ADVANCED INTELLIGENT SYSTEMS
Open Access
www.advintellsyst.com

Improving Long-Term Glucose Prediction Accuracy with Uncertainty-Estimated ProbSparse-Transformer

Wei Huang, Ni Fan, Weiping Wang, Jinqiang Wang, Xiaojuan Qi, and Shiming Zhang**

Accurate prediction of blood glucose (BG) with precise data recorded by continuous glucose monitoring (CGM) is essential to improve the safety of closed-loop insulin delivery systems for diabetic patients. However, predicting BG trends under long-term prediction horizons is challenging due to the dynamic complexity of glucose changes. In this work, a ProbSparse-Transformer model, which alleviates the long-term error spreading effect seen in traditional autoregressive models, is developed. This model incorporates a trustworthy uncertainty-estimation approach to reduce output variance, further improving predictive accuracy. Additionally, an open-source benchmark is established using four public datasets and five evaluation metrics to comprehensively assess model performance. This model shows significant improvements in both short-term (15–30 min) and long-term (45–60 min) BG predictions. In the 60 min task, it achieves root mean square error values of 10.86, 15.33, 20.46, and 13.74 mg dL^{-1} across four datasets, representing a 20%–39.4% improvement over previous methods. Finally, the model on edge devices is compressed and deployed, demonstrating its potential for practical application in real CGM systems.

1. Introduction

Diabetes mellitus, affecting over 500 million people globally, is characterized by high blood glucose (BG) levels stemming from inadequate insulin production (Type 1) or diminished insulin responsiveness (Type 2).[1] This metabolic disorder can lead to serious vascular and nonvascular complications, thereby increasing mortality rates.[2,3] Recent advancements in wearable technology, particularly continuous glucose monitorings (CGMs), have revolutionized diabetes management by enabling continuous tracking of BG levels.[2,4] However, while CGMs are adept at monitoring, they lack therapeutic functions, prompting the development of closed-loop systems that adjust insulin dosing based on glucose readings. The effectiveness of these systems heavily relies on the algorithms used to predict BG levels, which are still in early development stages.[5] Enhancing algorithm accuracy is crucial for improving the safety of closed-loop CGMs, yet the challenge lies in creating models that can effectively handle the uncertainties and long-term inaccuracies associated with BG predictions.[6–8]

Early BG prediction works primarily focused on the use of traditional machine learning algorithms, including models such as autoregressive (AR) moving average (MA),[9] support vector regression,[10] and random forest.[11] Compared with traditional mathematical modeling methods, machine learning can model nonlinear BG changes in a data-driven learning manner. In recent years, deep-learning technologies, driven by big data and backpropagation computations,[12] have been extensively utilized in BG prediction. Deep learning is adept at capturing nonlinear functional relationships in complex scenarios and mass data. Deep neural networks (DNNs) based on architectures like convolution neural network (CNN),[13] recurrent neural networks (RNNs),[14] long short-term memory (LSTM),[15] and gated recurrent unit (GRU)[16] are widely used in glucose prediction. Transformer[17] model also gained attention in BG task, which demonstrates outstanding performance in diverse time series prediction contexts, distinguished for its advanced contextual learning capabilities via a multi-head attention mechanism. Meanwhile, recent works[18,19] reported that the Transformer model achieves leading performance in BG prediction over vanilla RNN families.

However, due to the nonlinear and complexity of glucose fluctuations, it is a significant challenge to accurately forecast BG trends under long-term prediction horizons (LTPHs). As shown in **Figure 1**, we find that LSTM and existing advanced Transformer models for BG prediction suffer significant

W. Huang, X. Qi, S. Zhang
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong 999077, Hong Kong
E-mail: xjqi@eee.hku.hk; szhang@eee.hku.hk

N. Fan, W. Wang
Department of Pharmacology
The University of Hong Kong
21 Sassoon Road, Pokfulam, Hong Kong 999077, Hong Kong

J. Wang
Key Laboratory of Advanced Drug Delivery Systems of Zhejiang Province
College of Pharmaceutical Sciences
Zhejiang University
West Zunyi Road, Hangzhou, Zhejiang Province 310030, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202500235.

DOI: 10.1002/aisy.202500235

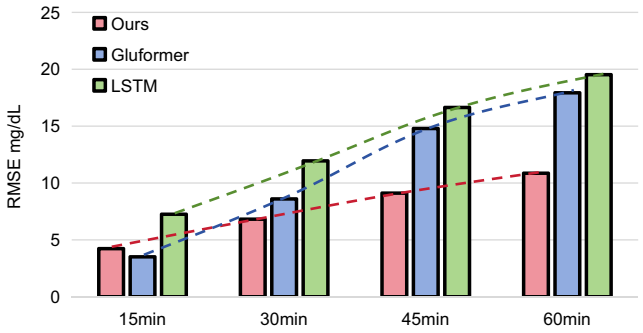

Figure 1. Prediction errors of the three models under different horizons of OhioT1DM. Dot-lines illustrate the error trend. Vanilla LSTM model commonly used in blood glucose prediction and the recent Gluformer model are selected.

accuracy degradation in longer prediction horizons (45 and 60 min). One perspective[18,20] is that the AR output modes in RNNs and vanilla Transformer architectures can lead to error spreading effects in prediction tasks, reducing the accuracy of tail-end forecasts. In the task of BG prediction, we have also observed this severe error spreading phenomenon caused by AR output (discussed in Section 6). This error spreading not only negates the advantages of long-term prediction in glucose management but may also lead to misjudgments in insulin dosage for individuals, resulting in serious consequences. Therefore, achieving more accurate LTPH BG performance becomes a challenging task.

To that end, we introduce a ProbSparse-Transformer structure[20,21] (as illustrated in **Figure 2**) to BG prediction tasks. By employing sparse sampling computation, the ProbSparse self-attention mechanism can successfully enhance the modeling capability for long sequences,[20] effectively alleviate errors in longer horizons, and significantly reduce the computational burden and memory requirements of the Transformer. To further increase the credibility of the ProbSparse self-attention in BG prediction, we quantify the uncertainty of the output distribution to enhance the prediction credibility. Meanwhile, to mitigate the issue of error spreading inherent in AR prediction modes, we have adopted a simple yet effective one-step generative head to generate prediction results. With only one-time inference to generate all predicted values, the one-step generation head further enhances the prediction efficiency.

We evaluate our proposed model on 4 datasets with 81 clinical individuals. The results show that our model demonstrates superior BG prediction capabilities, particularly excelling in LTPH tasks, compared to existing works. We also conduct transfer learning approach to improve its capability for personalized prediction and show the deployment capability on real edge devices. We believe the accurate long-term BG prediction can provide a valuable preemptive window, reducing the risk of abnormal BG levels and further advancing the treatment of diabetes.

2. Related Works

The prediction task is complicated by the highly nonlinear and nonstationary characteristics of BG variations. Nevertheless, analysis of extensive datasets of glucose measurements enables machine learning algorithms to discern complex patterns and relationships. Kamuran et al.[22] used the ARMAX prediction algorithm that utilizes the AR and MA components to model

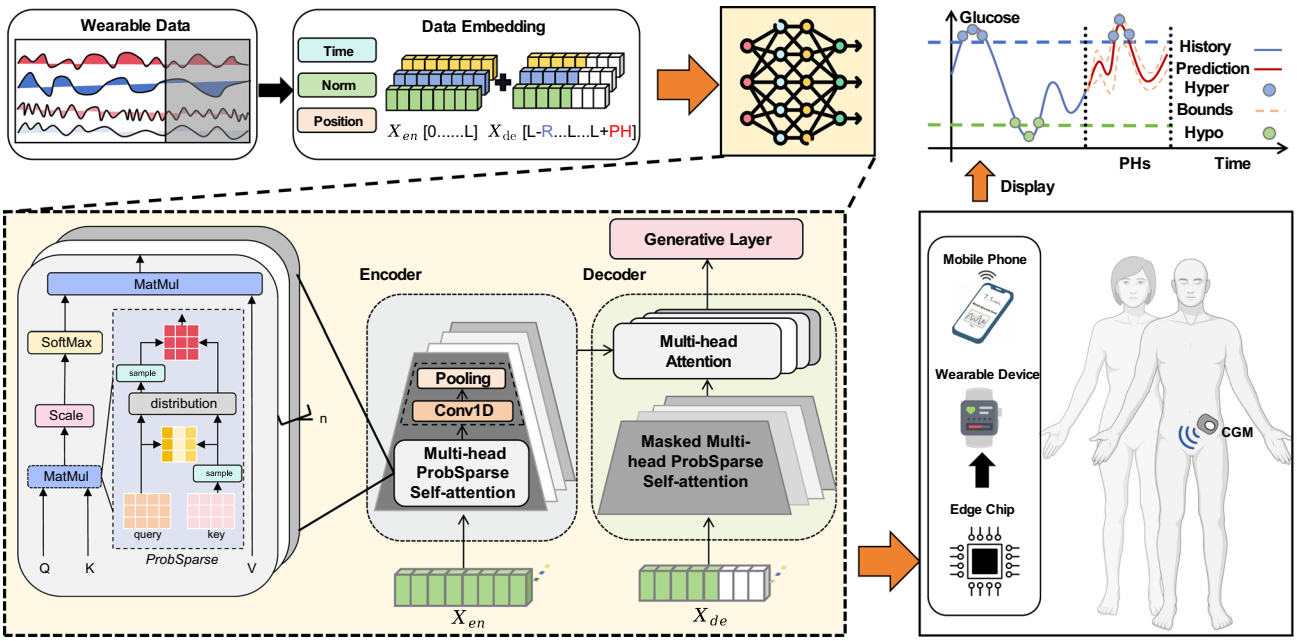

Figure 2. Framework of our BG prediction system. Wearable devices collect blood glucose and other physiological data, which serve as inputs to the prediction model. After data embedding, we introduce a structure utilizing ProbSparse self-attention and a one-step generative head within a Transformer-based model, which is concurrently designed for deployment on edge devices, enabling real-time analysis and providing precise BG monitoring for diabetic individuals.

the external condition (X) and accurately track BG variations. This method effectively addressed the nonlinear challenges in BG prediction and provided reliable warnings for the risk of hypoglycemia within the next 30 min (root mean square error [RMSE] = 18.55 mg dL^{-1}) and 60 min (RMSE = 38.06 mg dL^{-1}) in their experimental study.

