Population-Specific Glucose Prediction in Diabetes Care with Transformer-Based Deep Learning on the Edge

Taiyu Zhu, *Member, IEEE*, Lei Kuang, *Graduate Student Member, IEEE*, Chengzhe Piao, *Graduate Student Member, IEEE*, Junming Zeng, *Member, IEEE*, Kezhi Li, *Member, IEEE*, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*— Leveraging continuous glucose monitoring (CGM) systems, real-time blood glucose (BG) forecasting is essential for proactive interventions, playing a crucial role in enhancing the management of type 1 diabetes (T1D) and type 2 diabetes (T2D). However, developing a model generalized to a population and subsequently embedding it within a microchip of a wearable device presents significant technical challenges. Furthermore, the domain of BG prediction in T2D remains under-explored in the literature. In light of this, we propose a population-specific BG prediction model, leveraging the capabilities of the temporal fusion Transformer (TFT) to adjust predictions based on personal demographic data. Then the trained model is embedded within a system-on-chip, integral to our low-power and low-cost customized wearable device. This device seamlessly communicates with CGM systems through Bluetooth and provides timely BG predictions using edge computing. When evaluated on two publicly available clinical datasets with a total of 124 participants with T1D or T2D, the embedded TFT model consistently demonstrated superior performance, achieving the lowest prediction errors when compared with a range of machine learning baseline methods. Executing the TFT model on our wearable device requires minimal memory and power consumption, enabling continuous decision support for more than 51 days on a single Li-Poly battery charge. These findings demonstrate the significant potential of the proposed TFT model and wearable device in enhancing the quality of life for people with diabetes and effectively addressing real-world challenges.**

***Index Terms*— Artificial intelligence, deep learning, diabetes, edge computing, glucose prediction, low power wearable device, Transformer.**

I. Introduction

This work was supported by EPSRC EP/P00993X/1, EPSRC EP/S021612/1, and President's Ph.D. Scholarship 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).

L. Kuang, J. Zeng, and P. Georgiou are with Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom (e-mail: {lei.kuang18, junming.zeng16, pantelis}@imperial.ac.uk).

C. Piao and K. Li are with Institute of Health Informatics, University College London, London, United Kingdom. (e-mail: {chengzhe.piao.21, ken.li}@ucl.ac.uk).

ACCORDING to the International Diabetes Federation, the estimated global prevalence of diabetes has surpassed half a billion [1]. Type 1 diabetes (T1D), accounting for around 10% of all cases, is an autoimmune condition wherein the immune system erroneously targets and annihilates the pancreatic β-cells [2]. People with T1D are dependent on exogenous insulin and necessitate consistent blood glucose (BG) monitoring and daily management [3]. Type 2 diabetes (T2D), representing over 90% of all cases, primarily arises due to a combination of deteriorating insulin secretion and insulin resistance, or either condition in isolation [4].

To enhance the tracking of BG trends and gain insights into glucose variability, real-time continuous glucose monitoring (CGM) systems have been introduced in recent years, and the technology has been swiftly evolving. These systems have been demonstrated significantly improved glycemic control in T1D [5]–[7] and T2D care [8], [9]. CGM sensors typically measure glucose concentration in the interstitial fluid and subsequently convert these measurements to BG levels through built-in algorithms. This process inevitably introduces the time delay between the CGM readings and the actual BG levels [10], thereby emphasizing the importance of BG prediction. Accurate prediction of BG levels is pivotal in both T1D and T2D for initiating timely medical interventions. This proactive strategy holds significant promise for effectively minimizing the risks associated with hyperglycemia and hypoglycemia [11]. Hence, it has the potential to mitigate the severity and incidence of diabetes-related complications, such as cardiovascular disease, kidney disease, retinopathy and diabetic ketoacidosis [12]. These complications not only damage multiple physiological systems but could also lead to life-threatening events if left unaddressed.

By continuously providing BG readings at frequent intervals, CGM generates high-resolution time series data that are valuable for developing machine learning algorithms. In this context, deep learning algorithms, have emerged as the state-of-the-art approaches across various tasks in diabetes care [13], especially in BG prediction [14], [15]. The complex architectures of deep neural networks are particularly well-suited for capturing the nonlinear relationships and temporal dependencies present in raw CGM data while minimizing the need for extensive feature engineering work [16]. Re-

