, 2020. [Online]. Available: https://ceur-ws.org/Vol-2675

[6] T. Zhu, L. Kuang, C. Piao, J. Zeng, K. Li, and P. Georgiou, "Population-specific glucose prediction in diabetes care with transformer-based deep learning on the edge," *IEEE Transactions on Biomedical Circuits and Systems*, 2024.

[7] A. R. Nasser, A. M. Hasan, A. J. Humaidi, A. Alkhayyat, L. Alzubaidi, M. A. Fadhel, J. Santamaría, and Y. Duan, "Iot and cloud computing in health-care: A new wearable device and cloud-based deep learning algorithm for monitoring of diabetes," *Electronics*, vol. 10, no. 21, p. 2719, 2021.

[8] G. M. Bhat and N. G. Bhat, "A novel iot based framework for blood glucose examination," in *2017 international conference on electrical, electronics, communication, computer, and optimization techniques (ICEECCOT)*. IEEE, 2017, pp. 205–207.

[9] S. Aminizadeh, A. Heidari, S. Toumaj, M. Darbandi, N. J. Navimipour, M. Rezaei, S. Talebi, P. Azad, and M. Unal, "The applications of machine learning techniques in medical data processing based on distributed computing and the internet of things," *Computer methods and programs in biomedicine*, p. 107745, 2023.

[10] B. R. Marling C., "The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020," *CEUR workshop proceedings*, pp. 71–74, 2020.

[11] T. Prioleau, A. Bartolome, R. Comi, and C. Stanger, "Diatrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions," *Scientific Data*, vol. 10, no. 1, p. 556, 2023.

[12] R. H. Shumway, D. S. Stoffer, R. H. Shumway, and D. S. Stoffer, "Arima models," *Time series analysis and its applications: with R examples*, pp. 75–163, 2017.

[13] R. McShinsky and B. Marshall, "Comparison of forecasting algorithms for type 1 diabetic glucose prediction on 30 and 60-minute prediction horizons." in *KDH@ ECAI*, 2020, pp. 12–18.

[14] A. Bhimireddy, P. Sinha, B. Oluwalade, J. W. Gichoya, and S. Purkayastha, "Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks," in *KDH@ECAI*. CEUR Workshop Proceedings, 2020.

[15] J. Xie and Q. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for type i diabetes in comparison with classical time-series models," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3101–3124, 2020.

[16] Y. Xing, H. Ye, X. Zhang, W. Cao, S. Zheng, J. Bian, and Y. Guo, "A continuous glucose monitoring measurements forecasting approach via sporadic blood glucose monitoring," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 860–863.

[17] J. Chen, K. Li, P. Herrero, T. Zhu, and P. Georgiou, "Dilated recurrent neural network for short-time prediction of glucose concentration." in *KDH@ IJCAI*, 2018, pp. 69–73.

[18] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, "Blood glucose prediction with variance estimation using recurrent neural networks," *Journal of Healthcare Informatics Research*, vol. 4, pp. 1–18, 2020.

[19] K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 603–613, 2019.

[20] J. Martinsson, A. Schliep, B. Eliasson, C. Meijner, S. Persson, and O. Mogren, "Automatic blood glucose prediction with confidence using recurrent neural networks," in *3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ IJCAI-ECAI 2018, 13 July 2018*, 2018, pp. 64–68.

[21] Y. Xue, S. Guan, and W. Jia, "Bgformer: An improved informer model to enhance blood glucose prediction," *Journal of Biomedical Informatics*, vol. 157, p. 104715, 2024.

[22] Q. Bian, A. As' arry, X. Cong, K. A. b. M. Rezali, and R. M. K. b. Raja Ahmad, "A hybrid transformer-lstm model apply to glucose prediction," *PLoS One*, vol. 19, no. 9, p. e0310084, 2024.

[23] T. Zhu, T. Chen, L. Kuangt, J. Zeng, K. Li, and P. Georgiou, "Edge-based temporal fusion transformer for multi-horizon blood glucose prediction," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2023, pp. 1–5.

[24] R. Sergazinov, M. Armandpour, and I. Gaynanova, "Gluformer: Transformer-based personalized glucose forecasting with uncertainty quantification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] E. Acuna and R. Aparicio, "Predicting the blood glucose level using transformers," in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2023, pp. 1392–1399.

[26] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[27] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[28] X. Zheng, S. Ji, and C. Wu, "Predicting adverse events for patients with type-1 diabetes via self-supervised learning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1526–1530.

[29] G. Rigamonti, M. P. Barbato, D. Marelli, and P. Napoletano, "Improving detection of type-1 diabetes adverse events using gru networks," in *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*. IEEE, 2024, pp. 79–84.

[30] M. M. H. Shuvo and S. K. Islam, "Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1612–1623, 2023.

[31] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[32] S. Grossberg, "Recurrent neural networks," *Scholarpedia*, vol. 8, no. 2, p. 1888, 2013.

[33] B. Kovatchev and C. Cobelli, "Glucose variability: timing, risk analysis, and relationship to hypoglycemia in diabetes," *Diabetes Care*, vol. 39, no. 4, pp. 502–510, 2016.

[34] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

[35] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[36] H. Hameed and S. Kleinberg, "Comparing machine learning techniques for blood glucose forecasting using free-living and patient generated data," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 871–894.

[37] C.-H. Lin and C.-L. Liu, "Prediction of blood glucose concentration based on optiscanner and xgboost in icu," *IEEE Access*, vol. 11, pp. 116 524–116 533, 2023.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[39] W. Clarke, C. DC, L. Gonder-Frederick, W. Carter, and S. Pohl, "Evaluating clinical accuracy of systems for self-monitoring of blood glucose," *Diabetes care*, vol. 10, pp. 622–8, 09 1987.

[40] M. Xu, W. C. Ng, W. Y. B. Lim, J. Kang, Z. Xiong, D. Niyato, Q. Yang, X. Shen, and C. Miao, "A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 656–700, 2022.

[41] J. P. Shivers, L. Mackowiak, H. Anhalt, and H. Zisser, ""turn it off!": diabetes device alarm fatigue considerations for the present and the future," *Journal of diabetes science and technology*, vol. 7, no. 3, pp. 789–794, 2013.