The emergence of DNNs has markedly simplified complex prediction tasks, primarily due to their capacity to transform inputs into high-dimensional latent spaces and discern temporal associations.[6] In the realm of glucose forecasting, DNNs adeptly employ historical glucose data to aid in future predictions, as evidenced by previous studies.[23,24] This showcases their proficiency in detecting subtle patterns and trends that are essential for precise forecasting. Perez et al.[24] demonstrated predictive modeling efficacy using linear layers, achieving RMSE values of 9.74 mg dL^{-1} over a 15 min time window, 14.75 mg dL^{-1} over 30 min, and 25.08 mg dL^{-1} over 45 min. In contrast to conventional feed-forward neural networks, RNN features recursive connections that incorporate outputs from previous or future time steps as current inputs. However, standard RNNs often encounter issues with vanishing and exploding gradients during backpropagation. LSTM and GRU networks address these RNN challenges in long sequence prediction by employing a gating mechanism. Martinsson et al.[25] developed an RNN-based system for predicting BG over 30 and 60 min horizons, utilizing the OhioT1DM dataset.[26] They also quantified the uncertainty of their predictions by using the standard deviation (SD), derived from a parameterized univariate Gaussian distribution over the outputs. The mean and SD of the RMSE across six T1DM patients using their model were 18.867 ± 1.794 mg dl^{-1} and 31.403 ± 2.078 mg dl^{-1} for the 30 and 60 min, respectively. Sun et al.[27] integrated LSTM with bidirectional LSTM to enhance the prediction accuracy. This integration aimed at deepening the network architecture and improving its bidirectional sequence learning capabilities. Li et al.[28] proposed the GluNet prediction framework using CNNs structure, achieving an RMSE of 8.88 ± 0.77 mg dL^{-1} with a short time lag of 0.83 ± 0.40 min for a 30 min prediction horizon (PH) on a virtual dataset. Furthermore, they obtained an RMSE of 19.90 ± 3.17 mg dL^{-1} with a time lag of 16.43 ± 4.07 min for a 60-min PH, validated on a real dataset. Attention-based methods[29–31] have been used for glucose prediction. Zhu et al.[32] integrated RNN with attention to effectively capture both local and global information in BG sequences. Additionally, they developed the ARISES (Adaptive, Real-time, and Intelligent System to Enhance Self-care) system for implementing these advanced algorithms on a mobile phone platform. Furthermore, they employed a *meta*-learning strategy to transfer the model across datasets, achieving an RMSE of 35.28 ± 5.77 mg dL^{-1} for 60 min PH on a private dataset. The Transformer model adheres to the conventional encoder–decoder framework but distinguishes itself by employing multi-head self-attention mechanisms instead of vanilla RNN. This architecture excels in parallelism and adeptly captures global relationships within sequences across multiple semantic spaces. In the context of glucose prediction, Lee et al.[19] implemented the Transformer model to execute both prediction and classification tasks through autoregression. This model was validated on a private dataset and achieved a mean absolute percentage error (MAPE) of 17.88 in the OhioT1DM dataset. Sergazinov et al.[18] also employed the Transformer model for glucose prediction and innovatively modified the model's dropout layer to quantitatively assess its uncertainty. This adaptation significantly enhanced the Transformer's performance in BG prediction tasks. Fine-tuning one small-sampled data can enhance the model's capability for specific and personalized applications. Deng et al.[33] systematically studied three neural network architectures, various loss functions, four transfer learning strategies, and four data augmentation techniques, including hybrid and generative models. Through transfer learning, they improved the accuracy of predicting abnormal BG levels within an hour.

3. Methodology

As shown in Figure 2, we retain Transformer's encoder–decoder structure. Initially, we embed the BG and related physiological signals and then train model's prediction capabilities through the ProbSparse self-attention mechanism and an one-step generative head.[20]

3.1. Sequence Data Embedding

Contrary to RNN-based models, vanilla Transformers employ a pointwise self-attention mechanism and lack inherent temporal continuity information. Consequently, it is necessary to supplement them with specific location and timing data to maintain the time series' temporal properties.

Assuming we have the input $\chi^t = \{x_1^t, \ldots, x_L^t | x_i^t \in \mathbb{R}^D\}$ at time t. L is the input length, and D is the dimension of x_i^t, consisting of glucose data, insulin dose, carbohydrate intake, and other wearable data. We first preserve the local context by using a fixed position embedding and get an uniform representation of d_{model}. This module is aligned with the definition in Transformer.[17]

$$\begin{aligned} \text{PE}_{(pos,2i)} &= sin\left(\frac{pos}{2L^{2i/d_{\text{model}}}}\right) \\ \text{PE}_{(pos,2i+1)} &= cos\left(\frac{pos}{2L^{2i/d_{\text{model}}}}\right) \end{aligned} \tag{1}$$

where j $\in$ [1,...,d_{model}/2] and *pos* denotes the data position in input sequence. Additionally, the data from the public dataset, derived from clinical records, typically records input CGM at intervals of 5 or 15 min.[26,34–36] Therefore, as shown in Figure 2, we embedded the data by a time encoder. We utilized a input size of 60 for minute-level granularity (SE_{min}) and 24 for hourly granularity (SE_{hour}). Let date^t be the time stamps of χ^t from the wearable device.

$$e = \text{SE}_{\text{min}}(\text{date}^t) + \text{SE}_{\text{hour}}(\text{date}^t) \tag{2}$$

Due to the multimodal features of the input, we employed normalization to ensure uniformity in processing across different feature dimensions. Thus, we get the input

$$\chi_{\text{en}(i)}^t = Conv1D\big(\text{Norm}\big(\chi_i^t\big)\big) + e_i + \text{PE}_{(L\times(t-1)+i,)} \tag{3}$$

where *Conv*1D is a 1D convolution operator. The input data is embedded into a uniform latent space with a convolution layer,

so the *Conv*1D (kernel width = 3, stride = 1) have the same output dimension of d_{model}.

Figure 2 illustrates the standard encoder–decoder architecture, where the encoder's input, χ_{en}, is processed by embedding flow. And χ_{de} is the input of the decoder which is represented as

$$\chi_{\text{de}} = \text{Concat}\left(\left[\chi^t_{\text{en}(L-R)}, \ldots, \chi^t_{\text{en}(L)}\right], \left[o^t_{L+1}, \ldots, o^t_{L+PH}\right]\right) \tag{4}$$

where PH denotes the prediction horizon, R denotes a portion of the sequence intercepted from *L* that is closer to the current timepoint, and Concat(·) is the matrix splicing operation. The o sequence contains only the timestamp and relative position (e and PE) of the target sequence.

3.2. ProbSparse Self-Attention Structure

The standard multi-head self-attention mechanism utilizes Query (Q), Key (K), and Value (V) for calculating attention scores through scaled dot products. This process is often described as $\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$, where $Q \in \mathbb{R}^{L_Q \times d}$, $K \in \mathbb{R}^{L_K \times d}$, and $V \in \mathbb{R}^{L_V \times d}$. However, the atomic operations of self-attention require $O(L^2)$ in terms of time and memory complexity, which consequently results in significant computational consumption for Transformer on resource-constrained edge devices for CGMs.

Previous works[20,21,37] have attempted to uncover the inherent sparsity in the probability distribution of self-attention. Consequently, the query attention for each row can be defined as a probabilistic form of a kernel smoother.[38]

$$f(q_i) = \sum_j \frac{\exp\left(q_i k_j^T \Big/ \sqrt{d}\right)}{\sum_l \exp\left(q_i k_l^T \Big/ \sqrt{d}\right)} v_j \tag{5}$$

$$p(k_j|q_i) = \frac{\exp\left(q_i k_j^T \Big/ \sqrt{d}\right)}{\sum_l \exp\left(q_i k_l^T \Big/ \sqrt{d}\right)} \tag{6}$$

It was discovered that the attention scores exhibit a long-tail distribution effect,[20,39] whereby a small number of query weights predominantly dictate the generation of results. Then, selective strategies can be designed for $p(k_j|q_i)$ to reduce the computational redundancy of Transformer. Therefore, we employ the ProbSparse[20] self-attention mechanism to select core attentions, thereby decreasing computational complexity. This approach also reduces the risk of overfitting, enhancing robustness for clinical scenarios. The process of ProbSparse is shown in Figure 2 and Kullback–Leibler (KL) divergence was used to distinguish the salient queries.

$$F(q_i, K) = \max_j \left\{\frac{q_i k_j^T}{\sqrt{d}}\right\} - \frac{1}{L_K}\sum_{j=1}^{L_k} \frac{q_i k_j^T}{\sqrt{d}} \tag{7}$$

Zhou et al.[20] have proven that the computation of $F(q_i,K)$ holds the boundary relaxation and is also under the long tail distribution. So, we only randomly sample $U_K = c \times \log L_K$ weights from K to do dot-product for $F(q_i,K)$, where c is a constant. Finally, the attention computation is given as

$$\text{PSA}(Q, K, V) = \text{Softmax}\left(\frac{Q' K^T}{\sqrt{d}}\right)V \tag{8}$$

where Q' is the selected query weight matrix, which is of the same size as Q but contains only U_q queries. Under the control by c, $U_q = c \times \log L_Q$, and the remaining queries are not involved in the computation, being set to the mean of *V*. This approach ensures that the final computational complexity is only O(L log L). Finally a 1D convolution and pooling is applied in each encoder layer for downsampling to extract important attentional features in the BG time series, which can be seen in Figure 2. The first attention block in decoder uses the same ProbSparse self-attention as in encoder, and the second block is a vanilla cross-attention computation.