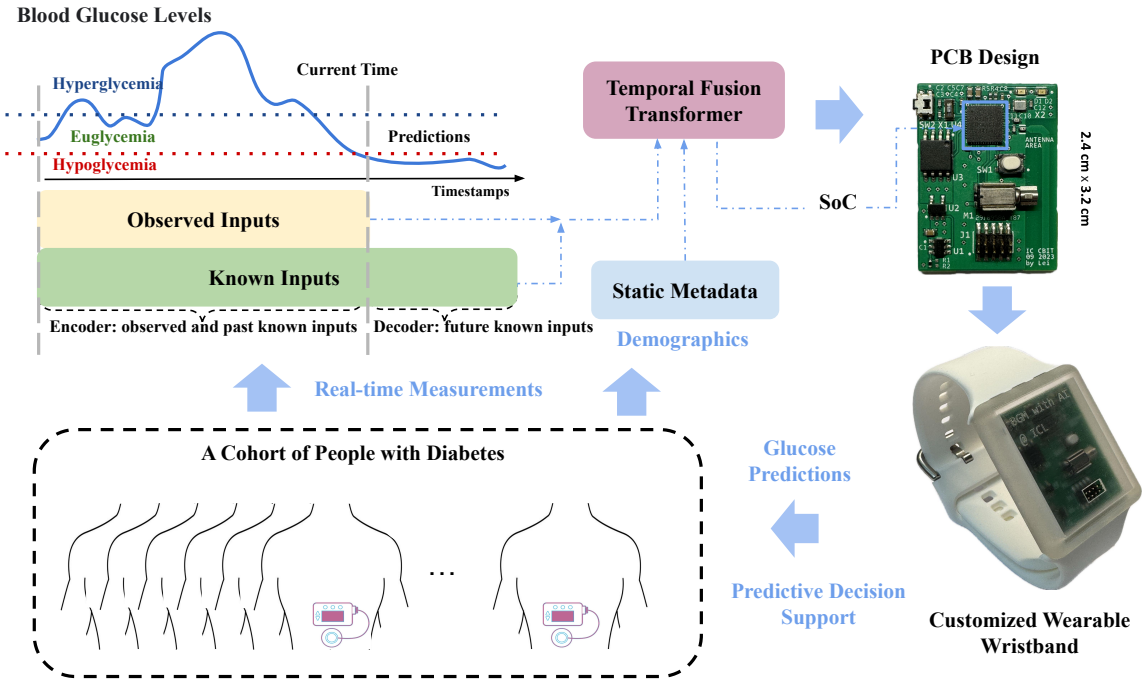


Fig. 1: Architecture of the proposed population-specific diabetes management system that integrates real-time CGM measurements and static metadata, utilizing a temporal fusion Transformer (TFT) for real-time BG predictions. The encoder of TFT processes the past known and observed features, while the decoder focuses on future known features. The model is subsequently embedded on a system-on-a-chip (SoC) of a customized printed circuit board (PCB), enabling seamless interaction with CGM and real-time decision support for a cohort of people with diabetes.

lying on their capacity to maintain internal states for capturing hidden sequential patterns, recurrent neural networks (RNNs) have been widely employed as many-to-one models for single-horizon BG prediction. Specifically, long short-term memory (LSTM) [14], [15], [17] and gated recurrent units (GRUs) [18]–[20] are frequently utilized, as they address some of the limitations of vanilla RNNs, such as the vanishing and exploding gradient problems.

Aiming to further improve predictive performance, recent efforts have focused on the integration of RNNs with conventional attention mechanisms to harness important state information [21], [22]. Emerging studies have adopted meta-learning for personalizing BG prediction models from a population model [22], [23]. Despite this advancement, such approaches typically necessitate separate models catering to different subjects or prediction horizons (PHs). The majority of existing research has concentrated on patient-specific and single-horizon models, predominantly utilizing conventional deep neural networks for T1D management without consideration for hardware deployment [13]. This leaves a significant gap when it comes to the actual application in real-world healthcare systems for broader applications and for the majority of people with diabetes, i.e., T2D. Our work diverges from this trend by employing a unified model that accommodates multiple PHs and both T1D and T2D subjects, thus streamlining model complexity and facilitating deployment.

Recently, the spotlight in deep learning has been captured by Transformers that are fundamentally built on self-attention mechanisms [24]. The Transformer-based models have served as the foundational architecture for a range of large language models, including BERT [25] and GPT [26]. These models have reached a level of performance comparable to human expertise in the field of natural language processing [27], which are regarded as early-stage manifestations of artificial general intelligence [28]. Given their broad applicability and proficiency across multiple tasks and domains, particularly in sequence processing, a variety of Transformer variants focused on time series forecasting have been proposed in recent literature [29], such as Informer [30], FEDformer [31], and Crossformer [32]. In BG prediction, the available data extends beyond merely time series CGM readings. Additional information, such as demographics in the form of static data, is also accessible, which can offer a more comprehensive approach to model personalized glucose dynamics. In this scenario, the temporal fusion Transformer (TFT) offers a specialized solution for processing multi-modal data by employing gating mechanisms [33]. This encompasses observed data features, future known information, and static features, making it highly applicable for BG prediction that involves diverse data sources.

Pioneering studies have deployed trained deep learning models on software based computation platforms, such as smartphones with iOS or Android operating systems [19], [20], [22], using existing software libraries, such as TensorFlow Lite. However, Smartphones and smartwatches, given their multifunctional nature, encounter several challenges. Frequent software updates can disrupt service continuity and lead to erroneous medical interpretation [34]. Battery life is another

critical issue; devices often cannot sustain the demands of real-time algorithms and continuous Bluetooth connectivity without frequent recharging [35]. Furthermore, data privacy emerges as a paramount concern, with the potential for unauthorized data access deterring users from adopting medical apps [35]. In light of these issues, the need for a dedicated, low-power, and low-cost wearable wristband in diabetes management becomes evident. Such a device would alleviate the aforementioned concerns and has precedent in the success of dedicated handheld devices, which have demonstrated their efficacy in large-scale clinical trials [36], [37]. Meanwhile, it is challenging to implement deep learning models in wearable medical devices with intensive computational resources and memory constraints [18], [38], [39]. This challenge is further magnified if the device is designed to host multiple personalized models for population-level prediction. Therefore, the urgent need arises for edge computing-enabled Transformer models that can operate on systems on a chip (SoCs), since these models are typically characterized by a large number of model parameters and complex architectures.

