Multi-Horizon Glucose Prediction Across Populations with Deep Domain Generalization

Taiyu Zhu, *Member, IEEE*, Ioannis Afentakis, Kezhi Li, *Member, IEEE*, Ryan Armiger, Neil Hill, Nick Oliver, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*— Real-time continuous glucose monitoring (CGM), augmented with accurate glucose prediction, offers an effective strategy for maintaining blood glucose levels within a therapeutically appropriate range. This is particularly crucial for individuals with type 1 diabetes (T1D) who require long-term self-management. However, with extensive glycemic variability, developing a prediction algorithm applicable across diverse populations remains a significant challenge. Leveraging meta-learning for domain generalization, we propose GPFormer, a Transformer-based zero-shot learning method designed for multi-horizon glucose prediction. We developed GPFormer on the REPLACE-BG dataset, comprising 226 participants with T1D, and proceeded to evaluate its performance using three external clinical datasets with CGM data. These included the OhioT1DM dataset, a publicly available dataset including 12 T1D participants, as well as two proprietary datasets. The first proprietary dataset included 22 participants, while the second contained 45 participants, encompassing a diverse group with T1D, type 2 diabetes, and those without diabetes, including patients admitted to hospitals. These four datasets include both outpatient and inpatient settings, various intervention strategies, and demographic variability, which effectively reflect real-world scenarios of CGM usage. When compared with a group of machine learning baseline methods, GPFormer consistently demonstrated superior performance and achieved the lowest root mean square error for all the evaluated datasets up to a prediction horizon of two hours. These experimental results highlight the effectiveness and generalizability of the proposed model across a variety of populations, demonstrating its substantial potential to enhance glucose management in a wide range of practical clinical settings.**

***Index Terms*— Deep learning, diabetes, domain generalization, glucose prediction, Transformer.**

This work was supported by EPSRC EP/P00993X/1, President's Ph.D. Scholarship, and UKRI Centre for Doctoral Training in AI for Healthcare (EP/S023283/1) at Imperial College London. *(Corresponding author: K. Li)*

T. Zhu was with Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom. He is now with Department of Psychiatry, University of Oxford, Oxford, United Kingdom (e-mail: taiyu.zhu@psych.ox.ac.uk).

I. Afentakis, R. Armiger, and P. Georgiou are with Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom. R. Armiger is also with National Institute for Health Research, Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Imperial College London, London, United Kingdom. (e-mail: {i.afentakis20, ryan.armiger15, pantelis}@imperial.ac.uk).

K. Li is with Institute of Health Informatics, University College London, London, United Kingdom. (e-mail: ken.li@ucl.ac.uk).

N. Hill and N. Oliver are with Department of Metabolism, Digestion and Reproduction, Faculty of Medicine, Imperial College London, United Kingdom. (e-mail: neilhill76@gmail.com, nick.oliver@imperial.ac.uk).

I. Introduction

TYPE 1 diabetes (T1D) is a long-term condition characterized by elevated blood glucose levels, which affects millions of people worldwide [1]. This condition results from an absolute deficiency of insulin and requires people living with T1D to adopt lifelong management [2]. Effective T1D management primarily involves the administration of exogenous insulin and glucose monitoring, with the core objective of maintaining blood glucose levels within a specified target range. By mitigating episodes of hypoglycemia and hyperglycemia, these management approaches aid in the prevention of a spectrum of diabetes-associated complications [3].

In light of this, the rapid evolution of real-time continuous glucose monitoring (CGM) over the past decades has demonstrated its potential to support optimal self-management and increase the percentage of time that blood glucose within the target range [4], [5]. Consequently, CGM has been endorsed as a standard element of T1D management [6], [7]. Furthermore, the use of CGM extends beyond outpatient T1D care; it has increasingly been adopted across a diverse range of clinical scenarios, including its use in hospital settings [8], [9], especially during the COVID-19 pandemic [10], [11], and among individuals without diabetes [12], such as aiding in dietary strategies aimed at ameliorating metabolic complications [13]. CGM is also increasingly adopted for type 2 diabetes (T2D) and has shown effectiveness in improving HbA1c [14]. Despite these advances, a notable time delay persists, ranging from five to ten minutes, between changes in blood glucose concentrations and detection by interstitial CGM sensors [15]. Moreover, the onset of action for insulin varies, typically starting within 15 minutes and extending up to several hours, depending on the formulation [16]. Therefore, there is a significant need for robust glucose prediction models to facilitate proactive self-management, effectively reducing the risk of potential adverse glycemic events. These models should also possess the versatility to generalize across populations and different clinical scenarios.

Building on the development of CGM, which has yielded a substantial volume of time series glucose data, machine learning methodologies have been employed across a variety of tasks related to diabetes management [17]–[19]. Empowered by various architectures of deep neural networks (DNNs), deep learning has achieved the state of the art in glucose prediction [20]. In the literature, most of the existing work relies on traditional DNN models, such as recurrent

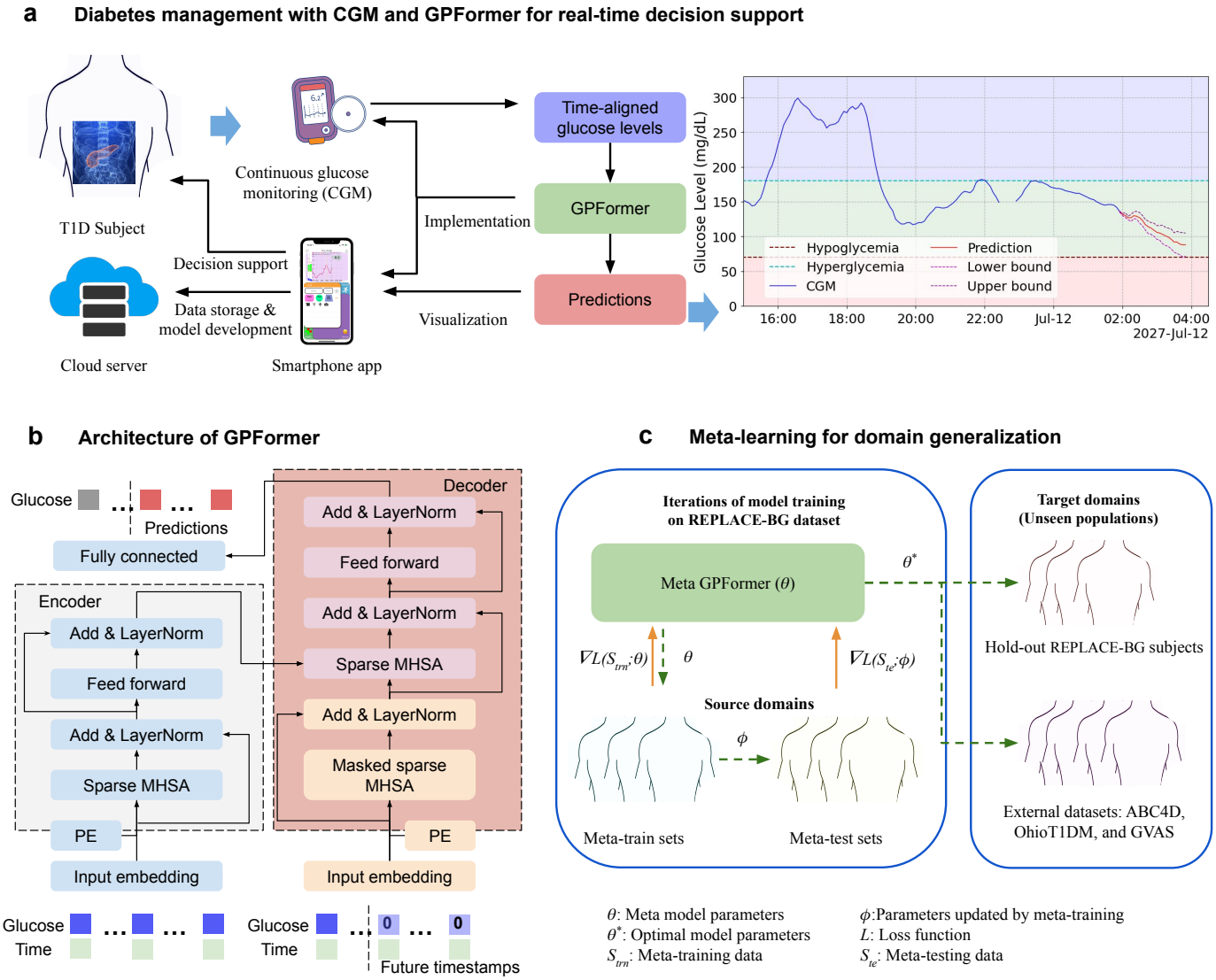


Fig. 1: Overview of developing and applying GPFormer for real-time decision support. **a** Demonstration of diabetes management systems for a T1D subject. CGM measures glucose levels and provides input data to GPFormer, which is developed in cloud server and can be implemented on smartphone apps or CGM transmitters. **b** Architecture of GPFormer based on Transformer mechanisms, including positional encoding (PE) multi-head self-attention (MHSA), and residual connections. **c** Diagram of meta-learning for domain generalization framework that develops prediction models generalized across different populations, where each individual is treated as a unique learning domain.

neural networks [21] that incorporate long short-term memory (LSTM) [22] or gated recurrent units [23]–[26], as well as convolutional neural networks [26], [27]. Recent advances in Transformer-based models [28] are reshaping the landscape of deep learning, demonstrating remarkable improvements in the field of natural language processing [29], which have become the backbone of many large language models, such as GPT-3 [30]. Built upon the self-attention mechanism [28], Transformers are highly effective at handling sequence data, and it can perform parallel computation to capture long-term dependencies. Hence, the inherent capabilities of Transformers make it attractive for time series forecasting, leading to the development of Transformer-based variants specifically aimed at this task in recent research [31].

The predominant trend in current research is to design personalized models tailored for individual subjects in one or two datasets [21], [22], [25]–[27], due to the limited availability of expansive CGM datasets. Given the considerable inter- and intra-subject glycemic variability [32]–[34], developing a glucose prediction model capable of achieving robust generalization across diverse populations is a challenging task. That is, a population-level model, once trained on an extensive dataset, has the capability to generate accurate predictions for unseen individuals from new datasets that may encompass varied clinical settings, distinct from the original training dataset. This aligns with the concept of domain generalization in the field of deep learning, which refers to the ability of a

model to perform effectively on data from domains or settings that it has not encountered during its training phase [35]. Recognizing this, harnessing vast historical CGM datasets and advanced domain generalization techniques to learn the complex glycemic and develop models at a population level emerges as a pivotal strategy in diabetes management. Such models are particularly advantageous for inpatient settings and new CGM users who lack individual data. However, there is a noticeable gap in research focused on glucose prediction for hospitalized patients, despite the fact that approximately 27.3% of them are living with diabetes [36]. Proactively preventing extreme glucose fluctuations in this population is critically important, not only for reducing mortality and morbidity but also for minimizing healthcare costs and optimizing resource utilization.