3.3. Uncertainty Estimation

Recent works[18,40] on time series prediction using autoregression Transformer is based on quantile regression for output computation, allowing the use of uncertainty analysis. Also, we observed that the initial sampling process of K in ProbSparse self-attention is random, which is akin to performing a dropout operation within the attention kernel during inference.[41–43] Such a computational pattern endows our model with uncertainty characteristics[42] (shown in **Figure 3**). We define that y^* is the observed output corresponding to input x^*, and the input and output sets are X, Y. Thus the approximate predictive distribution of is given by

$$q(y^*|x^*, X, Y) = \int p(y^*|x^*, w)p(w|X, Y)dw \tag{9}$$

where $w \in \{W_i\}_{i=1}^m$ is the initial random variables for a model with m layers. Typically, analytical evaluation of the distribution $p(w|X, Y)$ is difficult. Therefore, following the methodologies of previous work,[41] we define an approximate variational distribution q(w), which is structurally more amenable to evaluation. Our aim is for this approximate distribution to closely resemble the posterior distribution obtained from the full Gaussian process. To achieve this, we approximate the distribution through KL divergence

$$\text{KL}(q(w)|p(w|X, Y)) \tag{10}$$

leading in the approximate predictive distribution

$$q(y^*|x^*) = \int p(y^*|x^*, w)q(w)dw \tag{11}$$

Specifically, we sampled *T* sets of realization vectors from the Bernoulli distribution, denoted as $z_1^t, \ldots, z_{Lt=1}^{t\,T}$ under $\{W_1^t, \ldots, W_m^t\}_{t=1}^T$. So, we can estimate the expectation as

$$\mathbb{E}_{q(y^*|x^*)}(y^*) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{y}^*(x^*, W_1^t, \ldots, W_{\text{m}}^t) \tag{12}$$

This is a Monte Carlo integration.[43] Eventually, the network is approximated to fit the desired output by performing *T* forward

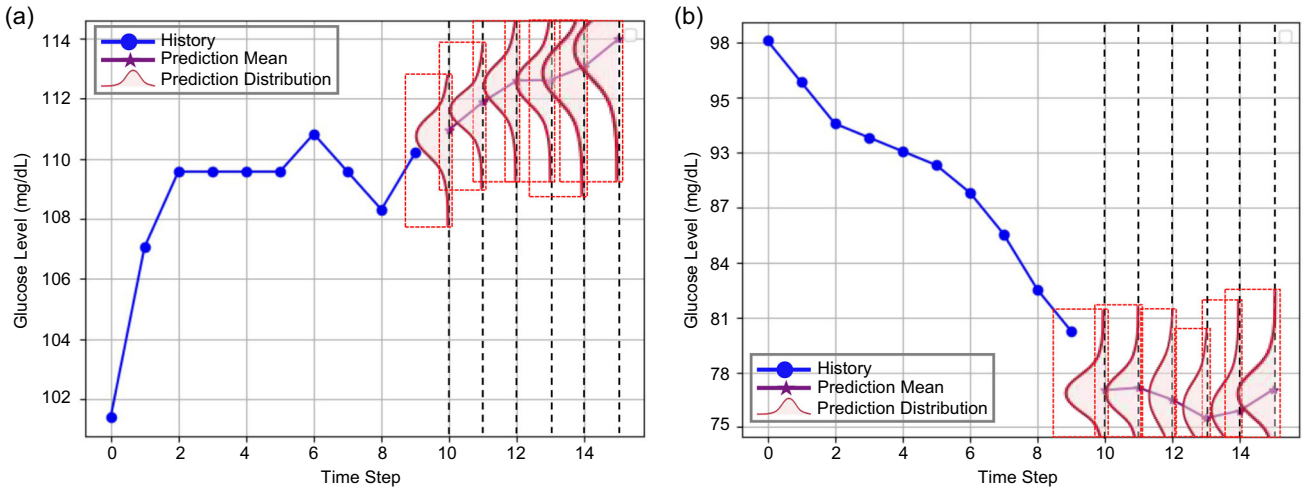

Figure 3. Uncertainty distribution schematics of forecast results.

passes in the inference phase and averaging the results as output. To quantify the uncertainty analysis of the model predictions and further improve the BG prediction confidence of the model, we estimate the prediction log-likelihood by Monte Carlo integration

$$\log p(y^*|x^*, X, Y) \approx \log\left(\frac{1}{T}\sum_{t=1}^{T} p(y^*|x^*, w_t)\right) \tag{13}$$

where $w_t \in q(w)$. And for the regression task[41] as BG prediction, we have

$$\log p(y^*|x^*, X, Y) \approx \log \text{sumexp}\left(-\frac{1}{2}\tau||y - y^*||^2\right) - \log T - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \tau^{-1} \tag{14}$$

where τ is the precision parameter. Excessive uncertainty, characterized by substantial observation noise or, conversely, minimal model precision τ, incurs significant penalties due to the final term in the predictive log-likelihood. Contrarily, a model exhibiting overconfidence, characterized by disproportionate precision in comparison to its mean estimation, is subject to penalization by the initial term. The integration of the ProbSparse self-attention mechanism into the Transformer model has facilitated the quantification and subsequent reduction of uncertainty, thereby augmenting the model's predictive reliability. This development holds significant implications in clinical settings, especially concerning therapeutic interventions and routine BG monitoring in patients.

3.4. One-Step Generative Head

To provide patients with sufficient time and early warnings for potential abnormal glucose events, it is necessary to predict future BG changes over as long a horizon as possible. Traditional methods predominantly employ RNN[14] and their variants[15,16] to develop sequence-to-sequence architectures. As shown on the left in **Figure 4**, whether it is RNN series models or the dynamic decoding form of the Transformer,[17] AR methods are used for

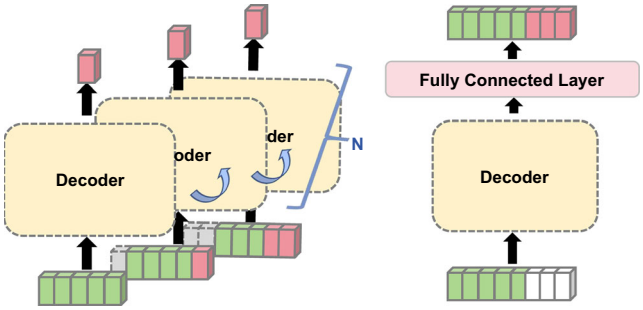

Figure 4. Comparison of autoregressive prediction (left) and one-step generative prediction (right).

prediction. While these can achieve precise short-term predictions, in LTPH tasks of 45 and 60 min, performance often degrades due to error spreading. Therefore, we adopt the same generative prediction head as recent works,[20,44,45] specifically through a learnable fully connected layer to generate future predictions in one step (Figure 4 right). The advantages of one-step head will be further discussed in Section 6.

4. Benchmark Construction

4.1. Datasets

To evaluate the performance of different BG prediction models, we developed a benchmark evaluation entirely based on open-source datasets. In pursuit of creating a balanced and exhaustive evaluation dataset, we incorporated three clinical datasets along with a dataset comprising virtual patients. **Table 1** lists the demographic and clinical characteristics of these datasets.

4.1.1. OhioT1DM Dataset

The OhioT1DM dataset[26] encapsulates a comprehensive collection of data from 12 individuals with T1D over an eight-week

Table 1. Datasets characteristics.

Characteristics	OhioT1DM	UVA-Padova	ShanghaiT1DM	D1NAMO
Age [years]	20–80	7–68	57.83 ± 11.12	20–79
Patients	12	30	30	9
Mean glucose level [mg dL^{-1}]	158.53(±16.90)	122.07(±12.67)	166.51(±31.94)	151.41(±36.43)
Median glucose level [mg dL^{-1}]	151.75(±18.52)	121.40(±15.36)	161.04(±32.63)	143.44(±37.14)
Insulin regimen (CSII/MDI)	MDI + CSII	CSII	CSII	MDI
Gender (female/male)	5/7	*/*	7/5	3/6
TIR [%]	62.6(±9.9)	85.4(±11.2)	54.7(±14.5)	57.7(±18.4)
TBR [%]	4.0(±3.1)	7.1(±7.0)	7.5(±7.0)	12.4(±18.6)
TAR [%]	33.4(±18.52)	7.5(±7.9)	37.8(±18.8)	30.0(±17.5)
Monitoring interval [min]	5	5	15	5
Total measurement length [h]	15 967	40 322	1308	1550

CSII: continuous subcutaneous insulin infusion; MDI: multiple daily injection.
TIR: time in range ($\geq$70, $\leq$180 mg dL^{-1}); TBR: time below range($<$70 mg dL^{-1}); and TAR: time above range($>$180 mg dL^{-1}).

period. This dataset includes CGM records, insulin dosing information, and physiological sensor data from devices like the Medtronic Enlite CGM and Medtronic 530 or 630 G insulin pumps. Some participants additionally utilized wearable devices such as the Basis Peak or Empatica Embrace wristbands to gather vital sign data. It is structured into a training set and a test set, constituting ≈80% and 20% of the total data, respectively.

4.1.2. D1NAMO

The D1NAMO dataset was collected as part of the D1NAMO project.[35] This dataset comprises diverse data from 29 individuals, including 20 healthy subjects and 9 diabetic subjects. All participants were equipped with uniform wearable devices, the Zephyr BioHarness 3 chest strap. The dataset is renowned for its comprehensiveness, encompassing not only CGM and insulin data but also 34 physiological indicators such as electrocardiogram (ECG) signals, respiratory patterns, accelerometer output, skin temperature, and annotated food images.

4.1.3. ShanghaiT1DM

The ShanghaiT1DM datasets[34] originate from 12 patients with type 1 diabetes in Shanghai. These datasets amalgamate a comprehensive array of information, including the patients' clinical profiles, laboratory results, and medication records. Notably, they feature CGM readings spanning 3 to 14 days, along with daily dietary data.

4.1.4. UVA-Padova

The UVA/Padova T1DMS[36,46] is an advanced simulation tool capable of accurately replicating real-life scenarios, including variations in dietary intake, scheduling, and insulin dosages. It also facilitates the detection and quantification of episodes of hyperglycemia and hypoglycemia. By precisely controlling experimental parameters and minimizing the calibration phase, the UVA/Padova T1DMS enhances the efficiency of diabetes research.