In an effort to tackle these challenges, we propose a population-specific mutli-horizon BG prediction model based on the TFT architecture, as shown in Fig. 1. Additionally, we propose a hardware framework designed for the implementation of the trained models on SoCs, which enables real-time communication with CGM systems via Bluetooth and facilitates predictive decision support tailored for a cohort of individuals with diabetes through edge computing. The proposed methodology was evaluated on two publicly available CGM datasets and achieved the best prediction accuracy when compared against a set of machine learning baseline methods. The preliminary results focusing on 12 T1D subjects of the OhioT1DM dataset have been previously published in the 2023 IEEE International Symposium on Circuits and Systems (ISCAS) [40]. In the present work, we have expanded the scope to include scenarios involving T2D care and developed the model on 24 T1D subjects and 100 T2D subjects by incorporating additional static data features. Moreover, we have upgraded the wearable device with a new design for a more compact printed circuit board (PCB). The shift to a Li-Poly battery from the earlier coin cell battery, coupled with a vibration motor, has considerably improved its usability and portability.

II. Methodology

A. Problem Formulation

Given the current timestep t and the target scalar BG time series represented as $\mathbf{y}$, the objective of multi-horizon BG prediction with a specified PH of τ is to forecast the future BG time series $\mathbf{y}_{t:t+\tau}$. To facilitate this prediction, various data features can be utilized. These include a sequence of historical BG levels, observed inputs, known inputs, and static features, such as the age and gender of an individual.

B. Temporal Fusion Transformers in Glucose Prediction

Traditional deep learning models are generally limited to utilizing historical BG levels and observed inputs through a multivariate input. To effectively manage multi-modal data, TFT incorporates two novel modules: gated residual networks (GRNs) and variable selection networks (VSNs), built upon a gate mechanism known as gated linear units (GLUs) [41], as shown in Fig. 2. By leveraging these modules, the TFT aims to appropriately weigh and integrate the diverse types of input data. Specifically, the GLU provides the flexibility to modulate nonlinear contributions by merging a linear feature transformation with a gated layer that employs a Sigmoid activation function σ, which is denoted as

$$\mathrm{GLU}(\mathbf{z}) = \sigma(\mathbf{W}_g\mathbf{z} + \mathbf{b}_g) \odot (\mathbf{W}_l\mathbf{z} + \mathbf{b}_l), \tag{1}$$

where $\mathbf{z}$ is the input data; $\odot$ is the element-wise product; $\mathbf{W}$ and $\mathbf{b}$ are the weights and biases. By integrating the GLU with residual connections and layer normalization LN, the GRN is designed to determine the relationship between primary input $\mathbf{p}$ and optional exogenous inputs $\mathbf{e}$, which is formulated as

$$\mathrm{GRN}(\mathbf{p}, \mathbf{e}) = \mathrm{LN}(\mathbf{p} + \mathrm{GLU}(\mathbf{W}_g\mathbf{a} + \mathbf{b}_g))), \tag{2}$$

$$\mathbf{a} = \mathrm{ELU}(\mathbf{W}_p\mathbf{p} + \mathbf{W}_e\mathbf{e} + \mathbf{b}_a), \tag{3}$$

where $\mathbf{a}$ represents the activated output, which is obtained through exponential linear unit (ELU) activation function that is robust to reduce the impact of outliers or noise in time series data. It is noted that $\mathbf{z}$, $\mathbf{p}$, $\mathbf{e}$ in Equations (1)-(3) can represent various data types, including temporal input, static features, or sub-layer outputs, given the extensive application of GRN and GLU within the TFT model. By integrating GRNs with a Softmax layer, the VSN is configured to generate variable selection weights across the feature dimension for each timestep. These weights are then used to merge the input features, thereby optimizing the model's focus on the most relevant features. Similarly, the covariate encoder in Fig. 2 also employs GRNs to produce distinct context vectors for different functional modules within TFT, including VSN, initial states of LSTM encoder, and the static enrichment layer. The state of LSTM decoder is initialized by the output of LSTM encoder.

In addition to optimizing at the feature level, TFT also incorporates a multi-head self-attention (MHSA) layer [24] to learn long-term dependencies across timesteps while enhancing temporal interpretability by aggregating the outputs of the heads, which is formulated as follows:

$$\mathrm{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i]\mathbf{W}_O, \tag{4}$$

$$\mathbf{H}_i = \mathcal{A}(\mathbf{Q}\mathbf{W}_Q, \mathbf{K}\mathbf{W}_K, \mathbf{V}\mathbf{W}_V), \tag{5}$$

$$\mathcal{A}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = Softmax(\frac{\hat{\mathbf{Q}}\hat{\mathbf{K}}^\mathsf{T}}{\sqrt{D_k}})\hat{\mathbf{V}}, \tag{6}$$

where $\mathcal{A}$ the self-attention mechanism; $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ stand for queries, keys, values, respectively, and $\hat{\mathbf{Q}}$, $\hat{\mathbf{K}}$, $\hat{\mathbf{V}}$ are queries, keys, values combined with head-specific weights, as defined in Equation (5); D_k denotes the dimension of the keys. Here $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the outputs derived from the GRN-based static enrichment layer, integrating insights from static as well as temporal features. The output from the i-th head is symbolized by $\mathbf{H}_i$. In the point-wise feed-forward layer, GRN is further utilized to extract non-linear patterns from