In addition, a significant proportion of these studies adopt a single-horizon predictive setting [20]–[22], [25], [26], also known as sequence-to-one regression. In this approach, a model outputs a single data point for a specific prediction horizon, necessitating multiple distinct models for different horizons. Deploying multiple predictive models on wearable devices for long-term prediction introduces practical complications, as these devices are often constrained by computational and memory capabilities [24]. In this context, multi-horizon prediction models, known as sequence-to-sequence regression, are more effective for real-world applications by generating a glucose trajectory that encompasses all considered horizons. They provide a comprehensive view of future outcomes and enhance decision-making processes [37], particularly when anticipating a range of future scenarios such as hypoglycemia, euglycemia, and hyperglycemia. Multi-horizon prediction is inherently more challenging because it requires optimizing multiple targets. However, employing a Transformer encoder-decoder architecture, as illustrated in Fig. 1, enables an effective sequence-to-sequence model. This architecture is particularly suited for handling the complexities of multi-horizon prediction by capturing long-term dependencies and relationships within the data.

In this study, to tackle the aforementioned challenges, we propose GPFormer (Glucose Prediction transFormer) for long-term multi-horizon glucose prediction, which is generalized across populations. In the context of glucose prediction, pioneering efforts have been made to apply Transformers, including Transformer [38], Glucose Transformer [39], Gluformer [40], and our prior work on Temporal Fusion Transformer [41]. However, all these existing studies leverage the vanilla multi-head self-attention (MHSA) mechanism, using either part of or the entire original Transformer architecture [28]. In contrast to these existing works, we apply a customized sparse MHSA mechanism for distilling major attention weights, incorporates a meta-learning framework to improve the model's ability for domain generalization, and utilizes a quantile loss function to better predict adverse glycemic events. Our proposed approach has demonstrated further improved prediction performance compared to vanilla MHSA (i.e., without sparsity), as demonstrated in Fig. 7 in the Appendix. The overall architecture is illustrated in Fig. 1. The population model was trained on the REPLACE-BG dataset [42], a publicly available T1D dataset comprised of 226 adult participants. The model was then evaluated using three external clinical datasets. Among these, OhioT1DM dataset is also publicly available [43], while ABC4D and GVAS are proprietary datasets (Imperial College London, London, UK). GVAS is an inpatient dataset, which includes CGM data from both people with and without diabetes. The proposed model consistently outperformed the considered baseline methods across all evaluation datasets.

II. Methodology

A. Data Preprocessing

The data partition for developing a population model was performed on individual basis (Appendix Fig. 8). We first divided the REPLACE-BG dataset into a training set with 180 T1D subjects and a hold-out testing set consisting of 46 T1D subjects. Stratified random sampling based on age and gender was used to ensure that the demographic profile of the testing set closely matched that of the entire REPLACE-BG dataset. The OhioT1DM, ABC4D, and GVAS datasets were utilized as hold-out testing sets. We did not specifically select outliers (i.e., significantly shifted domains) for the REPLACE-BG testing set, considering that we already had three external validation cohorts with considerable differences (Table I). For each individual in the REPLACE-BG, OhioT1DM, and ABC4D datasets, we retained only the initial three days of data for three primary reasons. Firstly, we aimed to simulate a data-limited scenario in hospital settings, where health records for a large number of patients are available, but each patient has limited days wearing a CGM device. Secondly, it is important to note that these datasets varied considerably in length, particularly the inpatient GVAS dataset, which has an average length of three days. By retaining three days of data, we ensure better consistency in dataset lengths and provide a more equitable evaluation. Finally, combining all the six-month training data for the 135 REPLACE-BG subjects within the framework of domain generalization would result in extensive memory usage, potentially causing an out-of-memory issues.

Subsequently, we scaled the glucose levels by standardization, while the 24-hour timestamps were encoded via sine and cosine transformations to extract cyclical temporal patterns, thus forming a three-dimensional vector. In the context of multi-horizon glucose prediction, a sequence of historical data with a fixed window size L is used to predict the corresponding sequence of future glucose levels with a prediction horizon of τ. Therefore, we derived input and output sequences using a sliding window approach. We applied linear interpolation to fill the gaps that appeared in the middle of glucose sequences, and extrapolated gaps at the tail end to avoid leakage of future information. At a timestep of t, the input of encoder is denoted as $\mathbf{X}_t^{enc} = [\mathbf{x}_{t-L+1}, \mathbf{x}_{t-L+2}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{3\times L}$, which combines the glucose and timestamp data. For the decoder, the similar input is used, denoted as $\mathbf{X}_t^{dec} = [\mathbf{x}_{t-L'+1}, \mathbf{x}_{t-L'+2}, \ldots, \mathbf{x}_\tau] \in \mathbb{R}^{3\times(L'+\tau)}$, where future glucose levels are represented via zeros. Following standardization, these zero values are effectively the mean value of

glucose data. The output of the decoder is represented as $\hat{\mathbf{y}}_t = [\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots, \hat{y}_{t+\tau}] \in \mathbb{R}^{\tau}$.

B. Transformer with Sparse Self-Attention Mechanism

In this work, we introduce GPFormer, a Transformer designed to process CGM data and predict glucose levels on a population scale, as depicted in Fig. 1. Inspired by the input representation approaches used in recent Transformer-based forecasting models [31], glucose sequences and cyclical time sequences were first processed by two distinct embedding layers that converted the feature dimension to d_m through a 1-D convolution layer and a linear layer, respectively. Then positional encoding vectors of the same dimension were generated using sine and cosine functions at varying frequencies to indicate the relative position of the elements ($\mathbf{x}_i$) within the sequence. By adding these components together, we obtain the input of self-attention mechanism: query $\mathbf{Q} \in \mathbb{R}^{L \times d_m}$, key $\mathbf{K} \in \mathbb{R}^{L \times d_m}$, and value $\mathbf{V} \in \mathbb{R}^{L \times d_m}$, where $\mathbf{q}_i$, $\mathbf{k}_i$, $\mathbf{v}_i$ stand for the elements on the i-th row. The output of self-attention module scales the values $\mathbf{V}$ by self-attention scores $\mathcal{A}$, which is denoted as

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathcal{A}(\mathbf{Q}, \mathbf{K})\mathbf{V} = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_m}})\mathbf{V}. \tag{1}$$

Previous research has shown that the self-attention matrix $\mathcal{A}$ tends to be sparse [31], resulting in a long tail distribution for self-attention scores [44], i.e., the elements of the self-attention matrix. Hence, with the aim of enhancing the generalization and robustness of the population model, we selected the attention scores that made significant contribution towards output, which preserved common temporal patterns of glucose trajectories, while discarding the trivial ones. To achieve this, we selected rows of the query associated dominant attention scores, i.e., those with the smallest sparsity. Denoting the function of a dot-product pair for the i-th query and j-th key as $f_k(\mathbf{q}_i, \mathbf{k}_j) = \mathbf{q}_i \mathbf{k}_j^T / \sqrt{d_m}$, the corresponding $Softmax$ score can be represented in probabilistic form as $p_A = \exp(f_k(\mathbf{q}_i, \mathbf{k}_j) / \sum_l \exp(f_k(\mathbf{q}_i, \mathbf{k}_l)$ [45]. Compare it with a uniform distribution of $p_U = 1/L$, the sparsity S can be measured by Kullback-Leibler divergence D_{KL}, as follows.

$$\begin{aligned} S(\mathbf{q}_i, \mathbf{K}) &= D_{KL}(p_U || p_A) = \sum_{j=1}^{L} \frac{1}{L} \log(\frac{\sum_l e^{f_k(\mathbf{q}_i, \mathbf{k}_l)}}{L e^{f_k(\mathbf{q}_i, \mathbf{k}_j)}}) \\ &= \log(\sum_{j=1}^{L} e^{f_k(\mathbf{q}_i, \mathbf{k}_l)}) - \frac{1}{L}\sum_{j=1}^{L} f_k(\mathbf{q}_i, \mathbf{k}_j) - \log(L). \end{aligned} \tag{2}$$

Following this, we assessed the sparsity of each row within the query using max-mean approximation [44], and selected the top n rows, denoted by indices I_n, based on this ranking. The remaining rows were filled with zeros to yield a sparse query $\bar{\mathbf{Q}}$. The process is as follows.

$$\bar{\mathbf{Q}} = [\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \ldots, \bar{\mathbf{q}}_L], \quad \text{where} \quad \bar{\mathbf{q}}_i = \begin{cases} \mathbf{q}_i & \text{if } i \in I_n \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

In order to enhance the model's learning capability, we incorporated a MHSA module [28] that allows the model to concurrently learn different hidden representations and focus on varied positions of the input sequence. Integrating it with the sparse query, an h-head sparse MHSA is denoted as follows.

$$\begin{aligned} MHSA(\bar{\mathbf{Q}}, \mathbf{K}, \mathbf{V}) = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_h]\mathbf{W}^H, \quad \text{where} \\ \mathbf{H}_i = Attention(\bar{\mathbf{Q}}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \end{aligned} \tag{4}$$

where $\mathbf{H}_i \in \mathbb{R}^{L \times d_a}$ stands for output of the i-th attention head, and $d_a = d_m/h$ represents the corresponding output dimension; $\mathbf{W}_i^Q \in \mathbb{R}^{d_m \times d_a}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_m \times d_a}$, and $\mathbf{W}_i^V \in \mathbb{R}^{d_m \times d_a}$ are head-specific projection parameters. The output weights $\mathbf{W}^H \in \mathbb{R}^{hd_a \times d_m}$ linearly processes the concatenated output of parallel attention layers.

The sparse MHSA modules were incorporated into both encoder and decoder, which allow them to attend to information from all positions of encoder input, encoder output, and decoder input. It is important to note that we also implemented a autoregressive mask in the decoder (Fig. 1) to ensure the each prediction does not attend the outputs at positions after it, which is structured as an upper-triangular matrix filled with $-\infty$. The remaining modules, including position-wise feed-forward networks, layer normalization, and residual connections are consistent with those in canonical Transformer [28].