5. Experimental Section

5.1. Prediction Tasks and Metrics

Our analysis encompassed diverse PHs, including intervals of 15, 30, 45, and 60 min. In this pursuit, we adopted comprehensive metrics in the sphere of BG prediction. These included the RMSE, mean absolute error (MAE), and MAPE, calculated by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{15}$$

$$\mathrm{MAE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \overline{y}_i|} \tag{16}$$

$$\mathrm{MAPE} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \overline{y}}{y_i}\right|\right) \times 100\% \tag{17}$$

where y denotes the true value of BG and $\overline{y}$ denotes the forecasting outputs. We also included Clark error grid (CEG) to build a confident benchmark metric, which were critical factors in real-world applications, where the timeliness of predictions could have substantial implications on patient outcomes.[47]

5.2. Model Configurations

In UVA-Padova stimulator, we generated 12 virtual patients (4 children, 4 adolescents, and 4 adults) and produced 56 days of CGM data using Dexcom(seed = 1). The input for the OhioT1DM, UVA-Padova, and ShanghaiT1DM datasets comprised 3D time series data (CGM, insulin, carbohydrate intake), while the D1NAMO dataset includes 25D data (incorporating

additional physiological information). We engaged in multimodal signal learning, with the output being future BG. To preserve the originality and fairness of the datasets, no artificial preprocessing was applied to the data. All datasets were divided into training, validation, and test sets in a ratio of 6.4:1.6:2. The OhioT1DM dataset had already been split into training and test sets in an 8:2 ratio, and we further subdivided 20% of the training set to serve as the validation set.

Gluformer[18] and LSTM[15] were chosen as baseline models. Because Gluformer is the most recent BG prediction model based on Transformer and achieved state-of-the-art (SOTA) performance prior to our work, and LSTM had the widest application as the model backbone many BG prediction efforts. Then, we surveyed the latest and best-performing works on different datasets for comparison. Within the OhioT1DM dataset, we selected GluNet,[28] Fast-adaptive and Confident Neural Network (FCNN),[32] Seasonal trend integrated predictor (STIP),[48] and Transformer-Glucose[19] for comparison. In the UVA-Padova dataset, ISRKN[49] was chosen. Currently, there were no deep-learning models available for comparison in the D1NAMO dataset. For the ShanghaiT1DM dataset, Heterogeneous temporal representation (HETER) Predic[50] was chosen for comparison. Since some studies have not released their models and only validated on specific metrics and prediction horizons, we limited our comparison to the results reported in previous studies. The recent emergence of general time series models has demonstrated formidable predictive capabilities. TimesNet,[44] in particular, has exhibited superior performance across multiple time series datasets. Consequently, we also adapted TimesNet to the task of BG prediction, facilitating a comparison with our model.

We developed out model with Pytorch 2.0.1 on NVIDIA GTX 2080 Ti GPUs, and the Adam optimizer was used for training. The number of training epochs was set to 30, with a patience setting of 5, employing an early stop mechanism to reduce the risk of model overfitting and enhance generalization ability. Mean squared error was used as the universal loss function for model training.

5.3. Results

5.3.1. Prediction Performance on 4 Datasets

Table 2–5 respectively summarize the results of our model and the comparative models under 4 different PHs across the OhioT1DM, UVA-Padova, D1NAMO, and ShanghaiT1DM datasets, based on 5 evaluation metrics. Notably, in the LTPH tasks (45 min, 60 min) and the 30 min prediction task across the 4 datasets, our model achieved the best performance on all evaluation metrics. Compared to previous works in 60 min BG prediction on 4 datasets, our model achieved RMSE of 10.86, 15.33, 20.46, and 13.75 mg dL^{-1}, respectively, which demonstrated at least a 20% improvement than previous BG models and also achieved the superior leading results on MAE and MAPE. This indicated that our model's predictions more closely aligned with the actual trends in glucose changes. In the CEG analysis, the prediction errors of our model were concentrated in zone A, enhancing the safety of BG prediction by 0.8%–14.4%, which was highly significant for clinical diagnosis safety. Meanwhile, we observed that Gluformer achieved superior performance in the 15 min tasks on three datasets, but the gap with our method was minimal. Even our model showed higher safety in the CEG results for the 15 min prediction in OhioT1DM. This was attributed to Gluformer's AR prediction method, which exceled in shorter time horizon but could cause error spreading in the more crucial LTPH tasks. The TimesNet model also exhibited commendable performance in BG tasks, ranking just behind our model in LTPH predictions. Particularly in the 45 min prediction task for the D1NAMO dataset, it achieved a higher proportion in Zone A of the CEG. This was attributed to the InceptionNet[51] convolutional feature extraction capability incorporated in TimesNet, which enhanced its learning ability for datasets like D1NAMO that consisted of dozens of physiological features. However, this came at the cost of increased model size and longer inference time.

It was worth noting that our experiments revealed the same findings as previous study,[32] in the CEG, Zone C is typically very small, while Zone D exhibits spiked values. Zone D primarily represents scenarios where the actual CGM data indicate extremely high or low BG levels, far outside the normal range, while the model's predicted values remain within the standard range. Due to these extreme reference values, errors in Zone D tend to escalate significantly. In contrast, Zone C represents cases where the actual values are within the normal range, but the model's predictions deviated considerably. However, errors in Zone C were usually minimal because the model performed well within the normal BG range.

Figure 5 illustrates the 30 min BG prediction trajectories for a patient in the OhioT1DM dataset over 2 days. Compared to the traditional LSTM model (left panel), our model (right panel) demonstrated a superior ability to capture future trends and accurately predict hyperglycemic or hypoglycemic events within the next 30 min. A comparison of the confidence interval predictions for different models is presented in **Table 6**, where our architecture achieved more robust confidence levels across various prediction horizons. Moreover, **Figure 6** compares the RMSE performance of our model with Gluformer, TimesNet, and LSTM across 4 datasets. Consistent with the results in Table 1, 6, and 2, our model demonstrated lower prediction errors and higher predictive robustness. This was particularly evident in LTPH tasks, where our model showed lower mean (dashed line in Figure 6) and median (solid line in Figure 6) errors, as well as reduced prediction variance. In general, our model shows better performance than previous BG prediction works on 4 public CGM datasets, greatly improving the accuracy of the LTPH task.

5.3.2. Transfer and Fine-Tuning Results

Even though a model might achieve excellent performance on the test set within the same dataset, pretrained models still faced challenges when applied to real patient scenarios. An important solution was to use transfer learning and fine-tuning learning techniques[32,33] to achieve patient-based personalized predictions.

Then, we pretrained four representative models on 12 virtual patients in the UVA-Padova simulator, including vanilla LSTM, Gluformer, TimesNet, and the framework proposed in this work. We then conducted cross-validation tests on 12 actual patients in the OhioT1DM dataset. The results of the cross validation,

Table 2. Prediction performance on OhioT1DM.

Prediction Horizon	Methods	RMSE [mg dL^{-1}]	MAE [mg dL^{-1}]	MAPE [%]	CEG distribution [%]				
					A	B	C	D	E
15 min	LSTM	7.25	5.06	4.27	99.23	0.71	0.00	0.06	0.00
	Gluformer[18]	**3.51**	**3.51**	**2.03**	99.73	0.24	0.00	0.03	0.00
	TimesNet[44]	4.77	3.15	2.63	99.71	0.28	0.00	0.01	0.00
	Ours	4.23	2.72	2.29	**99.78**	0.21	0.00	0.01	0.00
30 min	GluNet[28]	10.73	–	–	–	–	–	–	–
	FCNN[32]	18.64	13.25	–	89.80	8.96	0.01	1.22	0.01
	STIP[48]	13.70	–	–	90.20	9.4	0.40	0.00	0.00
	LSTM	11.94	8.73	7.37	94.92	4.75	0.00	0.33	0.00
	Gluformer[18]	8.60	5.79	4.67	98.71	1.20	0.00	0.09	0.00
	TimesNet[44]	7.41	4.66	3.89	98.66	1.29	0.00	0.05	0.00
	Ours	**6.82**	**4.49**	**4.49**	**98.90**	1.07	0.00	0.03	0.00
45 min	LSTM	15.59	11.65	9.89	88.55	10.74	0.00	0.71	0.00
	Gluformer[18]	14.79	11.02	9.95	97.19	2.61	0.00	0.20	0.00
	TimesNet[44]	9.83	6.31	5.26	96.70	3.20	0.00	0.10	0.00
	Ours	**9.11**	**5.94**	**4.97**	**97.30**	2.62	0.00	0.08	0.00
60 min	GluNet[28]	22.65	–	–	–	–	–	–	–
	FCNN[32]	31.07	22.86	–	72.58	24.39	0.16	2.85	0.02
	Transformer-Glucose[19]	17.88	–	–	–	–	–	–	–
	STIP[48]	21.79	–	–	76.60	22.10	1.20	0.00	0.00
	LSTM	18.33	13.93	11.89	82.98	16.23	0.00	0.79	0.00
	Gluformer[18]	17.93	13.41	11.01	94.78	4.94	0.00	0.28	0.00
	TimesNet[44]	11.99	7.83	6.54	94.25	5.58	0.00	0.17	0.00
	Ours	**10.86**	**7.05**	**6.07**	**95.54**	4.19	0.00	0.03	0.00

The original work has not yet reported this performance metric.

displayed in **Figure 7**a, indicated that Gluformer maintained stable performance at the 15 min prediction interval but experienced significant losses at 45–60 min. TimesNet remained relatively stable, and our model exhibited higher cross-prediction accuracy overall. Figure 7b compares the accuracy drop with the results in Table 3, showing a performance decline across all models; however, our model consistently demonstrated stability across four prediction horizons. The results displayed in Figure 7 highlighted the robustness of our model in BG prediction, showcasing its predictive performance across different training and testing datasets. Additionally, although LSTM generally performed modestly, it did not exhibit significant performance degradation in cross-testing, likely due to its simpler structural design and fewer involved parameters.

We then fine-tuned the models that were pretrained on silicon patients. The specific steps involved were as follows: first, we selected data from 10 days from 12 patients in OhioT1DM training dataset. This data was then used to fine-tune the pretrained models. Our model utilized a decoder-only fine-tuning approach, while the LSTM, Gluformer, and TimesNet models underwent full-model fine-tuning. Finally, we validated them on the test dataset. As illustrated in Figure 7c, after fine-tuning with a small sample, our model still exhibits the best performance across different PHs, with a notably superior lead in LTPH. In Figure 7d, we customized the pretrained prediction model for each patient. Post-finetuning, it was observed that the model maintained a lower RMSE error. However, the unified fine-tuning model for the 12 patients demonstrated higher prediction accuracy than individual patient fine-tuning, likely due to the larger data volume and more diverse BG dynamic features provided by regional and group data, resulting in more effective fine-tuning (as shown in Figure 7c). This insight suggested that in the future, group-based personalized fine-tuning for patients with similar ages and physiological characteristics could enhance the precision and individualization of BG predictions.