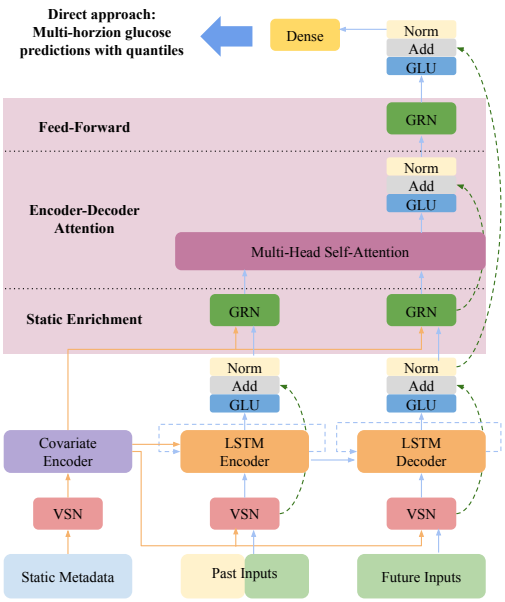


Fig. 2: Diagram of the proposed temporal fusion Transformer. Static demographic data serves as the input for the covariate encoder. The LSTM encoder handles both observed and past known features, while the LSTM decoder solely focuses on future known inputs. Modules with gated mechanisms are employed to learn non-linear relationships from multi-modal data. The predictions of future BG levels are simultaneously generated through a direct approach.

the outputs of MHSA. It is worth noting that GLU and residual connections are implemented to potentially skip over LSTM modules, MHSA, and whole Transformer block. Theses direct paths provide model with the flexibility to bypass these modules if a less complex model proves more effective.

The TFT model leverages a direct approach to generate the predicted sequence in a single forward step. This method predicts all future BG values simultaneously, enhancing computational speed and minimizing the propagation of errors that typically occur in recursive approaches. In contrast to traditional BG prediction models, the proposed model leverages a quantile loss [42] to obtain lower and upper bounds for each predictive value. This not only provides a richer informational context but also grants the flexibility to adjust the sensitivity of predictive warnings and to align the model closely with real-world preferences. This adaptability was underscored as a desirable feature requested by individuals with diabetes in one of our previous clinical focus groups [20]. Given a prediction vector $\hat{\mathbf{y}}_{t:t+\tau}$, the loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \frac{1}{\tau} \sum_{i=t}^{t+\tau} \sum_{q_l \in R} (1 - q_l)(\hat{y}_i - y_i)^+ + q_l (y_i - \hat{y}_i)^+, \quad (7)$$

where $\hat{y}_i$ and y_i is i-th element in target BG time series and the prediction vector, respectively; $R = \{0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98\}$, representing the range of considered quantiles; and $()^+$ denotes the ramp function.

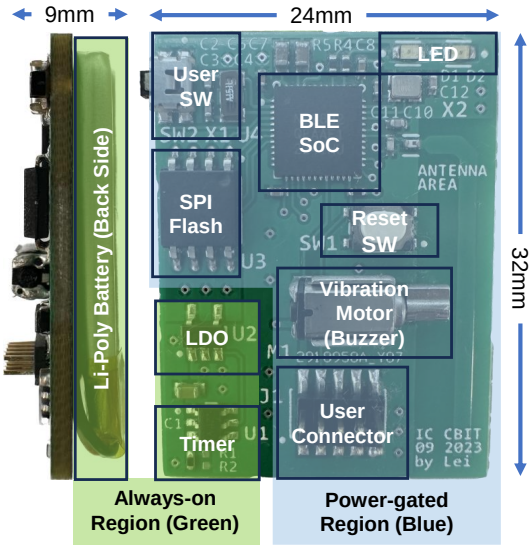


Fig. 3: Block diagram of the optimized device powered by a Li-Poly battery. To further enhance user notification, a vibration motor can be integrated to provide flexibility for individuals, including those with hearing impairments. A power-gating circuit, including a timer and a voltage regulator, is employed to maintain ultra-low power when the device is idle.

C. Wearable Device for Edge Computing

To offer more reliable real-time monitoring services that will not be limited by inconsistent Internet connectivity, as well as low-latency decision-making for patients, a wearable device that employs a Bluetooth low energy (BLE) SoC (Nordic nRF52832) for wireless communication and edge computing is developed. This device is miniaturized into a dimension of 32 mm x 24 mm x 9 mm, as shown in Fig. 3, where a Li-Poly battery is attached to the back side of the PCB with a capacity of 200 mAh to power the device.

Benefiting from the advancement in Li-Poly batteries, the device now supports different capacities of battery life at the cost of increasing in volume, such as 2 mm thicker for another 100 mAh capacity. Given that the nominal voltage of the Li-Poly battery is 3.7 V, a 3 V low-dropout (LDO) regulator is used for voltage regulation. As a result, the generated voltage will be more stable compared to a standalone DC-DC converter which is used for voltage ramping up when using the coin battery. In addition, to cater to a broader range of users, especially those with hearing impairments, we provide an option to select between a coreless vibration motor haptic feedback or a buzzer for low-power sound alerts. Various buzzer modes, vibration patterns, and LED flash modes are designed to alert users of adverse glycemic events, such as hypoglycemia and hyperglycemia. These notifications can be customized according to user preferences. In the end, the device is packed into a 3D-printed case with strips so that it can be wearable.

III. Experiments

A. Clinical Datasets

We employed two publicly available CGM datasets: OhioT1DM [43] and ShanghaiDM [44]. These datasets were specifically curated to promote data-driven algorithms in diabetes research, which encompass diverse populations and clinical scenarios.