C. Model Training with Domain Generalization

The primary objective of this work is to develop a robust GPFormer model based on data of historical T1D subjects, which can provide reliable glucose prediction for unseen subjects with CGM data, even in the face of significant inter-subject variability. By conceptualizing each T1D subject as a distinct learning domain, we transformed population glucose prediction as a domain generalization problem, aiming to mitigate the domain shift between the historical T1D cohort and a new cohort. However, the majority of current domain generalization algorithms are dependent on specific applications [35], such as tasks in computer vision. Fortunately, meta-learning, also known as learning-to-learn, has been increasingly adopted in domain generalization, which has also been demonstrated to enhance few-shot learning in personalized glucose prediction in our previous work [23]. Therefore, in this work, we applied meta-learning for domain generalization (MLDG) [46], a model agnostic meta-learning framework, to tackle this more challenging scenario of zero-shot learning.

As shown in Fig. 1, the central concept of MLDG is to expose the model to domain shift during training by randomly splitting the historical cohort into meta-train sets and meta-test sets. This process essentially mirrors the real-world situation where the model encounters unseen data. The model parameters are updated through a bi-level optimization process. The details of training and testing GPFormer within MLDG are presented in Algorithm 1.

D. Regression with Quantile Loss

Reflective insights gleaned from our prior clinical trials underline the instrumental value of featuring multi-horizon

Algorithm 1 Training GPFormer within meta-learning for domain generalization

Input: T1D subjects in the training set of REPLACE-BG as source domains $\mathcal{S}$, randomly initialized parameters of global GPFormer θ, loss function $\mathcal{L}$, number of iterations T, loss factor β, learning rates α and γ

Output: Optimized model θ^* for glucose prediction

for steps $t \in 1, 2, \ldots, T$ **do**
  Partition $\mathcal{S}$ into meta-train domains $\mathcal{S}_{trn}$ and meta-test domains $\mathcal{S}_{te}$ on a individual basis
  Sample a mini-batch $\mathcal{B}_{trn}$ from T1D subjects in $\mathcal{S}_{trn}$
  Obtain loss in meta-train $\mathcal{L}(\mathcal{B}_{trn}; \theta)$
  Copy the global model to a local model and update its parameters $\phi = \theta - \alpha \nabla \mathcal{L}(\mathcal{B}_{trn}; \theta)$
  Sample a mini-batch $\mathcal{B}_{te}$ from T1D subjects in $\mathcal{S}_{te}$
  Obtain loss with the local model in meta-test $\mathcal{L}(\mathcal{B}_{te}; \phi)$
  Update global GPFormer $\theta \leftarrow \theta - \gamma \Delta(\mathcal{L}(\mathcal{B}_{trn}; \theta) + \beta \mathcal{L}(\mathcal{B}_{te}; \phi))$
end for

predictions supplemented with upper and lower confidence bounds, marking a commendable enhancement in diabetes care functionalities [25]. Therefore, to obtain these bounds and provide a comprehensive understanding of the prediction uncertainty, we applied a quantile loss function [47] that not only predicted glucose levels but also provided the associated lower and upper bounds to estimate intervals reflecting the potential range of future values. It is denoted as follows.

$$\mathcal{L} = \frac{1}{\tau} \sum_{q_l \in \mathcal{Q}} \sum_{i=t}^{t+\tau} f_L(y^i, \hat{y}^i, q_l), \tag{5}$$

$$\text{where} \quad f_L(y^i, \hat{y}^i, q_l) = \begin{cases} (1 - q_l)|y^i - \hat{y}^i|, & \text{if } y^i < \hat{y}^i, \\ q_l |y^i - \hat{y}^i|, & \text{if } y^i \geq \hat{y}^i. \end{cases}$$

where $\mathcal{Q} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ represents the set of quantiles.

III. Experiments

A. Clinical Datasets

In this research, we employed an array of clinical datasets to validate GPFormer, which include two publicly available datasets, namely REPLACE-BG [42] and OhioT1DM [43], alongside two proprietary ones, ABC4D and GVAS. These datasets were collected in clinical trials conducted across various settings with a diverse range of participants, thereby ensuring a comprehensive and robust evaluation.

In particular, the REPLACE-BG dataset contains data of 226 adult T1D participants, which was collected over 26 weeks of clinical trial. Each participant was equipped with the Dexcom CGM system and a personal insulin pump. The objective of this study is to investigate the safety and effectiveness of the use of CGM, particularly without the need for confirmatory blood glucose monitoring. The OhioT1DM dataset encompasses data of 12 T1D participants who wore Medtronic Enlite CGM and Medtronic insulin pumps, which was collected in an eight-week clinical trial. The aim of this study is to facilitate the research on data-driven glucose prediction in diabetes management. Therefore, for each participant, the dataset offers two distinct files for training and testing purposes, and we solely employed the testing data in this work.

We collected the ABC4D dataset during a controlled crossover study [48], which provided 33 participants with a smartphone app that employed case-based reasoning for personalizing insulin bolus recommendations. We recruited adult participants who had been living with T1D for a minimum of three years, as confirmed by clinical features and a c-peptide level below 200 pmol/L. Eligible participants had been following an intensified multiple daily injection regimen for at least six months, with an HbA1c level ranging from 53 mmol/mol (7.0%) to 75 mmol/mol (9.0%), and had completed structured diabetes education. After excluding incomplete records, we proceeded with data from 22 subjects. The study was under the protocol (13/LO/0264) approved by London - Chelsea Research Ethics Committee in 2013.

The GVAS dataset was collected from our previous observational study aiming to assess glucose and its associations during an initial 72 hours after a confirmed stroke in two groups: people with and without diabetes [49]. The study includes 67 adult participants with clinically suspected ischemic stroke within 12 hours of symptoms. The inclusion criteria were adults aged over 18 years admitted with arterial ischemic stroke within first 72 hours. Exclusion criteria were patients with hemorrhagic stroke, pregnancy, terminal illness or life expectancy less than a year. The study aims to provide insights into glucose dynamics during the early phase of stroke and potential differences between people with an acute stroke with and without diabetes. We excluded the data of participants who had less than 20 hours of data, and also removed any missing or duplicated records, leaving 45 subjects for analysis. The study was under the protocol (20/SC/0214) approved by South Central - Berkshire Research Ethics Committee in 2020.

B. Performance Evaluation

To assess the distribution of glucose data, we selected 24-hour CGM measurements with less than 10% missing data. For the principal component analysis (PCA) analysis [50], we chose three principal components to enable effective visualization while maintaining a high cumulative explained variance ratio of 58.7%, ensuring that a substantial amount of the original variance was captured. Considering the complexity of glucose dynamics and significant glycemic variability, we opted for a 3-dimensional visualization for t-distributed stochastic neighbor embedding (t-SNE) [51]. This decision enables a more detailed representation of the high-dimensional data. We set the perplexity parameter to 40, a choice that allows a balance between emphasizing local and global structures in the data.

During model validation phase of GPFormer, five-fold cross validation (Appendix Fig. 8) and Hyperband [52] optimization were introduced to fine-tune hyperparameters, where the training sets and the validation sets were also partitioned on an individual basis. Following the model training and validation, we evaluated the GPFormer model on the hold-out

TABLE I: Demographic characteristics (Mean ± SD) of the subjects in four clinical datasets

Demographic	REPLACE-BG	OhioT1DM	ABC4D	GVAS
Age (years)	44 ± 14	50 ± 10	47 ± 17	74 ± 12
Gender (male/female)	114/112	7/5	10/12	27/18
Insulin regimen (CSII/MDI)	226/0	12/0	0/22	N/A
Mean glucose level (mg/dL)	159.9 ± 26.4	162.3 ± 20.0	177.5 ± 17.6	140.3 ± 48.2
GMI (%)	7.1 ± 0.6	7.2 ± 0.5	7.6 ± 0.4	6.7 ± 1.2
TBR (<54 mg/dL) (%)	1.5 ± 2.3	0.4 ± 0.4	0.3 ± 0.3	1.2 ± 3.8
TBR (54-69 mg/dL) (%)	3.5 ± 3.3	2.3 ± 2.0	1.6 ± 1.2	2.5 ± 10.5
TIR (70-180 mg/dL) (%)	62.5 ± 16.0	62.9 ± 13.2	53.8 ± 10.8	80.6 ± 28.5
TAR (181-250 mg/dL) (%)	22.5 ± 10.1	25.7 ± 8.0	29.8 ± 5.7	10.2 ± 15.0
TAR (>250 mg/dL) (%)	10.1 ± 9.9	8.8 ± 8.5	14.4 ± 7.3	5.5 ± 18.2
Low blood glucose index	1.3 ± 1.1	0.7 ± 0.4	0.6 ± 0.3	1.0 ± 2.3
High blood glucose index	7.6 ± 4.4	7.4 ± 3.3	10.0 ± 3.0	4.4 ± 8.0
Inter-day CV (%)	37.1 ± 8.6	34.3 ± 4.6	35.7 ± 4.0	18.8 ± 6.9
Intra-day CV (%)	34.0 ± 7.8	29.2 ± 4.5	31.2 ± 3.4	16.0 ± 6.9

MDI: multiple daily injection, CSII: continuous subcutaneous insulin infusion

testing set from REPLACE-BG, as well as on three additional external CGM datasets. We mainly used root mean square error (RMSE) to assess the accuracy of the prediction, which is calculated as $RMSE_i = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t^i - y_t^i)^2}$ mg/dL, where N is the total number of trajectories and, and i represents the ith timestep in the prediction horizon, such as $i = t + 6$ for a 30-minute prediction horizon. Clark error grid (CEG) analysis analysis [53] was performed by scatter plot.

To ensure a fair comparison, we implemented deep learning baseline methods using MLDG, specifically N-BEATS [54] and Bi-LSTM [55], while preserving their original architectures. The support vector regression (SVR) model, a classic machine learning approach, was constructed based on the radial basis function kernel [56]. The orders of the ARIMA model, p, m and d, of the autoregressive (AR), moving average (MA) and integrated (I) parts respectively, were determined via cross validation over the sets $P = \{1, 2, ..., 6\}$, $M = \{0, 1, ..., 6\}$ and $D = \{0, 1, ..., 6\}$.

The training, validation, and testing procedures for all baseline methods were conducted on identical datasets as those used for GPFormer. To determine if the improvements of GPFormer are significant, we assessed statistical significance. First, we confirmed whether the differences were normally distributed using the D'Agostino-Pearson test. Depending on the outcome, we performed either a paired t-test or a Wilcoxon signed-rank test. Both tests used the null hypothesis that the mean difference in improvements on metrics between the paired observations is zero. A p-value less than 0.05 indicates significant improvements, meaning the differences are statistically different from zero. All statistical tests were performed using SciPy version 1.8.0.