In our approach to transfer learning, we fine-tuned only the decoder while keeping the encoder parameters unchanged. This method required adjusting only a small number of parameters, allowing the model to adapt to new data more quickly. We compared three strategies: using only the encoder, using only the decoder, and fine-tuning the entire model. **Table 7** demonstrates that fine-tuning only the decoder achieved performance comparable to full-model of encoder-only fine-tuning on small datasets, while preserving the model's generalization ability on large-scale data.

Table 3. Prediction performance on UVA-Padova.

Prediction Horizon	Methods	RMSE [mg dL^{-1}]	**MAE** [mg dL^{-1}]	**MAPE** [%]	CEG distribution [%]				
					A	B	C	D	E
15 min	LSTM	16.81	12.31	15.43	0.00	85.36	11.84	2.80	0.00
	Gluformer[18]	**5.21**	**4.07**	**6.04**	97.77	1.76	0.00	0.47	0.00
	TimesNet[44]	8.79	5.97	8.04	96.76	2.43	0.00	0.81	0.00
	Ours	5.48	4.79	6.10	**97.88**	1.59	0.00	0.53	0.00
30 min	ISRKN[49]	11.14	–	–	–	–	–	–	–
	LSTM	26.20	19.16	25.07	68.09	24.24	0.00	7.67	0.00
	Gluformer[18]	12.81	10.23	11.57	91.64	6.71	0.00	1.65	0.00
	TimesNet[44]	14.08	9.66	12.85	89.83	8.00	0.00	2.16	0.00
	Ours	**10.07**	**7.65**	**10.48**	**92.06**	5.98	0.00	1.96	0.00
45 min	LSTM	31.23	22.89	31.02	60.32	29.09	0.00	10.57	0.03
	Gluformer[18]	18.79	14.19	14.67	87.49	10.17	0.00	2.34	0.00
	TimesNet[44]	17.13	12.01	16.20	84.96	11.91	0.00	3.12	0.00
	Ours	**13.42**	**8.48**	**13.71**	**88.03**	8.90	0.00	3.07	0.00
60 min	ISRKN[49]	16.00	–	–	–	–	–	–	–
	LSTM	34.28	25.7	36.46	54.58	31.68	0.02	13.68	0.03
	Gluformer[18]	25.16	19.41	22.08	83.71	13.19	0.00	3.09	0.01
	TimesNet[44]	19.46	13.79	18.67	81.04	15.00	0.00	3.95	0.00
	Ours	**15.33**	**11.21**	**16.78**	**85.33**	10.92	0.00	3.75	0.00

The original work has not yet reported this performance metric.

Table 4. Prediction performance on D1NAMO.

Prediction Horizon	Methods	RMSE [mg dL^{-1}]	MAE [mg dL^{-1}]	MAPE [%]	CEG distribution [%]				
					A	B	C	D	E
15 min	LSTM	15.57	12.89	13.40	81.38	8.54	0.00	10.08	0.00
	Gluformer[18]	9.47	**7.02**	**6.69**	**96.67**	2.52	0.00	0.81	0.00
	TimesNet[44]	10.87	7.45	7.10	93.83	5.53	0.00	0.64	0.00
	Ours	**9.41**	7.11	6.95	95.71	3.24	0.00	1.04	0.00
30 min	LSTM	19.54	15.81	16.24	71.71	18.38	0.00	9.91	0.00
	Gluformer[18]	18.29	12.49	10.46	86.46	11.96	0.00	1.58	0.00
	TimesNet[44]	14.54	9.85	9.48	88.98	10.00	0.00	1.00	0.01
	Ours	**14.31**	**9.73**	**9.98**	**86.62**	10.55	0.00	2.83	0.00
45 min	LSTM	23.37	18.64	19.12	64.83	25.16	0.00	10.01	0.00
	Gluformer[18]	21.48	16.20	13.98	77.83	17.64	0.00	4.51	0.01
	TimesNet[44]	17.69	11.84	12.37	**83.92**	14.34	0.00	1.71	0.03
	Ours	**17.58**	**12.89**	**12.64**	78.48	17.64	0.00	3.88	0.00
60 min	LSTM	26.29	20.5	21.08	59.87	30.16	0.00	9.97	0.00
	Gluformer[18]	24.91	19.22	17.28	78.75	18.77	0.01	2.42	0.05
	TimesNet[44]	21.12	14.08	15.29	74.03	21.43	0.00	4.53	0.00
	Ours	**20.46**	**13.79**	**19.67**	**81.04**	15.00	0.00	3.95	0.00

The original work has not yet reported this performance metric.

Table 5. Prediction performance on ShanghaiT1DM.

Prediction Horizon	Methods	RMSE [mg dL−1]	MAE [mg dL^{-1}]	MAPE [%]	CEG distribution [%]				
					A	B	C	D	E
15 min	HETER[50]	5.70	4.06	3.30	–	–	–	–	–
	LSTM	7.54	5.35	4.10	99.04	0.89	0.00	0.07	0.00
	Gluformer[18]	9.47	7.02	6.69	96.67	2.52	0.00	0.81	0.00
	TimesNet[44]	6.81	4.85	3.79	99.26	0.51	0.00	0.23	0.00
	Ours	**5.32**	**3.76**	**3.01**	**99.77**	0.23	0.00	0.00	0.00
30 min	HETER[50]	10.11	6.72	5.40	–	–	–	–	–
	STIP[48]	14.56	–	–	90.60	8.90	0.50	0.00	0.00
	LSTM	13.96	10.00	7.84	93.75	5.75	0.00	0.50	0.00
	Gluformer[18]	9.26	6.21	4.68	97.79	2.16	0.00	0.05	0.00
	TimesNet[44]	14.54	9.85	9.82	88.98	10.00	0.00	1.00	0.01
	Ours	**8.32**	**5.47**	**4.36**	**98.52**	1.34	0.00	0.14	0.00
45 min	LSTM	18.98	13.67	10.82	86.73	12.51	0.00	0.75	0.00
	Gluformer[18]	21.21	14.28	14.21	78.75	18.77	0.01	2.42	0.05
	TimesNet[44]	12.17	7.99	6.02	95.29	4.59	0.00	0.12	0.00
	Ours	**11.19**	**7.43**	**5.72**	**96.17**	3.56	0.00	0.28	0.00
60 min	HETER[50]	16.59	10.98	8.90	–	–	–	–	–
	STIP[48]	24.12	–	–	75.70	21.90	2.30	0.10	0.00
	LSTM	23.08	16.81	13.45	79.70	19.34	0.00	0.97	0.00
	Gluformer[18]	21.48	16.20	13.98	77.83	17.64	0.00	4.51	0.01
	TimesNet[44]	15.17	9.97	7.52	92.42	7.40	0.00	0.18	0.00
	Ours	**13.74**	**9.01**	**6.94**	**94.10**	5.55	0.00	0.34	0.00

The original work has not yet reported this performance metric.

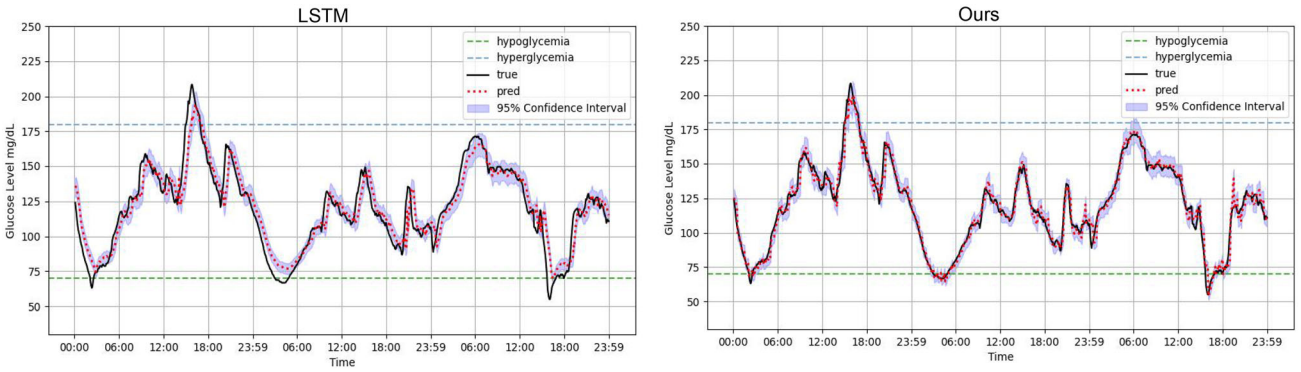

Figure 5. Prediction visualization (PH = 30 min). The left figure visualizes the LSTM prediction results, while the right side depicts our model. The red-dashed line represents the predicted BG values, the solid black line indicates the actual BG levels, the green-dashed line marks the hypoglycemic threshold, and the blue-dashed line denotes the hyperglycemic threshold. The purple-shaded area signifies the 95% prediction confidence interval.

5.3.3. Model Quantization and Deployment

Edge computing needed more reliable real-time services on wearable devices with exceptionally low latency in output, unfettered by internet connectivity constraints. Such advancements were particularly crucial in the realm of intelligent BG management systems, catering to both everyday monitoring and clinical treatment. Our aim is to deploy these networks on edge devices, like smartphones, smartwatches, and CGMs, to facilitate real-time computation.[52,53] As shown in **Figure 8**, we quantized 32-bit floating-point arrays into 8-bit and 4-bit fixed-point by post-trained quantization methods with per-tensor type. This process substantially reduced the model size while maintaining stable accuracy,[54] which achieved the 3.67 and 6.13 times model

Table 6. Comparison of 95% confidence interval coverage proportions on OhioT1DM.