1) OhioT1DM: This dataset was collected over an eight-week clinical trial period, which contains the data of 12 T1D individuals. These participants were equipped with Medtronic Enlite real-time CGM and Medtronic insulin pumps. BG levels were systematically recorded at five-minute intervals, alongside the data of insulin dosages from the pumps. Self-reported events, such as meal content and exercise, were logged via a custom smartphone app.

2) ShanghaiDM: This is a recently released dataset that encompasses data of 12 T1D individuals and 100 T2D individuals. These participants were equipped with Abbott FreeStyle Libre flash CGM, which consistently recorded BG levels every 15 minutes. Dietary information and insulin dosages were also self-reported by the participants. This clinical trial spanned 14 days. Each participant underwent a physical examination, responded to a standardized questionnaire, and provided laboratory measurements from medical records, which contributed to a rich set of clinical characteristics.

B. Experimental Setup

1) Data Preprocessing: In this research, we developed two population-specific models: one targeting the OhioT1DM cohort and the other for ShanghaiDM. The OhioT1DM dataset provides a training set and a testing set separately for each participant, spanning approximately six and two weeks of data, respectively. For each individual in the ShanghaiDM dataset, we performed a 80/20 data split, where the initial 80% of data was used for a training set and the latter 20% for a testing set. For both OhioT1DM and ShanghaiDM datasets, the final 25% of each training set was employed as a validation set. This two-step data split is widely used in the existing work on machine learning-based BG prediction [19]–[22].

By aggregating individual data sets, we generated comprehensive population-level training, validation, and testing sets for both two datasets. CGM measurements occasionally present missing values, especially in the OhioT1DM dataset, which arise from factors such as sensor calibrations and occurrence of artifacts. To fill these gaps without drawing on future information, we adopted a linear extrapolation approach and ensured the values adhere to the sensor's functional range of 40-400 mg/dL. The timestamp spanning 24 hours was normalized to the range [0,1], and standard normalization is applied to all the other data features.

2) Model Development: Through the use of a look-back sliding window, we merge observed features from the last 120 minutes, known features, and static demographic data, with the objective of forecasting the subsequent 60-minute BG levels. This is a standard PH in existing literature that aids in proactive interventions [13]. Consequently, we generated mini-batches that encompass both the model's input data and the corresponding target values. Once the data is fed into the TFT model, the parameters are automatically updated through back-propagation, utilizing the Adam optimizer to minimize the quantile loss. To mitigate overfitting, we implement an early stopping mechanism, setting a patience level of 20 across 200 training epochs.

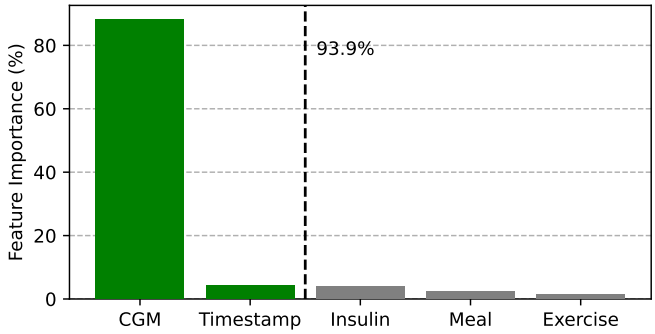


Fig. 4: Feature importance scores derived from the variable selection weights of the VSN in the LSTM encoder [40]. Features represented by green bars were selected for inclusion in the final TFT model.

On the OhioT1DM dataset, we first trained a TFT model with all the available exogenous features. Subsequently, we undertook feature selection based on variable selection weights within the encoder's VSN, which acted as a post-hoc interpretation technique. As depicted in Fig. 4, CGM and timestamp contribute to 93.9% of the overall selection weights. Given their significance, we opted to retain only these two features and proceeded with retraining the model on the OhioT1DM dataset. One notable benefit of exclusively utilizing CGM and timestamp is the feasibility of implementing the model on SoCs, such as the BLE SoC within CGM transmitters or other wearable devices, eliminating the need for manual input. In light of this advantage, we persisted with this setup while training the TFT model on the ShanghaiDM dataset. Hence, in the present study, the observed input data consisted of CGM readings, while the known input data were the timestamps. However, the static demographic features varied between the two datasets based on their availability. For the OhioT1DM model, gender and age were considered, while the ShanghaiDM model extended to gender, age, body mass index, and specific diabetes types. The hyperparameters of each model were fine-tuned using the corresponding validation set and the HyperBand tuner [45].

3) Model Implementation: The TFT architecture was initially developed in Python and then translated into C models. These C models leverage the CMSIS-DSP library that provides high-performance APIs for mathematical functions. The firmware development was carried out using the nRF5 SDK v17.0.2. Parameters from the PyTorch model were allocated in the SoC's FLASH memory, formatted as 32-bit floats, ensuring the fidelity of model inference [18] and achieving a negligible difference of less than 10^{-3} mg/dL between the Python and C models on both datasets. These efforts finally resulted in the embedded population-specific TFT (EPS-TFT). We assessed the BG prediction performance of EPS-TFT by feeding CGM

readings into the SoC and retrieving predictions through a universal asynchronous receiver-transmitter. This was intended to emulate the practical application of the proposed wearable device. When evaluating the power efficiency and the system behaviour, a CGM emulator was employed to transmit real-time CGM measurements to the wearable device through Bluetooth.