C. Participant Characteristics of Clinical Datasets

Table I provides a comprehensive overview of the demographic characteristics associated with the four clinical datasets employed in this work. In particular, we present a collection of the standardized CGM metrics that have been recognized by international consensus for their applications in clinical diabetes care [7], which include glucose management indicator (GMI) [57], time below range (TBR), time in range (TIR), time above range (TAR), blood glucose index, and coefficient of variation (CV). The recently proposed GMI

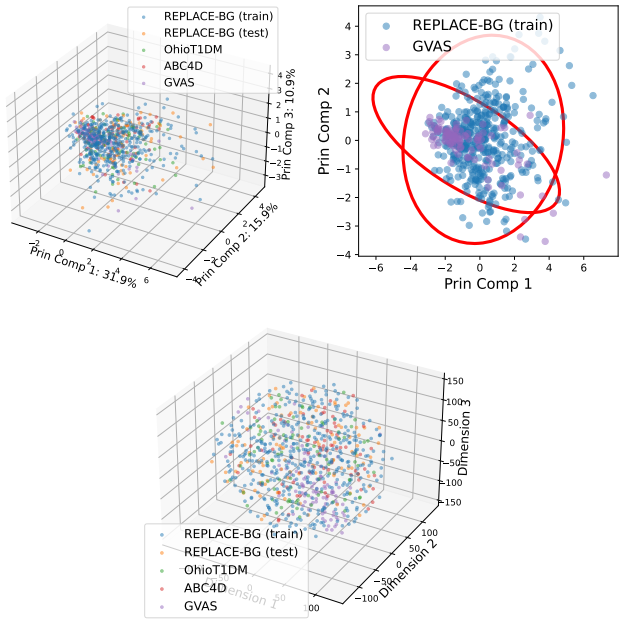


Fig. 2: Visualization of daily glucose profiles of the four clinical datasets. The top row displays PCA results, while the bottom row shows t-SNE plots.

serves as an alternative to the traditional laboratory-measured HbA1c when CGM data is available.

D. Visualization of CGM Glucose Profiles

Fig. 2 illustrates the three-dimensional distributions of CGM glucose profiles [58] derived by dimensionality reduction techniques, namely PCA and t-SNE. Each dot in the plots stands for a 24-hour trajectory of glucose levels. PCA visualizes high-dimensional data using a linear method, while t-SNE employs a non-linear technique. In both plots, we observe overlaps in the data distributions from the four clinical datasets, indicating the presence of shared patterns that the deep learning model can utilize. However, we also detect notable domain shifts, particularly between the REPLACE-BG and GVAS datasets, as illustrated by the red confidence ellipses. This visualization serves as a preliminary evaluation, suggesting the potential in developing population-based glucose predictors.

E. Performance of Glucose Prediction

The proposed GPFormer model was benchmarked against four established machine learning algorithms previously used in glucose prediction studies. It is worth noting that GPFormer consistently outperformed the other models by achieving the smallest RMSE on the hold-out testing subjects in the REPLACE-BG, and on three external datasets, as depicted in Fig. 3 and 4, respectively. Furthermore, fewer outliers with large RMSE values, represented by data points above the whiskers in the box plots, are observed in the results obtained from GPFormer, indicating its robustness when handling subjects that deviate from the characteristics of the training population. Notably, over a 30-minute prediction horizon, GPFormer achieved RMSE (mean±SD) of 19.2 ± 3.8, 22.9 ± 3.9, 18.9 ± 3.2, and 15.9 ± 4.0 for the REPLACE-BG, ABC4D, OhioT1DM, and GVAS datasets, respectively.

Table II presents the regression metrics used to evaluate glucose prediction across different prediction horizons. Specifically, we utilized RMSE, prediction time delay (PTD), glucose variability impact index (GVII), and glucose prediction consistency index (GPCI) [59]. PTD measures the lag in prediction (in minutes), while GVII and GPCI incorporate glucose variability to assess prediction accuracy. For all these metrics, lower values indicate better performance. It is worth noting that GPFormer demonstrated consistent performance across all the evaluation scenarios, achieving the lowest metrics in 59 out of 64 instances across the four metrics, four different datasets, and four prediction horizons. Fig. 5 displays both the predicted glucose levels and actual CGM readings over a 60-minute prediction horizon. The integration of MLDG, coupled with the establishment of lower and upper bounds, significantly enhanced GPFormer's ability in forecasting hypoglycemic and hyperglycemic events. Table III presents the results of hypoglycemia prediction, a challenging task involving the prediction of hypoglycemic events within the next two hours. This is a highly imbalanced classification problem, as indicated by the very small TBR in Table I. To provide a comprehensive evaluation, we include a group of metrics: accuracy (ACC %), sensitivity (SEN %), specificity (SPEC %), precision (PREC %), and F1 scores. The F1 score, in particular, is a balanced metric that reflects the trade-off between precision and recall. Notably, GPFormer achieved the most balanced performance with the highest F1 scores and precision across all four datasets, indicating the highest probability of correctly detecting hypoglycemia. For regression metrics, when compared with the best-performing baseline methods across the four prediction horizons, i.e., Bi-LSTM for GVAS and N-BEATS for the others, GPFormer showed significant improvements in RMSE with $p < 0.05$. In hypoglycemia prediction, GPFormer also demonstrated significant improvements in F1 scores with $p < 0.05$ when compared with N-BEATS for OhioT1DM and ARIMA for the others.

To further assess the performance of GPFormer across diverse populations, we conducted Clarke Error Grid (CEG) analysis on the four datasets, as shown in Fig. 6. For the GVAS dataset, we separated the results between participants with diabetes (T2D) and those without diabetes. Each dot on the plots represents a paired instance of actual and predicted glucose values. Notably, the majority of the dots cluster in zones A and B, indicating that the associated predictions would not lead to inappropriate treatment decisions. Specifically, 97.68%, 99.35%, and 99.60% of the predictions fall within CEG zones A and B for the REPLACE-BG testing set, ABC4D, and OhioT1DM, respectively. For the GVAS dataset, 99.95% of predictions for participants with diabetes and 99.52% for participants without diabetes are within CEG zones A and B. This underlines the strong generalizability of GPFormer across patient populations.

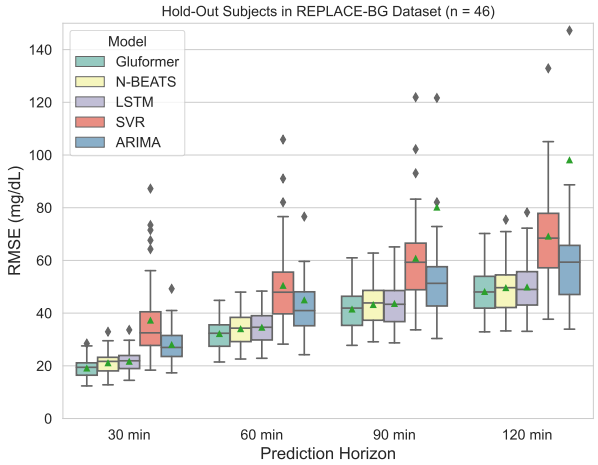


Fig. 3: Multi-horizon glucose prediction performance of GPFormer and the considered baseline method on the hold-out subjects in the REPLACE-BG dataset.

IV. Discussion

The novelty of this work lies in two primary aspects. Firstly, we designed a sparse MHSA mechanism inspired by Informer [44] and combined it with a quantile loss function and MLDG training framework to develop a glucose predictor particularly suitable for population-based prediction. This approach differs from existing research on Transformers and glucose prediction, which mainly utilized the vanilla MHSA and original Transformer blocks. Notably, in one of our previous studies, we explored the vanilla Transformer encoder as a baseline method [23]; however, it did not exhibit significant improvements when compared with Bi-LSTM models. Secondly, our multi-horizon model was evaluated on four diverse clinical datasets to demonstrate its generalization capability and potential for a wider range of real-world applications. This includes both individuals with and without diabetes, in inpatient and outpatient settings. However, existing studies considered relatively small sample sizes of individuals within one or two datasets and focused on single horizons.

In clinical settings, substantial glycemic variability is often encountered, which impacts not only patients with diabetes but also those without this condition [34]. This variability is particularly pronounced when these individuals are subject to different types of interventions, such as varying insulin delivery schemes [60], which can range from multiple daily

TABLE II: Regression metrics for evaluating the glucose prediction performance

	Method	30 minutes				60 minutes				90 minutes				120 minutes			
		RMSE	GVII	GPCI	PTD	RMSE	GVII	GPCI	PTD	RMSE	GVII	GPCI	PTD	RMSE	GVII	GPCI	PTD
REPLACE	GPFormer	**19.2**	**0.14**	**2.31**	**3.8**	**32.3**	**0.29**	**3.54**	**15.6**	**41.6**	**0.44**	**4.19**	**27.2**	**48.2**	**0.56**	**4.83**	**28.7**
	N-BEATS	21.2	0.17	2.56	6.0	34.1	0.35	3.67	19.3	43.3	0.52	4.31	30.0	49.7	0.66	4.91	30.4
	Bi-LSTM	21.7	0.18	2.53	6.6	34.7	0.39	3.62	20.4	43.7	0.56	4.39	30.1	50.0	0.71	5.10	32.2
	SVR	37.4	1.20	6.37	8.5	50.5	1.17	7.14	21.5	60.9	1.16	8.17	27.5	69.2	1.15	9.42	35.4
	ARIMA	28.1	0.18	4.03	12.4	45.1	0.32	14.39	33.9	80.2	0.76	10.83	57.5	98.2	1.06	16.47	79.8
ABC4D	GPFormer	**22.9**	**0.32**	**7.27**	**4.4**	**35.9**	**0.32**	**10.50**	**15.0**	**43.9**	0.29	**12.65**	**25.0**	**49.2**	**0.29**	**15.16**	26.9
	N-BEATS	24.9	0.34	7.57	6.8	37.6	0.33	10.77	17.7	45.8	0.32	12.95	28.5	50.9	0.31	15.26	**26.3**
	Bi-LSTM	25.7	0.34	7.96	6.3	38.1	**0.32**	10.93	17.7	45.9	**0.26**	13.28	27.3	50.8	0.27	15.59	29.2
	SVR	40.1	0.84	19.97	10.0	53.2	0.76	19.40	19.2	62.7	0.59	20.49	29.9	71.5	0.57	20.91	31.9
	ARIMA	33.9	0.64	10.81	10.7	72.3	0.99	19.64	31.9	96.8	2.54	28.91	51.2	103.0	2.06	28.26	69.3
OhioT1DM	GPFormer	**18.9**	**0.16**	7.58	**4.2**	**32.4**	**0.33**	**8.58**	**14.7**	**42.0**	**0.48**	10.33	**22.8**	**48.7**	**0.57**	13.07	**26.3**
	N-BEATS	21.1	0.26	7.70	4.6	34.4	0.44	8.66	16.9	44.1	0.58	**10.09**	25.9	50.4	0.66	**12.43**	32.7
	Bi-LSTM	21.9	0.25	**7.56**	4.7	35.2	0.45	8.43	15.0	44.7	0.61	10.82	22.8	50.8	0.72	13.34	31.8
	SVR	36.9	0.64	19.71	6.5	50.5	0.93	15.61	19.9	61.2	1.29	14.88	28.0	70.1	1.54	16.52	37.9
	ARIMA	27.9	0.17	8.09	8.6	42.2	0.50	12.44	26.7	57.5	0.78	23.47	44.4	60.8	0.74	21.27	61.5
GVAS	GPFormer	**15.9**	**0.45**	**2.28**	**7.7**	**22.3**	**0.89**	**2.94**	**14.9**	**26.7**	**1.29**	**3.76**	**19.6**	**29.2**	**1.57**	**4.67**	**23.9**
	N-BEATS	18.0	0.39	2.66	8.8	25.5	0.79	3.18	15.3	30.5	1.12	4.05	21.0	34.4	1.33	5.68	28.0
	Bi-LSTM	17.9	0.48	2.68	8.6	24.8	0.88	3.35	16.0	29.9	1.18	4.62	**19.6**	33.7	1.34	6.37	27.0
	SVR	25.0	1.69	15.22	9.1	33.7	2.00	15.20	17.2	39.9	2.19	15.63	23.1	40.6	2.67	16.80	29.1
	ARIMA	19.6	0.74	2.76	10.8	26.4	1.20	3.71	18.9	33.1	1.66	5.32	25.2	39.3	2.06	6.63	35.1