Prediction horizons	LSTM [95%]	Gluformer [95%]	TimesNet [95%]	Ours [95%]
15 min	89.0	94.5	93.0	**93.8**
30 min	85.5	90.8	92.0	**92.5**
45 min	80.5	85.0	91.5	**92.5**
60 min	76.0	80.0	88.1	**90.2**

size compression. On edge device inference, fixed-point computations enhanced processing speed and efficiency.[55]

We selected a field-programmable gate array (FPGA) as the platform for deploying quantized models on edge devices due to its high flexibility and low power consumption. These characteristics made it particularly suitable for deep-learning computations in edge environments, aligning with the requirements of edge-based boundary grid computing scenarios.[56,57] Deployment verification was conducted on the ECE-EMBD development board equipped with a ZYNQ 7020 chip. This board featured a dual-core ARM Cortex-A9 processor, a processing system with 512 MB DDR3 memory, and programmable logic equipped with an XC7Z020-CLG400-1 series chip. The model was deployed on the FPGA platform using a custom toolchain.[56] **Table 8** illustrates the accuracy (30 min) and model size compression following the quantitative deployment of both Int 8 and Int 4 bit width.

Specifically, to achieve on-chip deployment,[56] a two-step process is employed: first, performing 8-bit quantization and compilation after model training, and then converting each layer into its respective arithmetic module. To efficiently deploy neural network operators, the Img2Col operation was applied, converting layer data from noncontiguous to contiguous storage, thereby simplifying data transfer and computation. Parallel stacking and cascading mechanisms for multiplication and addition were then employed to enable simultaneous element-wise computations on large datasets. Large matrices were divided into smaller blocks, transferred to the chip, processed within these blocks, and subsequently reassembled, alleviating the resource constraints of the FPGA. As illustrated in Figure 8, we began by quantizing the model weights

$$W_{\text{int}} = \text{round}\left(\frac{W_{\text{float}}}{\text{s}}\right) \tag{18}$$

$$W_{\text{q}} = \text{clamp}\left(-2^{b-1}, 2^{b-1}-1, W_{\text{int}}\right) \tag{19}$$

where

$$\text{clamp}(a,b,x) = \begin{cases} a & if \quad W_i \le a \\ W_i & if \quad a \le W_i \le b \\ b & if \quad W_i \ge b \end{cases} \tag{20}$$

where W_{float} denotes the original float type weights; s represents the scale factor, which maps floating-point values to integers and is generally set as the maximum absolute value in the weight matrix; b denotes the quantization bit width, such as 4 or 8 bits; and W_{q} is the weight matrix post-quantization.

6. Discussion

We evaluated all the metrics on public datasets, enabling a comprehensive and fair comparison of BG prediction models, along with cross-validation and transfer fine-tuning. From the experimental analysis, it is evident that our proposed model effectively addresses the LTPH issue in BG prediction tasks. Compared to existing BG prediction models and other time series models, our model demonstrates leading performance across all 4 datasets. In

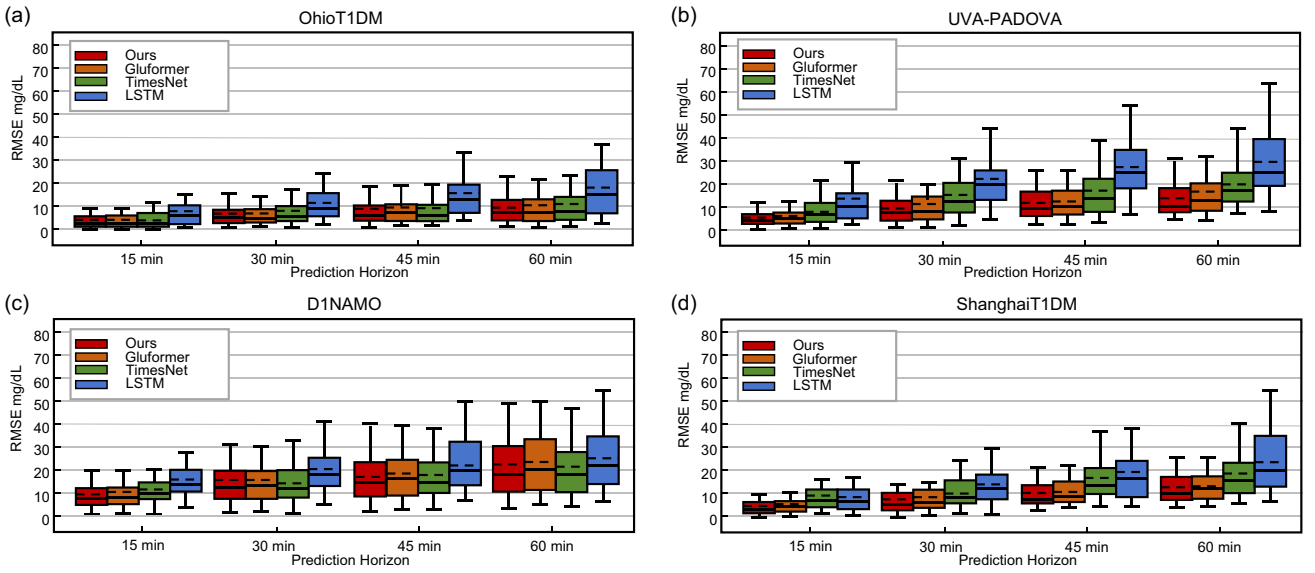

Figure 6. Comparative analysis of glucose level prediction performance. a–d) The RMSE box plots of four models across four time windows (15, 30, 45, 60 min). Dashed lines represent the mean RMSE values, solid lines indicate the median RMSE, and the boundaries of the box plots denote the error range. Our model consistently exhibits the lowest error across all evaluations and demonstrates a more concentrated and stable prediction range.

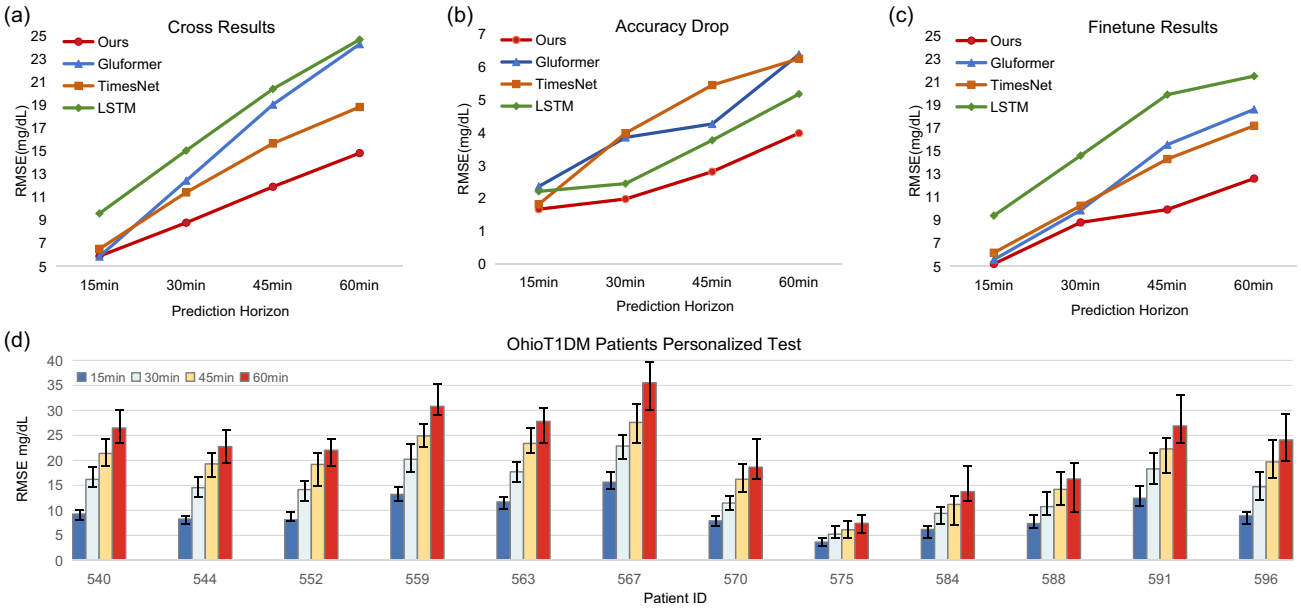

Figure 7. Transfer and fine-tuning comparison. a) The error of the 4 models trained in virtual patient on OhioT1DM dataset; b) the loss of 3 models performance tested across datasets; c) the prediction error after fine-tuning; and d) a plot of prediction error of our model after personalized fine-tuning by 10 days of specific patient data.

Table 7. Parameters and RMSE loss for transfer learning of different modules across datasets.

Tuning module	None	Full-model	Encoder-only	Decoder-only
Parameters [Mb]	–	46.57	18.21	**28.36**
RMSE [mg dL^{-1}, 60 min]	14.84	13.32	15.60	**12.63**

Table 8. Quantized performance of our proposed model on different edged devices.

Precision	RMSE 30 min [mg dL^{-1}]	Model size [Mb]	Compression ratio	Device
Ours(Float 32)	6.82	46.57	1.00×	Graphics Processing Unit (GPU)
Ours(Int 8)	6.83	12.68	3.67×	FPGA
Ours(Int 4)	7.51	7.63	6.13×	FPGA

the prediction of 60 min BG trends, RMSE values were obtained at 10.86, 15.33, 20.46, and 13.74 mg dL^{-1}, respectively. The maximum reduction in the error of glucose LTPH prediction was 39.4%. We attribute the superiority of our model's performance to two main factors.