4) Baseline Methods: To rigorously evaluate prediction performance, we benchmarked EPS-TFT against with a range of established machine learning baseline methods. In particular, we incorporated conventional machine learnig models, including support vector regression (SVR) [46], XGBoost [47], and linear regression (LR) [48]. We also integrated advanced deep learning models: LSTM [14], N-BEATS [49], and N-HiTS [50]. Utilizing blocks of fully connected layers, N-BEATS has demonstrated proficiency in handling various time-series prediction tasks. Augmenting this approach, N-HiTS introduces hierarchical interpolation and elements of multi-rate sampling to further enhance forecasting accuracy [50]. In our experimental setup, LSTM, DRNN, N-BEATS, and N-HiTS were reconfigured to serve as multi-horizon predictors. SVR, XGBoost, and LR require two separate models trained for single-horizon prediction to address the 30 and 60-minute PHs. All the deep neural network models were crafted using Python 3.9 and PyTorch 1.11. The training process was accelerated using an NVIDIA GTX 1080 Ti.

5) Evaluation Metrics: The standard statistical metrics in glucose prediction are root mean square error (RMSE) and mean absolute error (MAE) [13], which are defined in Equation (8) and (9), respectively. However, given the variability in BG levels across a diverse population, it becomes essential to consider metrics that can account for individual scales. To this end, we introduced the mean absolute percentage error (MAPE), as defined in Equation (10), a percentage-based metric that offers insights into relative prediction errors. Further delving into the clinical implications of prediction errors, we adopted the glucose-specific RMSE (gRMSE) [51], as shown in Equation (11). Based on Clark error [52], this metric implements penalties on predictions that could potentially lead to harmful clinical events, thereby emphasizing the clinical significance of prediction accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y_i}^{(t)} - y_i^{(t)})^2}, \tag{8}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\hat{y_i}^{(t)} - y_i^{(t)}|, \tag{9}$$

$$\text{MAPE} = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{(\hat{y_i}^{(t)} - y_i^{(t)})}{y_i^{(t)}}\right|, \tag{10}$$

$$\text{gRMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}P(\hat{y_i}^{(t)}, y_i^{(t)})(\hat{y_i}^{(t)} - y_i^{(t)})^2}, \tag{11}$$

where T stands for the total number of entries in BG time series, and i represents i-th position in the predicted and ground truth values, consistent with the PH.

TABLE I: Performance of the prediction methods evaluated on the OhioT1DM dataset

Method	RMSE (mg/dL)	MAE (mg/dL)	MAPE (%)	gRMSE (mg/dL)
PH = 30 minutes				
EPS-TFT	$\mathbf{19.1 \pm 2.5}$	$\mathbf{13.1 \pm 1.6}$	$\mathbf{8.5 \pm 1.6}$	$\mathbf{23.3 \pm 3.1}$
N-HiTS	19.6 ± 2.3	13.9 ± 1.6	9.5 ± 1.7	24.5 ± 2.8
N-BEATS	19.4 ± 2.3	13.4 ± 1.6	8.9 ± 1.6	23.8 ± 2.7
LSTM	20.2 ± 2.6	14.3 ± 1.7	9.7 ± 1.7	24.6 ± 3.3
XGBoost	22.1 ± 2.9	15.7 ± 1.9	10.6 ± 2.0	28.2 ± 3.9
SVR	23.5 ± 4.7	16.1 ± 2.3	10.6 ± 2.0	29.9 ± 6.7
LR	22.2 ± 2.8	15.9 ± 2.0	10.9 ± 2.1	27.7 ± 3.7
PH = 60 minutes				
EPS-TFT	$\mathbf{32.3 \pm 3.8}$	$\mathbf{23.2 \pm 2.8}$	$\mathbf{15.5 \pm 2.8}$	$\mathbf{40.1 \pm 4.9}$
N-HiTS	33.0 ± 3.7	24.5 ± 2.9	17.3 ± 3.1	42.5 ± 4.6
N-BEATS	33.8 ± 3.9	24.4 ± 3.0	16.3 ± 3.0	43.0 ± 4.9
LSTM	34.4 ± 4.2	25.4 ± 3.1	17.6 ± 3.3	42.7 ± 5.4
XGBoost	35.6 ± 4.7	26.4 ± 3.4	18.1 ± 3.7	46.6 ± 6.6
SVR	37.0 ± 5.6	27.0 ± 3.9	18.0 ± 3.7	48.5 ± 8.0
LR	36.0 ± 4.6	27.0 ± 3.7	18.7 ± 3.9	46.5 ± 6.6

C. Performance

1) BG Prediction: The results (Mean±STD) of BG prediction for the OhioT1DM dataset and the ShanghaiDM dataset are detailed in Table I and II. Notably, EPS-TFT outperformed all the considered baseline methods, registering the lowest RMSE, MAE, MAPE, and gRMSE across both 30 and 60-minute PHs for the two datasets. Such performance not only highlights the model's superior accuracy but also its promising clinical relevance. In addition, the multi-horizon N-BEATS, N-HiTS, and LSTM models demonstrated significant improvements over traditional single-horizon XGBoost, SVR, and LR models. Appendix A provides a detailed ablation study that invesigates the influence of TFT's submodules on the overall prediction accuracy, offering insights into the contributions of each component.