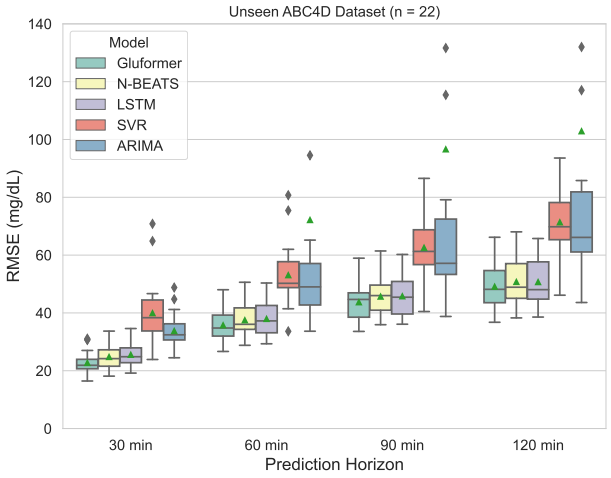
TABLE III: Performance of hypoglycemia prediction with a two-hour prediction horizon

	Method	ACC	SEN	SPEC	PREC	F1
REPLACE	GPFormer	**89.5**	74.9	**90.3**	**46.4**	**0.56**
	GPFormer*	89.1	85.9	88.7	32.3	0.46
	N-BEATS	89.3	87.3	88.8	26.8	0.39
	Bi-LSTM	88.3	**94.6**	87.7	23.6	0.35
	SVR	85.6	46.6	87.9	24.0	0.30
	ARIMA	89.3	84.2	89.1	29.4	0.41
ABC4D	GPFormer	**92.5**	56.6	**94.3**	**52.2**	**0.48**
	GPFormer*	**92.5**	80.8	92.6	39.1	**0.48**
	N-BEATS	91.5	85.3	91.4	29.1	0.39
	Bi-LSTM	90.1	**90.2**	90.0	21.3	0.32
	SVR	88.8	41.8	91.1	25.4	0.28
	ARIMA	91.7	81.7	91.8	33.0	0.43
OhioT1DM	GPFormer	**92.9**	64.7	**94.7**	**55.0**	**0.58**
	GPFormer*	92.2	91.0	92.0	31.6	0.45
	N-BEATS	92.4	86.5	92.2	29.3	0.43
	Bi-LSTM	89.2	**97.0**	88.9	26.6	0.40
	SVR	87.7	53.6	89.7	19.8	0.26
	ARIMA	91.9	79.1	92.1	28.4	0.40
GVAS	GPFormer	91.9	56.1	93.6	**57.2**	**0.51**
	GPFormer*	90.8	**90.1**	90.7	33.7	0.44
	N-BEATS	90.0	76.7	90.1	28.2	0.36
	Bi-LSTM	89.5	78.4	89.5	19.7	0.28
	SVR	**93.3**	30.5	**94.3**	23.3	0.18
	ARIMA	89.9	71.6	90.0	34.0	0.40

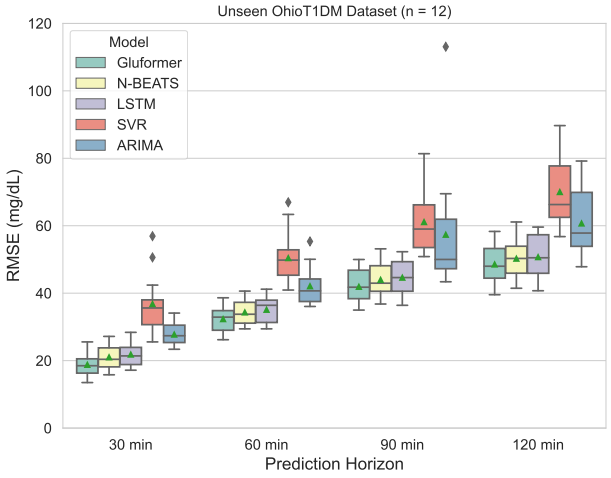
* indicates GPFormer without lower quantile bounds.

injections using insulin pens to continuous subcutaneous insulin infusion via insulin pumps. The extent of this variability is well-demonstrated in the clinical datasets utilized in this study. In addition, numerous internal factors also influence glycemic dynamics, such as diet, physical activities and individual metabolism. Hence, we observe notable discrepancies in demographics of the four clinical datasets, as outlined in Table I, which are further underscored by the significant variations in the four glucose trajectories depicted in Fig. 5. However, when we examine the daily CGM trajectories through PCA and t-SNE in Fig. 2, we observe both domain shifts and overlaps among the populations. Notably, overlaps are particularly notable in the t-SNE analysis, which is adept at capturing non-linear patterns. It suggests that, despite the variability, common linear and non-linear glycemic patterns exist in high dimensional spaces, which are shared across diverse populations. This observation motivated us to utilize domain generalization to develop population-level glucose prediction models. In Fig. 5, it is important to highlight that GPFormer demonstrated robust generalizability on the ABC4D and OhioT1DM datasets that consist of T1D subjects only. In these cases, the application of MLDG contributed to minor but noteworthy enhancements in performance. Nevertheless, for the GVAS dataset, which presents the most significant domain shift (Fig.2) and exhibits variable demographic characteristics (Table I), such as age and types of diabetes, the integration of MLDG leads to a significant enhancement in prediction accuracy, as marked in the areas circled in Fig. 5. To investigate demographic-based performance, we trained separate models for females and males using the REPLACE-BG dataset and validated them respectively on females and males in the testing sets. The 30-minute RMSE for the females' model and males' model were 19.7 and 20.1 mg/dL, respectively. Due to the different sizes of the result groups, we performed the Wilcoxon Rank-Sum test, which yielded a p-value much larger than 0.05, indicating no significant difference between the two result groups. This suggests that our model performs comparably across demographic variables, although the RMSE slightly increased, mainly due to the reduced sample size in the MLDG framework. We also calculated the point biserial correlation coefficient between gender and glucose variability, obtaining an r-value of 0.02 and a p-value of 0.68. This indicates a random association between gender and glucose variability, which further supports our findings.

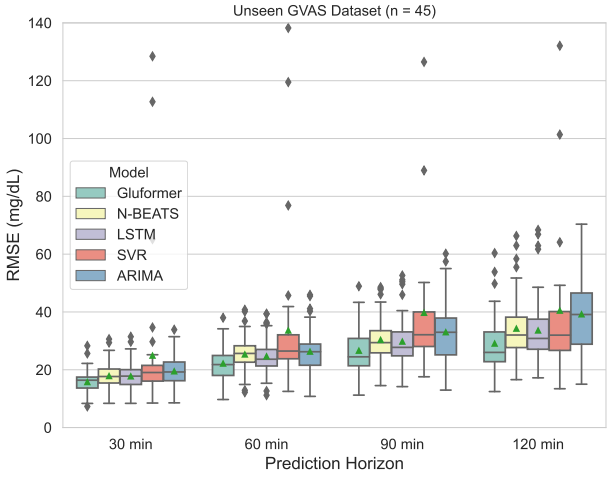
Unlike many existing studies on personalized glucose prediction [20], we did not use meal and insulin bolus data as model input. These data features are associated with noise and can further increase variability, due to various factors, such as inaccurate carbohydrate counting and missing manual records [61]. In our experiments on the REPLACE-BG dataset, including these data features resulted in a significant increase in mean RMSE and generated more outlier instances with large