First, the application of one-step generative head allows the model's output to be generated in one time. While traditional AR methods may perform well in short-term predictions horizons, the accumulation of errors in long-term can impact accuracy and introduce more risks in clinical applications. As shown in **Figure 9**, we compared the error spreading effect of Transformer using an AR head (top) versus an one-step generative head (bottom). The single-step generation head model maintains a relatively stable error at the 9th (45 min) and 12th (60 min) steps, whereas the error in the AR prediction head accumulates and diffuses, leading to performance loss in LTPH tasks. **Table 9** also highlights the differences in inference iterations and accuracy between the two prediction heads. Single-step inference enables more efficient and faster predictions while maintaining lower long-term error, whereas the AR method incurs higher computational costs. Specifically, for short-term predictions (e.g., prediction windows less than 15 min), the RMSE for the AR head is 3.60, compared to 4.35 for the single-step head—a difference of less than 1 mg dL^{-1}. However, for longer prediction windows (e.g., step sizes greater than 6 and up to 12), the single-step head significantly outperforms the AR head. This

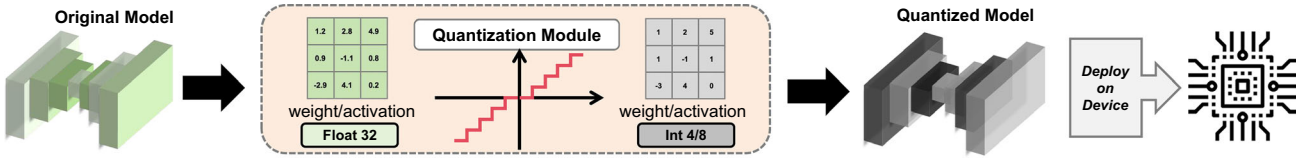

Figure 8. Network quantization and deployment flow. The full-accuracy model compresses the weights to 4/8 bits through quantization and can subsequently be deployed on edge computing chips for real-time glucose prediction.

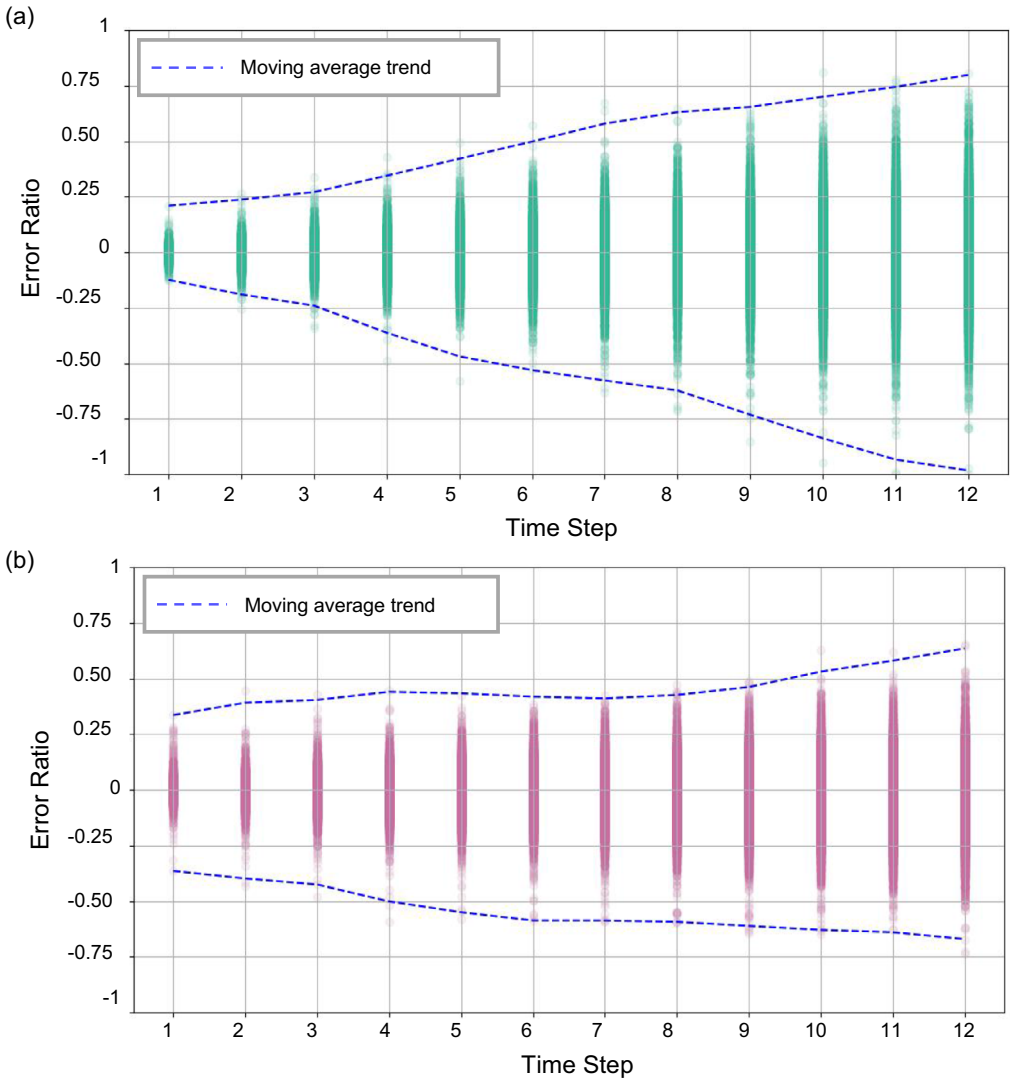

Figure 9. Illustration of prediction error spreading. The upper graph shows the error accumulation in autoregressive predictions, while the lower graph depicts the error accumulation in one-step generative head predictions. The vertical axis represents the ratio of each sample's error to the maximum RMSE in 60 min prediction (where the maximum error is the highest RMSE observed in two models).

indicates that the single-step generation head excels in long-term and more challenging prediction tasks, which is critical for BG forecasting. In BG monitoring and alert systems, the ability to predict abnormalities over longer time windows is particularly important for improving patient treatment and preventive care.

The second factor is the uncertainty-estimated analysis based on the ProbSparse self-attention structure, which has enhanced the predictive confidence of the model. The random selection of key in the ProbSparse self-attention computation introduces uncertainty distributions into the model's network. By quantifying the uncertainty of the output, we have reduced the error risk of the model's predictions, further ensuring the robustness of the model. **Figure 10** illustrates the calibration plot for the 60 min predictions of our model on the OhioT1DM dataset. We observe that the confidence levels of the model outputs are close to the expected levels, demonstrating the exceptional predictive robustness of our model. In the output phase, we employed the average of 5 sampling iterations as the prediction result. Although this introduces additional computational overhead,

Table 9. Error spreading analysis.

PHs [min]	Autoregression head		One-step generative head	
	RMSE 60 min [mg dL^{-1}]	Inference Steps	RMSE 60 min [mg dL^{-1}]	Inference Steps
15	3.6	3	4.35	1
30	8.55	6	6.97	1
45	12.38	9	9.16	1
60	16.25	12	11.34	1

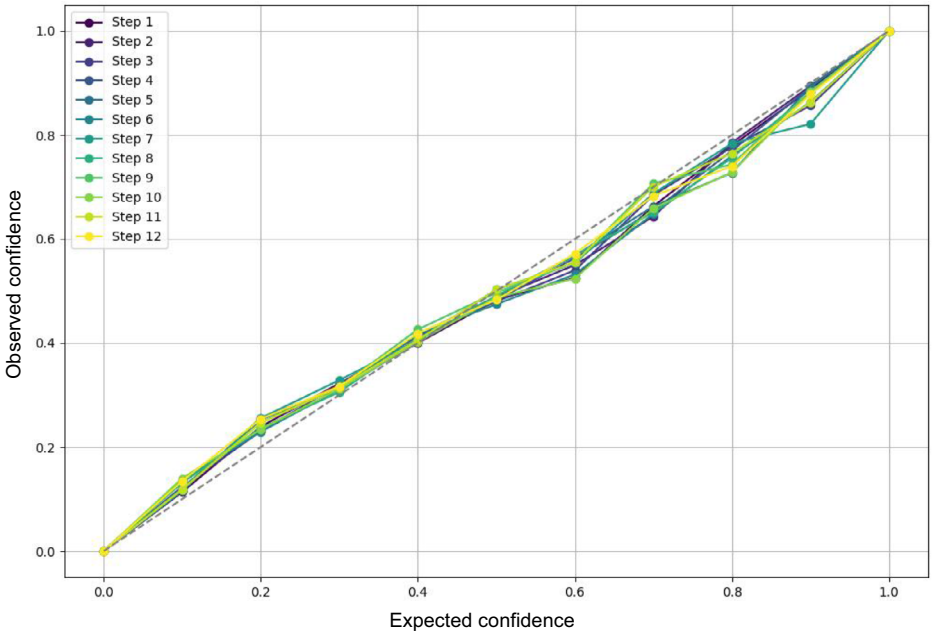

Figure 10. Calibration evaluation on 12 steps prediction (60 min). The curves is tested on OhioT1DM dataset and the gray dot-line (45°) shows the perfect calibration curve.

it still retains an advantage in comparison to the computational consumption of the AR prediction in LTPH. Simultaneously, due to the unique filtering mechanism employed in the ProbSparse self-attention computation, redundant calculations in the Transformer model are reduced. The computational time complexity is decreased from $O(L^2)$ to $O(L\ln L)$. This improvement makes the deployment of Transformer models on edge devices more feasible and user-friendly.

Finally, we validated the robustness of our model through comparative experiments and further confirmed its generalization ability and clinical application potential through fine-tuning with a small sample size. Our model demonstrates robust generalization and transfer capabilities, as evidenced by its exemplary performance on clinical datasets after being trained on a virtual patient dataset, shown in Figure 7a,c. The model's performance is further enhanced after fine-tuning. We observed that models achieving SOTA performance in the general time series prediction domain also exhibit commendable performance. As indicated in Table 5, the TimesNet model exhibits strong learning abilities in multivariable features, achieving a leading position in the 45 min CEG evaluation.

Additionally, to assess the real-time inference capabilities of BG prediction models on actual edge chips, we deployed our compressed model on an FPGA, verifying its inference capabilities in mobile scenarios. This confirms that our model is well-suited for edge computing in smartphones, smartwatches, or CGM devices, meeting the daily usage needs of patients and truly realizing intelligent BG prediction in daily life.

7. Conclusion

In this study, we address the challenge of enhancing the safety of closed-loop CGM systems by improving the accuracy of long-term glucose predictions. We propose a uncertainty-estimated ProbSparse-Transformer for BG prediction, yielding significant enhancements in LTPH tasks. This model integrates a ProbSparse self-attention mechanism with uncertainty analysis and a one-step generative head, adeptly countering the error spreading issues commonly encountered in RNNs and vanilla Transformers. Moreover, recognizing the constraints of limited dataset sizes and the need for personalized predictions, we refined our model's capabilities through fine-tuning with a small sample size. Finally, we established an open-source evaluation benchmark comprising 4 public datasets and 5 predictive metrics, contributing to the field of intelligent closed-loop CGMs to reduce diabetes-related risks.