In Fig. 5, we present two-day glucose trajectories for three indviduals from the two datasets, comparing actual BG levels with the 60-minute EPS-TFT predictions. The dotted blue and green lines stand for the upper and lower bounds, sourced from the 25th and 75th percentiles of the quantile forecasts. The large variability in glucose patterns among the three selected participants is notable. Specifically, the T1D subject from the OhioT1DM dataset experienced both hyperglycemia and hypoglycemia, accompanied by significant fluctuations and missing CGM data. On the other hand, the T1D individual from the ShanghaiDM dataset encountered solely hypoglycemia, whereas the T2D subject was predominantly affected by hyperglycemia. In each of these cases, the predictions exhibit a close alignment with the actual CGM readings and the majority of adverse glycemic events were successfully predicted by EPS-TFT. The quantile-based upper and lower bounds played a crucial role in identifying severe hypoglycemia and hyperglycemia events that the point predictions overlooked. Such consistent performance across diverse real-world clinical scenarios underscores the robust generalization capabilities of the proposed model.

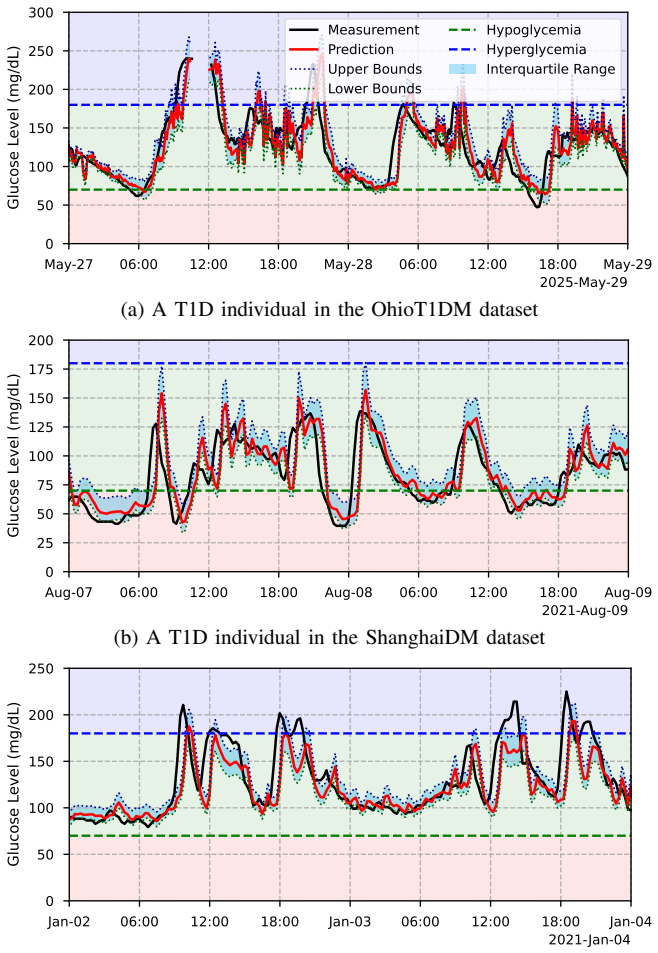
2) Memory Footprint: The final model is implemented and deployed onto the target BLE SoC. It occupies a total of 282.4 KB in Flash space and requires 14.7 KB of RAM memory for computation as indicated in Table III. To optimize

TABLE II: Performance of the prediction methods evaluated on the ShanghaiDM dataset for T1D and T2D subjects

Method	RMSE (mg/dL)	MAE (mg/dL)	MAPE (%)	gRMSE (mg/dL)
PH = 30 minutes				
EPS-TFT	$\mathbf{12.7 \pm 3.8}$	$\mathbf{8.8 \pm 2.8}$	$\mathbf{6.7 \pm 2.3}$	$\mathbf{14.8 \pm 4.9}$
N-HiTS	13.4 ± 4.1	9.3 ± 3.1	7.1 ± 2.3	15.6 ± 5.3
N-BEATS	12.9 ± 4.0	9.0 ± 3.0	6.9 ± 2.2	15.1 ± 5.2
LSTM	13.1 ± 3.7	9.2 ± 2.7	7.1 ± 2.2	15.4 ± 5.0
XGBoost	17.2 ± 7.0	13.1 ± 6.2	11.5 ± 8.3	20.8 ± 9.3
SVR	18.3 ± 9.4	13.8 ± 8.3	11.9 ± 10.3	22.2 ± 12.5
LR	17.7 ± 14.3	12.4 ± 5.6	9.5 ± 4.5	20.3 ± 17.4
PH = 60 minutes				
EPS-TFT	$\mathbf{21.7 \pm 6.9}$	$\mathbf{15.1 \pm 5.1}$	$\mathbf{11.2 \pm 4.1}$	$\mathbf{26.2 \pm 9.5}$
N-HiTS	22.5 ± 7.0	15.7 ± 5.3	11.8 ± 3.9	27.2 ± 9.5
N-BEATS	22.1 ± 7.0	15.4 ± 5.2	11.6 ± 3.7	26.9 ± 9.6
LSTM	22.5 ± 6.8	15.9 ± 5.0	12.0 ± 3.9	27.4 ± 9.4
XGBoost	27.1 ± 10.8	21.0 ± 9.9	18.0 ± 11.7	32.6 ± 13.6
SVR	26.3 ± 10.7	20.1 ± 9.6	16.9 ± 11.5	32.4 ± 14.4
LR	28.8 ± 16.4	21.0 ± 10.6	16.7 ± 7.6	34.1 ± 19.8