(a) Performance on the ABC4D dataset



(b) Performance on the OhioT1DM dataset

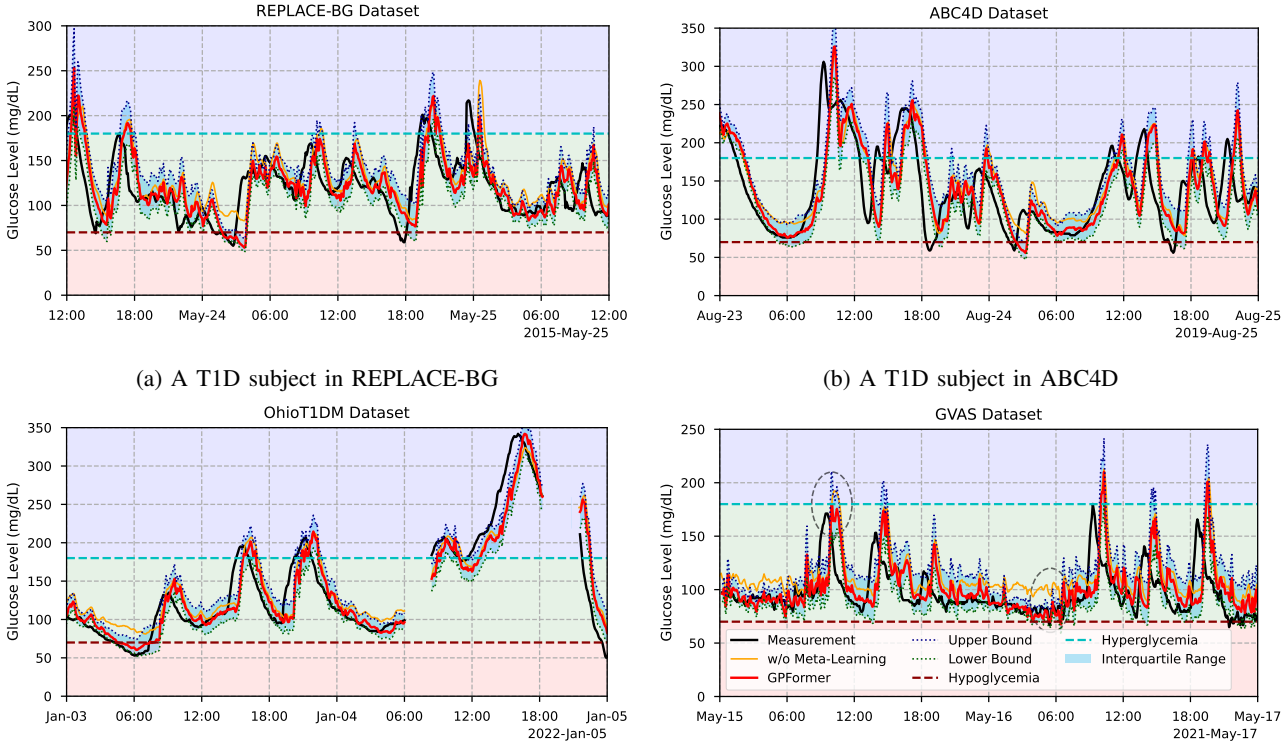


(c) Performance on the GVAS dataset

Fig. 4: Multi-horizon glucose prediction performance of GPFormer and the considered baseline methods on three completely external datasets. RMSE is utilized for evaluation.

RMSE. A similar degradation in performance was observed in the literature [59]. Using solely CGM data presents several key advantages. Firstly, it enables the possibility of embedding the model directly into CGM transmitters to provide on-device decision support through edge computing [24]. The model requires 9.8 MB of flash memory using the hardware pipeline proposed in our prior work [24] (Table IV in the Appendix). This makes it ideal for integration into microcontroller units of various edge devices, such as ESP32-WROVER-E, which features 16 MB of flash memory, Wi-Fi, and Bluetooth capabilities, or nRF52832, which includes Bluetooth and can be equipped with 128 MB of external flash memory. Secondly, it facilitates prediction without the need for manual input, thus avoiding potential artifacts and enabling automated closed-loop automated insulin delivery systems. Lastly, the proposed method is applicable to a broader array of clinical scenarios, including inpatient settings and monitoring for other conditions related to glucose metabolism, such as stroke, as demonstrated by non-diabetic and T2D patients with acute ischaemic stroke in the GVAS dataset.

In this work, a six-hour window size was primarily selected through hyperparameter tuning (Table V in the Appendix), where we evaluated various settings including two hours, four hours, six hours, and twelve hours. In cross-validation, our model achieved the lowest RMSE with a six-hour window, which we therefore adopted. This window length is also practical because the effects of insulin and meal intake typically last around six hours. Given that our model does not explicitly use insulin bolus and carbohydrate intake information, a six-hour window effectively captures the glucose level changes resulting from these events. Detailed ablation studies are provided in Fig. 7 in Appendix, where the inclusion of timestamps reduced RMSE and improved prediction. The Bi-LSTM model also demonstrated reduced RMSE when timestamps were added. However, due to the univariate nature of the N-BEATS model [54], timestamps were not included in this model. Moreover, model interpretability analysis (Fig. 9 in the Appendix) revealed that small cosine and sine timestamp values are associated with increased glucose predictions. This finding is consistent with the fact that cosine and sine values of -1 correspond to typical meal times when glucose levels are expected to rise due to food intake. In our analysis at the individual level, we observed notable improvements from timestamps for subjects with more consistent daily patterns with the population average and those with more fixed times for meal intake. However, many subjects exhibited considerable variability in their meal timing, and for these individuals, the inclusion of timestamps had less impact on prediction accuracy. This variability in meal patterns across the population led to minor overall improvements in Fig. 7. The ablation studies also indicate that the use of meta-learning significantly reduced RMSE and improved prediction accuracy across the four clinical datasets, resulting in better domain generalization. These findings align with the observations in Fig. 5, demonstrating that GPFormer trained with meta-learning shows enhanced capability in detecting adverse glycemic events. Although the inclusion of the quantile loss function resulted in a slight reduction in RMSE, using the lower bounds derived from this loss function significantly improved F1 scores for hypoglycemia prediction, as shown in Table III. These lower bounds can also be adjusted



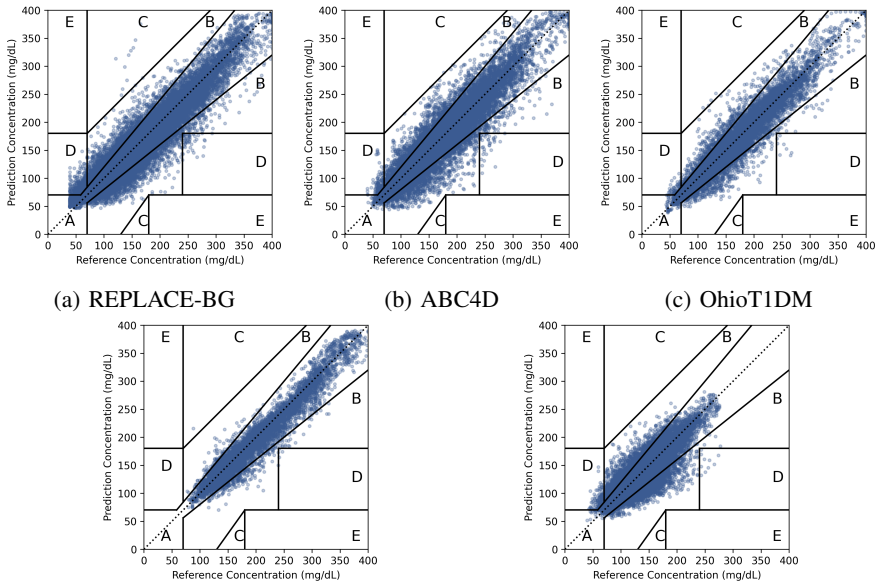
(a) A T1D subject in REPLACE-BG

(b) A T1D subject in ABC4D

(c) A T1D subject in OhioT1DM

(d) A T2D subject in GVAS

Fig. 5: Two-day trajectories of glucose predictions over the 60-minute prediction horizon, illustrated for individual subjects from each of the four clinical datasets. The solid black line represents actual CGM measurements, while the red line illustrates GPFormer predictions. The orange line depicts the results of GPFormer without the meta-learning framework. Blue dotted and green dotted lines represent the upper bounds (75th percentile) and lower bounds (25th percentile) derived via quantile regression. The background shading categorizes glucose levels into hypoglycemia, euglycemia, and hyperglycemia zones.



(a) REPLACE-BG

(b) ABC4D

(c) OhioT1DM

(d) Patients with diabetes in GVAS

(e) Patients without diabetes in GVAS

Fig. 6: Clark error grid (CEG) analysis that shows the comparison between CGM measurements (x-axis) and glucose prediction values (y-axis).

by clinicians and users by choosing different percentiles, allowing them to find an optimal balance between precision and sensitivity. This flexibility enables a customized approach to managing the trade-off between false alarms and missed

events, tailoring the model's predictions to individual patient needs or specific clinical contexts.

Upon evaluation across three external datasets, GPFormer consistently demonstrated superior performance with the smallest RMSE, as illustrated in Fig. 3 and 4. The deep learning baseline models, N-BEATS and Bi-LSTM, exhibited better performance when compared with traditional machine learning approaches including SVR and ARIMA, underlining the effectiveness of DNNs combined with meta-learning. Compared with deep learning baseline methods, GPFormer has 2.4 million trainable parameters and requires similar computational demands during both the training and inference stages (Table IV). The Bi-LSTM with a hidden size of 512 and two recurrent layers has 3.1 million trainable parameters, while N-BEATS has 4.9 million. GPFormer achieves a training speed of 0.23 iterations per second, compared to 0.43 for Bi-LSTM and 4.34 for N-BEATS. For inference speed, GPFormer operates at 0.02 seconds per iteration, while Bi-LSTM and N-BEATS operate at 0.05 and 0.02 seconds, respectively. We chose RMSE as the major evaluation metric due to its widespread use in the literature [20] and its direct representation of prediction accuracy. Regarding the GVAS dataset, which exhibits the large variability compared to the training set (Table I), we further employed CEG plots to evaluate the clinical efficacy sepeartely for people with and without diabetes, as depicted in Fig. 6. The results demonstrated that the proposed model can provide reliable decision support for both individuals with and without diabetes. However, drawing a head-to-head comparison between the proposed model and existing studies can be challenging due to differences in experimental settings, which encompass aspects such as the emphasis on population-level models, variations in the datasets used, and constraints related to data availability.

The effectiveness of GPFormer is grounded in the extensive population size offered by the REPLACE-BG dataset. As evidenced in Fig. 2, the diverse distribution of CGM trajectories allows GPFormer to learn a broad spectrum of glucose patterns. To the best of our knowledge, REPLACE-BG currently remains the largest publicly available dataset in terms of participant numbers. We have validated GPFormer on this dataset, as well as on three additional datasets - one publicly available and two proprietary. However, as a future direction, it would be insightful to extend our validation to a wider range of diverse and recent large-scale datasets, such as the ShanghaiT1DM & T2DM datasets [62] to offer valuable diversity in ethnic backgrounds. Incorporating these datasets may require adjustments to our model architecture to account for the different resolutions of CGM readings. This further validation would potentially enhance the generalizability and robustness of our model. The main backbone of our GPFormer model is based on the Transformer architecture, which has demonstrated remarkable success in various sequence modeling tasks, including recent large language models. To validate the efficacy of our approach, we conducted comprehensive ablation studies comparing GPFormer with variants using vanilla MHSA and without the MLDG framework, as detailed in Fig. 7. These studies demonstrate the superior performance of our proposed backbone. When compared with Bi-LSTM and N-BEATS across multiple datasets, our backbone consistently outperformed these existing methods in terms of glucose level prediction, hypoglycemia prediction, and generalization across different patient populations, as shown in Table II and III. In future work, we plan to incorporate this backbone model into larger architectures and explore its potential in federated learning settings as more extensive CGM datasets become available. Furthermore, the GPFormer model will be integrated into automated insulin delivery systems to augment decision support, such as facilitating predictive insulin suspension [63]. The effectiveness of this integration will be validated through both *in silico* simulations and real clinical trials. Given its adaptability for inpatient settings, we also aim to deploy GPFormer in hospital environments via cloud platforms [24], enhancing patient care in areas such as the intensive care unit. This extension of the application is a critical step towards achieving efficient glycemic control in diverse clinical contexts.