Acknowledgements

The authors acknowledge the Innovation and Technology Fund (Mainland-Hong Kong Joint Funding Scheme, MHP/053/21) and Shenzhen-Hong Kong-Macau Technology Research Programme (grant no. SGDX20210823103537034) for supporting this work.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Wei Huang: conceptualization (equal); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (lead); visualization (lead); writing—original draft (equal); and writing—review and editing (equal). **Ni Fan**: data curation (lead); formal analysis (equal); and writing—original draft (supporting). **Weiping Wang**: conceptualization (supporting); data curation (lead); and formal analysis (supporting). **Jinqiang Wang**: conceptualization (equal); formal analysis (equal); and

funding acquisition (equal). **Xiaojuan Qi**: formal analysis (lead); methodology (lead); software (equal); writing—original draft (lead); and writing—review and editing (equal). **Shiming Zhang**: conceptualization (lead); data curation (lead); formal analysis (lead); funding acquisition (lead); investigation (lead); methodology (supporting); writing—original draft (equal); and writing—review and editing (equal).

Data Availability Statement

Data available on request due to privacy/ethical restrictions.

Keywords

digital healths, edged computings, glucose predictions, intelligent medicines, wearables

Received: February 27, 2025
Revised: April 30, 2025
Published online: June 25, 2025

[1] A. M. N. Renzaho, GBD 2021 *DIabetes Collaborators*. **2021**, *402*, 10397.
[2] Emerging Risk Factors Collaboration, *The Lancet* **2010**, *375*, 2215.
[3] M. Tancredi, A. Rosengren, A.-M. Svensson, M. Kosiborod, A. Pivodic, S. Gudbjörnsdottir, H. Wedel, M. Clements, S. Dahlqvist, M. Lind, *New Engl. J. Med.* **2015**, *373*, 1720.
[4] C. Gorst, C. S. Kwok, S. Aslam, I. Buchan, E. Kontopantelis, P. K. Myint, G. Heatlie, Y. Loke, M. K. Rutter, M. A. Mamas, *Diabetes Care* **2015**, *38*, 2354.
[5] R. Sergazinov, E. Chun, V. Rogovchenko, N. Fernandes, N. Kasman, I. Gaynanova, Arxiv:2410.05780, **2024**.
[6] T. Zhu, K. Li, P. Herrero, P. Georgiou, *IEEE J. Biomed. Health Inf.* **2020**, *25*, 2744.
[7] E. Chun, N. J. Fernandes, I. Gaynanova, *Diabetes Technol. & Ther.* **2024**, *26*, 939.
[8] W. Huang, I. Pang, J. Bai, B. Cui, X. Qi, S. Zhang, *Adv. Intell. Syst.*, **2025**, 2400822.
[9] M. Eren-Oruklu, A. Cinar, L. Quinn, D. Smith, *Diabetes Technol. & Ther.* **2009**, *11*, 243.
[10] E. I. Georga, V. C. Protopappas, D. Ardigò, D. Polyzos, D. I. Fotiadis, *Diabetes Technol. & Ther.* **2013**, *15*, 634.
[11] J. Li, C. Fernando, *ICT Express* **2016**, *2*, 150.
[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, *Neural Comput.* **1989**, *1*, 541.
[13] J. Bouvrie, *Notes On Convolutional Neural Networks*. Citeseer, New Jeresey, USA **2006**.
[14] W. Zaremba, I. Sutskever, O. Vinyals. arXiv preprint arXiv:1409.2329, **2014**.
[15] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735.
[16] J. Chung, C. Gulcehre, K. H. Cho, Y. Bengio. NIPS 2014 Workshop on Deep Learning. **2014**.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
[18] R. Sergazinov, M. Armandpour, I. Gaynanova, in *ICASSP 2023-2023 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE **2023**, pp. 1–5.
[19] S.-M. Lee, D.-Y. Kim, J. Woo, *IEEE J. Biomed. Health Inf.* **2023**, *27*, 1600.
[20] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, in *Proc. of the AAAI Conf. on Artificial Intelligence* **2021**, Vol. *35*, pp. 11106–11115.
[21] R. Child, S. Gray, A. Radford, I. Sutskever, arXiv preprint arXiv:1904.10509, **2019**.
[22] K. Turksoy, E. S. Bayrak, L. Quinn, E. Littlejohn, D. Rollins, A. Cinar, *Indus. & Eng. Chem. Res.* **2013**, *52*, 12329.
[23] C. Zecchin, A. Facchinetti, G. Sparacino, G. D. Nicolao, C. Cobelli, *IEEE Trans. Biomed. Eng.* **2012**, *59*, 1550.
[24] C. Pérez-Ganda, A. Facchinetti, G. Sparacino, C. Cobelli, E. J. Gómez, M. Rigla, A. de Leiva, M. E. Hernando, *Diabetes Technol. & Ther.* **2010**, *12*, 81.
[25] J. Martinsson, A. Schliep, B. Eliasson, O. Mogren, *J. Healthcare Inf. Res.* **2020**, *4*, 1.
[26] C. Marling, R. Bunescu, in *CEUR Workshop Proc.*, NIH Public Access **2020**, Vol. *2675*, p. 71.
[27] Q. Sun, M. V. Jankovic, L. Bally, S. G. Mougiakakou, in *2018 14th symposium on neural networks and applications (NEUREL)*, IEEE **2018**, pp. 1–5.
[28] K. Li, C. Liu, T. Zhu, P. Herrero, P. Georgiou, *IEEE J. Biomed. Health Inf.* **2019**, *24*, 414.
[29] S. Mirshekarian, H. Shen, R. Bunescu, C. Marling, in *2019 41st Annual Inter. Conf. of the IEEE Engineering In Medicine And Biology Society (EMBC)*, IEEE **2019**, pp. 706–712.
[30] M. Armandpour, B. Kidd, Y. Du, J. Z. Huang, in *2021 Inter. Joint Conf. on Neural Networks (IJCNN)*, IEEE **2021**, pp. 1–8.
[31] A. Bertachi, L. Biagi, I. Contreras, N. Luo, J. Veh, *KHD@ IJCAI* **2018**, pp. 85–90.
[32] T. Zhu, K. Li, P. Herrero, P. Georgiou, *IEEE Trans. Biomed. Eng.* **2022**, *70*, 193.
[33] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, C. S. Mantzoros, *NPJ Digital Med.* **2021**, *4*, 109.
[34] Q. Zhao, J. Zhu, X. Shen, C. Lin, Y. Zhang, Y. Liang, B. Cao, J. Li, X. Liu, W. Rao, C. Wang, *Sci. Data* **2023**, *10*, 35.
[35] F. Dubosson, J.-E. Ranvier, S. Bromuri, J.-P. Calbimonte, J. Ruiz, M. Schumacher, *Inf. Med. Unlocked* **2018**, *13*, 92.
[36] C. Dalla Man, M. D. Breton, C. Cobelli, *Physical Activity Into The Meal Glucose—Insulin Model Of Type 1 Diabetes: In Silico Studies*, **2009**.
[37] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, X. Yan, *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
[38] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, R. Salakhutdinov, Conf. on Empirical MEthods in Natural Language Process. **2019**, pp. 4344 -4353.
[39] I. Beltagy, M. E. Peters, A. Cohan, arXiv preprint arXiv:2004.05150, **2020**.
[40] B. Lim, S. Ö. Ark, N. Loeff, T. Pfister, *Int. J. Forecast.* **2021**, *37*, 1748.
[41] Y. Gal, Z. Ghahramani, in *Inter. Conf. on Machine Learning*, PMLR **2016**, pp. 1050–1059.
[42] K. A. Sankararaman, S. Wang, H. Fang, arXiv preprint arXiv:2206.00826, **2022**.
[43] A. Damianou, N. D. Lawrence, *Artificial Intelligence And Statistics*, PMLR **2013**, pp. 207–215.
[44] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, M. Long. arXiv preprint arXiv:2210.02186, **2022**.
[45] A. Zeng, M. Chen, L. Zhang, Q. Xu, in *Proc. of the AAAI Conf. on Artificial Intelligence* **2023**, Vol. *37*, pp. 11121–11128.
[46] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, C. Cobelli, *J. Diabetes Sci. Technol.* **2014**, *8*, 26.
[47] A. Gani, A. V. Gribok, Y. Lu, W. K. Ward, R. A. Vigersky, J. Reifman, *IEEE Trans. Inf. Technol. Biomed.* **2009**, *14*, 157.
[48] W. Rao, G. Yang, Q. Zhao, Y. Liu, H. Zhu, M. Li, X. Li, Y. Zhang, in *Inter. Conf. on Advanced Data Mining and Applications*, Springer **2023**, pp. 437–450.
[49] S. Langarica, M. Rodriguez-Fernandez, F. J. Doyle, F. Núñez, *IEEE J. Biomed. Health Inf.* **2023**, *27*, 10.
[50] Y. Huang, Z. Ni, Z. Lu, X. He, J. Hu, B. Li, H. Ya, Y. Shi, *Front. Phys.* **2023**, *14*, 1225638.

[51] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, in *Proc. of the AAAI Conf. on Artificial Intelligence* **2017**, Vol. *31*.
[52] T. Zhu, T. Chen, L. Kuangt, J. Zeng, K. Li, P. Georgiou, in *2023 IEEE Inter. Symp. on Circuits and Systems (ISCAS)*, IEEE **2023**, pp. 1–5.
[53] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, P. Georgiou, *IEEE Internet Things J.* **2022**, *10*, 3706.
[54] R. Krishnamoorthi. arXiv preprint arXiv:1806.08342, **2018**.
[55] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, in *Proc. of the IEEE conference on computer vision and pattern recognition* **2018**, pp. 2704–2713.
[56] W. Huang, H. Qin, Y. Liu, J. Liang, Y. Ding, Y. Li, and X. Liu. arXiv preprint arXiv:2309.01945, **2023**.
[57] K. Guo, S. Zeng, J. Yu, Y. Wang, H. Yang, arXiv preprint arXiv:1712.08934, **2017**.