V. Conclusion

In this work, we introduce GPFormer, a glucose prediction model that harnesses the power of a sparse Transformer architecture and domain generalization techniques for real-time multi-horizon glucose prediction across diverse populations. One of the key strengths of this research lies in its inclusion of hospitalized subjects with limited data availability, presenting a realistic yet challenging scenario. When evaluated on four clinical datasets, GPFormer demonstrated robust performance, consistently outperforming four well-established machine learning-based methods in terms of minimizing RMSE, which highlights its potential to significantly improve patient care in various real-world settings.

VI. Appendix

A. Ablation Studies

To investigate the efficacy of the modules and techniques incorporated in GPFormer, we conducted ablation studies by individually excluding the quantile loss function, sparsity mechanism, timestamp input features, and meta-learning for domain generalization. The overall RMSE, calculated by averaging the mean square error for all prediction horizons from 5 minutes to 120 minutes, which is the same metric used in cross-validation, was employed to evaluate the performance of each ablation study. The results demonstrate that each component enhanced GPFormer's predictive power by reducing the RMSE on the four clinical datasets. Notably, the integration of MLDG led to the most significant improvement, aligning with the observations presented in Fig. 5.

B. Data Partitioning

Fig. 8 shows the data partitioning strategy used in our experiments to develop population-level models.

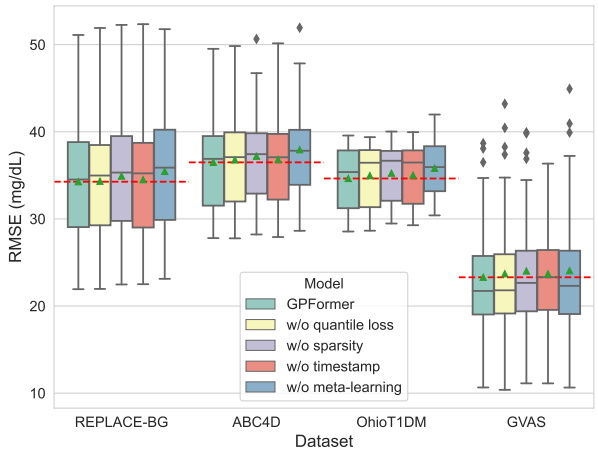


Fig. 7: RMSE performance of GPFormer variants with certain components ablated, illustrating the impact of each component on the model's predictive power. The green triangles show the mean RMSE for each ablated model, while the red dotted line represents the RMSE of the complete GPFormer model, serving as a benchmark for comparison.

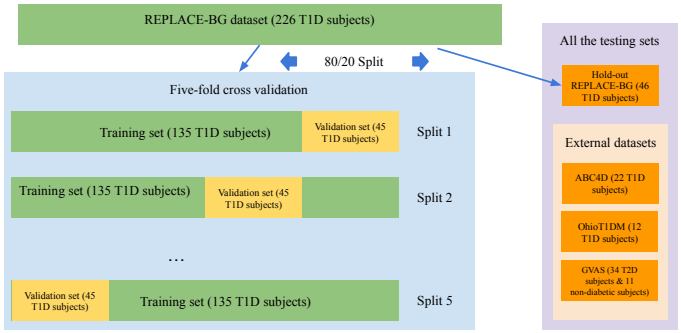


Fig. 8: Visualization of data partitioning strategy utilized in this study for the development of population-level models. Initially, we conducted an 80/20 split on the REPLACE-BG dataset, resulting in a training set and a separate hold-out testing set. Following this, a five-fold cross-validation was performed to facilitate model validation and hyperparameter optimization. To ensure unbiased evaluation, three external CGM datasets were also incorporated.

Layer	Output shape	Param (#)
DataEmbedding (Enc)	[:, 72, 128]	640
Encoder (Attention layer 1)	[:, 72, 128]	593,024
Encoder (Attention layer 2)	[:, 72, 128]	593,024
Encoder (Attention layer 3)	[:, 72, 128]	593,024
Encoder (Norm)	[:, 72, 128]	256
DataEmbedding (Dec)	[:, 30, 128]	640
Decoder (Attention layer 1)	[:, 30, 128]	659,328
Decoder (Norm)	[:, 30, 128]	256
Decoder (Projection)	[:, 30, 5]	645
CPU Inference time per batch (s):	0.98	
CPU latency (s):	0.01	
GPU Inference time per batch (s):	0.05	
GPU Throughput (samples/s):	5446.81	
GPU Training speed (iter/s):	0.23	
Total trainable params:	2,440,837 (9.8 MB for Float32)	

TABLE IV: Implementation details of GPFormer

Hyperparameter	Value
Batch size	256
Dimension of hidden state	128
Epoch number	50
Iterations per epoch	500
Learning rate	1×10^{-3}
Look-back window size L	72
Number of attention heads	8
Number of encoder layers	3
Number of decoder layers	1
Number of query domains	8
Number of support domains	16
Start token length	6
Weight decay	5×10^{-4}

TABLE V: Model hyperparameters

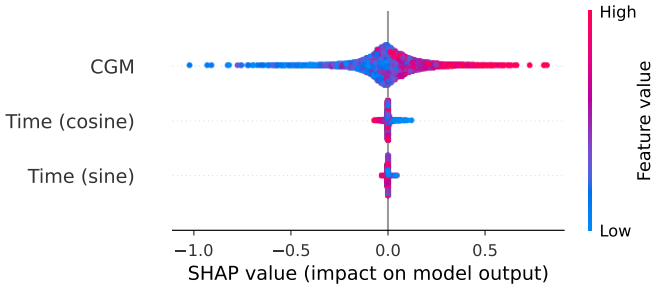


Fig. 9: SHAP values quantifying the contribution of each input feature to the model's predictions, using the magnitude and sign of the values to indicate the strength and direction of the feature's impact.

C. Details of Model Implementation

Table IV provides detailed information on the parameters and outputs of each layer of GPFormer, as well as its training speed, latency, and throughput. The output shape of [30, 5] consists of 5 quantiles in the last dimension. The output sequence, with a length of 30, comprises a label length L' of 6 and a prediction length τ of 24. Specifically, the last 24 outputs are used for predictions up to two hours.

Table V lists the hyperparameters used in our model. The deep learning models were developed using Python 3.9.12 and PyTorch 1.11.0, and were accelerated by an NVIDIA GeForce GTX 1080Ti GPU. The code implementation can be found here.

D. Model Interpretability

We performed SHapley Additive exPlanations (SHAP) [64] to interpret GPFormer and visualize the impact of input features on model predictions using the hold-out REPLACE-BG testing set, as depicted in Fig. 9. The results show that the majority of large CGM values lead to an increase in model output, while small CGM values lead to a decrease, which aligns with the existing literature [65]. Moreover, it is observed that small cosine and sine timestamps lead to an increase in glucose predictions. This finding is consistent with the fact that cosine and sine timestamps equal to -1 correspond to 12:00 and 18:00, respectively, which are typical times for lunch and dinner when food intake would elevate glucose levels.

VII. Acknowledgement

This research was funded by Engineering and Physical Sciences Research Council (EPSRC EP/P00993X/1). Taiyu Zhu was supported by President's Ph.D. Scholarship at Imperial College London. Ioannis Afentakis was supported by the UKRI Centre for Doctoral Training in AI for Healthcare (`http://ai4health.io`; grant no. EP/S023283/1). Ryan Armiger is affiliated to the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Infections in partnership with the UK Health Security Agency (previously PHE), in collaboration with Imperial Healthcare Partners, University of Cambridge and University of Warwick. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or the UK Health Security Agency.

The source of the REPLACE-BG data is Aleppo *et al.* (2017). REPLACE-BG RCT Protocol 10-27-15 v2.0. Retrieved from `https://public.jaeb.org/dataset/546`. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by Aleppo *et al.*

[1] A. Green, S. M. Hede, C. C. Patterson, S. H. Wild, G. Imperatore, G. Roglic, and D. Beran, "Type 1 diabetes in 2017: global estimates of incident and prevalent cases in children and adults," *Diabetologia*, vol. 64, no. 12, pp. 2741–2750, 2021.

[2] L. A. DiMeglio, C. Evans-Molina, and R. A. Oram, "Type 1 diabetes," *The Lancet*, vol. 391, no. 10138, pp. 2449–2462, 2018.

[3] E. W. Gregg, N. Sattar, and M. K. Ali, "The changing face of diabetes complications," *The Lancet Diabetes & Endocrinology*, vol. 4, no. 6, pp. 537–547, 2016.

[4] L. Heinemann, G. Freckmann, D. Ehrmann, G. Faber-Heinemann, S. Guerra, D. Waldenmaier, and N. Hermanns, "Real-time continuous glucose monitoring in adults with type 1 diabetes and impaired hypoglycaemia awareness or severe hypoglycaemia treated with multiple daily insulin injections (HypoDE): a multicentre, randomised controlled trial," *The Lancet*, vol. 391, no. 10128, pp. 1367–1377, 2018.

[5] M. Lind, W. Polonsky, I. B. Hirsch, T. Heise, J. Bolinder, S. Dahlqvist, E. Schwarz, A. F. Ólafsdóttir, A. Frid, H. Wedel *et al.*, "Continuous glucose monitoring vs conventional therapy for glycemic control in adults with type 1 diabetes treated with multiple daily insulin injections: the GOLD randomized clinical trial," *JAMA*, vol. 317, no. 4, pp. 379–387, 2017.

[6] C. A. van Beers, J. H. DeVries, S. J. Kleijer, M. M. Smits, P. H. Geelhoed-Duijvestijn, M. H. Kramer, M. Diamant, F. J. Snoek, and E. H. Serné, "Continuous glucose monitoring for patients with type 1 diabetes and impaired awareness of hypoglycaemia (IN CONTROL): a randomised, open-label, crossover trial," *The Lancet Diabetes & Endocrinology*, vol. 4, no. 11, pp. 893–902, 2016.

[7] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, E. Bosi, B. A. Buckingham, W. T. Cefalu, K. L. Close *et al.*, "Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.

[8] L. Bally, H. Thabit, S. Hartnell, E. Andereggen, Y. Ruan, M. E. Wilinska, M. L. Evans, M. M. Wertli, A. P. Coll, C. Stettler *et al.*, "Closed-loop insulin delivery for glycemic control in noncritical care," *New England Journal of Medicine*, vol. 379, no. 6, pp. 547–556, 2018.

[9] A. Wallia, G. E. Umpierrez, R. J. Rushakoff, D. C. Klonoff, D. J. Rubin, S. Hill Golden, C. B. Cook, B. Thompson, and D. C. G. M. in the Hospital Panel, "Consensus statement on inpatient use of continuous glucose monitoring," *Journal of Diabetes Science and Technology*, vol. 11, no. 5, pp. 1036–1044, 2017.

[10] G. M. Davis, E. Faulds, T. Walker, D. Vigliotti, M. Rabinovich, J. Hester, L. Peng, B. McLean, P. Hannon, N. Poindexter *et al.*, "Remote continuous glucose monitoring with a computerized insulin infusion protocol for critically ill patients in a COVID-19 medical ICU: proof of concept," *Diabetes Care*, vol. 44, no. 4, pp. 1055–1058, 2021.

[11] R. R. Longo, H. Elias, M. Khan, and J. J. Seley, "Use and accuracy of inpatient cgm during the covid-19 pandemic: an observational study of general medicine and icu patients," *Journal of Diabetes Science and Technology*, vol. 16, no. 5, pp. 1136–1143, 2022.

[12] D. C. Klonoff, K. T. Nguyen, N. Y. Xu, A. Gutierrez, J. C. Espinoza, and A. P. Vidmar, "Use of continuous glucose monitors by people without diabetes: an idea whose time has come?" *Journal of Diabetes Science and Technology*, p. 19322968221110830, 2022.

[13] S. Berry, K. Bermingham, A. Valdes, P. Franks, J. Wolf, and T. Spector, "Optimised glucose "time in range" using continuous glucose monitors in 4,805 non-diabetic individuals is associated with favourable diet and health: The ZOE PREDICT studies," *Current Developments in Nutrition*, vol. 6, p. 1108, 2022.

[14] C. Park and Q. A. Le, "The effectiveness of continuous glucose monitoring in patients with type 2 diabetes: a systematic review of literature and meta-analysis," *Diabetes Technology & Therapeutics*, vol. 20, no. 9, pp. 613–621, 2018.

[15] M. Sinha, K. M. McKeon, S. Parker, L. G. Goergen, H. Zheng, F. H. El-Khatib, and S. J. Russell, "A comparison of time delay in three continuous glucose monitors for adolescents and adults," *Journal of Diabetes Science and Technology*, vol. 11, no. 6, pp. 1132–1137, 2017.

[16] C. Mathieu, P. Gillard, and K. Benhalima, "Insulin analogues in type 1 diabetes mellitus: getting better all the time," *Nature Reviews Endocrinology*, vol. 13, no. 7, pp. 385–399, 2017.

[17] B. Bent, P. J. Cho, M. Henriquez, A. Wittmann, C. Thacker, M. Feinglos, M. J. Crowley, and J. P. Dunn, "Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches," *npj Digital Medicine*, vol. 4, no. 1, p. 89, 2021.

[18] E. Longato, G. P. Fadini, G. Sparacino, A. Avogaro, L. Tramontan, and B. Di Camillo, "A deep learning approach to predict diabetes' cardiovascular complications from administrative claims," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3608–3617, 2021.

[19] J. Theis, W. L. Galanter, A. D. Boyd, and H. Darabi, "Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 388–399, 2021.

[20] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2021.

[21] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *Journal of Healthcare Informatics Research*, vol. 4, no. 3, pp. 308–324, 2020.

[22] 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.

[23] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 193–204, 2023.

[24] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, and P. Georgiou, "IoMT-enabled real-time blood glucose prediction with deep learning and edge computing," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3706–3719, 2023.

[25] T. Zhu, C. Uduku, K. Li, P. Herrero, N. Oliver, and P. Georgiou, "Enhancing self-management in type 1 diabetes with wearables and deep learning," *npj Digital Medicine*, vol. 5, no. 1, p. 78, 2022.

[26] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *npj Digital Medicine*, vol. 4, no. 1, p. 109, 2021.

[27] S. M. A. Zaidi, V. Chandola, M. Ibrahim, B. Romanski, L. D. Mastrandrea, and T. Singh, "Multi-step ahead predictive model for blood glucose concentrations of type-1 diabetic patients," *Scientific Reports*, vol. 11, no. 1, p. 24332, 2021.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," *Proceedings of the 2020*

conference on empirical methods in natural language processing: system demonstrations, pp. 38–45, 2020.

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[31] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," in *International Joint Conference on Artificial Intelligence(IJCAI)*, 2023.

[32] L. Monnier and C. Colette, "Glycemic variability: should we and can we prevent it?" *Diabetes Care*, vol. 31, no. Supplement_2, pp. S150–S154, 2008.

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

[34] A. Ceriello, L. Monnier, and D. Owens, "Glycaemic variability in diabetes: clinical and therapeutic implications," *The Lancet Diabetes & Endocrinologyy*, vol. 7, no. 3, pp. 221–230, 2019.

[35] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[36] Y. Zhang, K. M. Bullard, G. Imperatore, C. S. Holliday, and S. R. Benoit, "Proportions and trends of adult hospitalizations with diabetes, united states, 2000–2018," *Diabetes Research and Clinical Practice*, vol. 187, p. 109862, 2022.

[37] C. Fan, Y. Zhang, Y. Pan, X. Li, C. Zhang, R. Yuan, D. Wu, W. Wang, J. Pei, and H. Huang, "Multi-horizon time series forecasting with temporal attention learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2527–2535.

[38] E. Acuna, R. Aparicio, and V. Palomino, "Analyzing the performance of transformers for the prediction of the blood glucose level considering imputation and smoothing," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 41, 2023.

[39] S.-M. Lee, D.-Y. Kim, and J. Woo, "Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1600–1611, 2023.

[40] 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.

[41] 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*, vol. 18, no. 2, pp. 236–246, 2024.

[42] G. Aleppo, K. J. Ruedy, T. D. Riddlesworth, D. F. Kruger, A. L. Peters, I. Hirsch, R. M. Bergenstal, E. Toschi, A. J. Ahmann, V. N. Shah *et al.*, "REPLACE-BG: a randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes," *Diabetes Care*, vol. 40, no. 4, pp. 538–545, 2017.

[43] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *The 5th KDH workshop, ECAI 2020*, 2020, pp. 71–74.

[44] 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," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11 106–11 115, 2021.

[45] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Transformer dissection: An unified understanding for transformer's attention via the lens of kernel," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

[46] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[47] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[48] R. Unsworth, R. Armiger, N. Jugnee, M. Thomas, P. Herrero, P. Georgiou, N. Oliver, and M. Reddy, "Safety and efficacy of an adaptive bolus calculator for type 1 diabetes: A randomized controlled crossover study," *Diabetes Technology & Therapeutics*, vol. 25, no. 6, pp. 414–425, 2023.

[49] L. Preechasuk, S. K. Rilstone, W. Xi Tang, J. Man, M. Yang, E. Zhao, L. Hoque, E. Tuncay, P. Wilding, O. Halse, S. Banerjee, N. Oliver, and N. E. Hill, "933-P: Glycemic Level and Glycemic Variability in Acute Ischemic Stroke and Functional Outcome at Discharge—A Continuous Glucose Monitoring Study," *Diabetes*, vol. 72, no. Supplement, p. 933, 2023.

[50] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[52] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

[53] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, "Evaluating clinical accuracy of systems for self-monitoring of blood glucose," *Diabetes Care*, vol. 10, no. 5, pp. 622–628, 1987.

[54] H. Rubin-Falcone, I. Fox, and J. Wiens, "Deep residual time-series forecasting: Application to blood glucose prediction." in *The 5th International Workshop on Knowledge Discovery in Healthcare Data in the 24th ECAI*, 2020, pp. 105–109.

[55] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting blood glucose with an lstm and bi-lstm based deep neural network," in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE, 2018, pp. 1–5.

[56] E. I. Georga, V. C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 71–81, 2012.

[57] R. M. Bergenstal, R. W. Beck, K. L. Close, G. Grunberger, D. B. Sacks, A. Kowalski, A. S. Brown, L. Heinemann, G. Aleppo, D. B. Ryan *et al.*, "Glucose management indicator (GMI): a new term for estimating a1c from continuous glucose monitoring," *Diabetes Care*, vol. 41, no. 11, pp. 2275–2280, 2018.

[58] M. L. Johnson, T. W. Martens, A. B. Criego, A. L. Carlson, G. D. Simonson, and R. M. Bergenstal, "Utilizing the ambulatory glucose profile to standardize and implement continuous glucose monitoring in clinical practice," *Diabetes Technology & Therapeutics*, vol. 21, no. S2, pp. S2–17, 2019.

[59] C. Mosquera-Lopez and P. G. Jacobs, "Incorporating glucose variability into glucose forecasting accuracy assessment using the new glucose variability impact index and the prediction consistency index: An lstm case example," *Journal of Diabetes Science and Technology*, vol. 16, no. 1, pp. 7–18, 2022.

[60] M. I. Maiorino, G. Bellastella, O. Casciano, P. Cirillo, V. Simeon, P. Chiodini, M. Petrizzo, M. Gicchino, O. Romano, P. Caruso *et al.*, "The effects of subcutaneous insulin infusion versus multiple insulin injections on glucose variability in young adults with type 1 diabetes: the 2-year follow-up of the observational metro study," *Diabetes Technology & Therapeutics*, vol. 20, no. 2, pp. 117–126, 2018.

[61] S. N. Mehta, N. Quinn, L. K. Volkening, and L. M. Laffel, "Impact of carbohydrate counting on glycemic control in children with type 1 diabetes," *Diabetes Care*, vol. 32, no. 6, pp. 1014–1016, 2009.

[62] Q. Zhao, J. Zhu, X. Shen, C. Lin, Y. Zhang, Y. Liang, B. Cao, J. Li, X. Liu, W. Rao *et al.*, "Chinese diabetes datasets for data-driven machine learning," *Scientific Data*, vol. 10, no. 1, p. 35, 2023.

[63] E. Bosi, P. Choudhary, H. W. de Valk, S. Lablanche, J. Castañeda, S. de Portu, J. Da Silva, R. Ré, L. Vorrink-de Groot, J. Shin *et al.*, "Efficacy and safety of suspend-before-low insulin pump technology in hypoglycaemia-prone adults with type 1 diabetes (SMILE): an open-label randomised controlled trial," *The Lancet Diabetes & Endocrinology*, vol. 7, no. 6, pp. 462–472, 2019.

[64] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[65] F. Prendin, J. Pavan, G. Cappon, S. Del Favero, G. Sparacino, and A. Facchinetti, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP," *Scientific Reports*, vol. 13, no. 1, p. 16865, 2023.