TABLE III: Details of Flash and SRAM Memory Footprint

Module	Input	Flash (B)	SRAM (B)	Time (ms)
Input Layer	(2, 36)	0	288	N/A
Encoder VSN	(2, 24)	11,448	14,976	142.8
Decoder VSN	(1, 12)	5,592	6,384	34.5
LSTM	(36, 48)	169,728	14,408	1520.6
Attention	(36, 48)	24,576	13,824	273.6
Feed-Forward	(12, 48)	76,416	9,600	419.2
Dense	(12, 48)	1,380	2,640	7.4
Output Layer	(12, 7)	0	336	N/A



(a) A T1D individual in the OhioT1DM dataset

(b) A T1D individual in the ShanghaiDM dataset

(c) A T2D individual in the ShanghaiDM dataset

Fig. 5: Two-day visualization of the results for three individuals across two datasets. The black solid line depicts actual CGM measurements, while the red line represents 60-minute EPS-TFT predictions on the wearable device. The sky-blue shaded region illustrates the interquantile range between the lower and upper quantile bounds. Background zones in light blue, green, and red stand for hyperglycemia, euglycemia, and hypoglycemia, respectively.

computational efficiency, we precomputed the outputs of the covariate encoder, specifically the context vectors, and stored them in the memory. This approach eliminates the need to recompute these vectors from demographic data during each iteration. Hence, covariate encoder is not shown in Table III. In this case, the computation of the model takes about 2.4 seconds in total.

The encoder VSN takes the 24 most recent CGM readouts with timestamps as the input, while the decoder VSN handles the subsequent 12 timestamps. To accommodate the output generated by both the encoder and decoder layers as input for the LSTM layer, and ensure sufficient room for all iterative results, it is essential to allocate 13.5 KB of RAM space. This allocation constitutes the predominant portion of RAM utilization and is strongly advised to be pre-allocated statistically to enhance system stability.

3) Power Analysis: To measure the energy consumption of the wearable device, a source meter Keithley 2606A is used to supply the voltage and monitor the power in real time. Fig. 6 presents the measurement result, where a complete run cycle typically involves system startup (power gating), BLE scanning and connection, CGM readout, external Flash memory access, edge computation of the prediction model, user notification, and device shutdown (power gating). Such a single operation usually lasts for 15 seconds, depending on the user's response time, achieving an average power that is less than 2 mW throughout the whole operation process. Thanks to the implementation of a power-gating circuit, the power consumption during idle periods is reduced to approximately 0.128 uW. It is noticeable that the most energy-intensive aspect of the process is the computation of the prediction model. During this phase, the BLE SoC accesses extensive weight data stored in its Flash memory and engages the on-chip digital signal processors for massive floating-point arithmetic operations, resulting in peak power consumption reaching approximately 2.7 mW.

The Li-Poly battery boasts a capacity of 200 mAh, providing the device with the capability to sustain continuous operation for a minimum of 88,000 cycles with the buzzer. When leveraging the motor to produce vibrations for user notification, it can still achieve 5,000 operation cycles. These projections are based on the worst-case scenario where notification events occur persistently, and each operation lasts for an average of 15 seconds. For individuals with T1D or T2D, they typically spend approximately 30% of their time in hypoglycemia or hyperglycemia range [53]. During these periods, they would

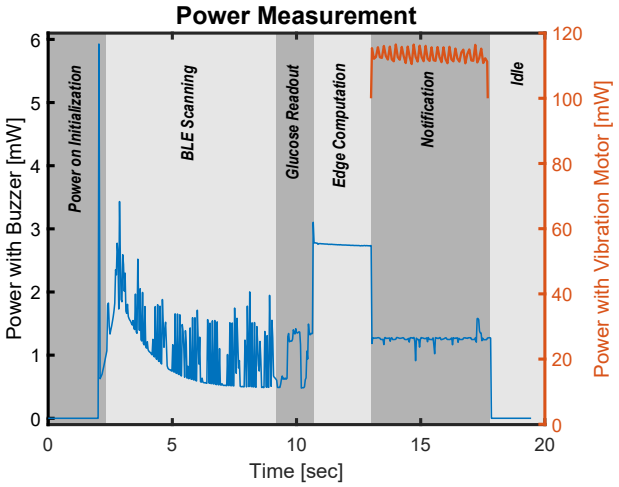


Fig. 6: The entire process encompasses several stages, including BLE scanning, sensor connection, data readout, external memory access, model computation, and user notification. Among these stages, edge computing primarily accounts for the energy consumption when using the buzzer for sound generation. In contrast, the utilization of the vibration motor results in a higher power consumption, peaking at 117 mW, but is typically required only for specific scenarios. On average, when the device employs a buzzer for user notification, its power consumption remains below 2 mW.

receive vibration notifications. In light of this, when paired with a CGM sensor with a five-minute resolution, our wearable device is capable of offering uninterrupted decision support for a span of 51 days with the vibration motor and up to 305 days when using the buzzer. For a CGM sensor with a 15-minute sampling rate, these durations increase to 153 days and 915 days, respectively.

IV. Discussion

In this study, we propose EPS-TFT, a population-specific multi-horizon BG prediction model based on TFT and edge computing. When evaluated on two publicly available clinical datasets encompassing both T1D and T2D individuals, the proposed model demonstrated superior performance. Subsequently, we embedded the trained model into a SoC within a low-power and low-cost wearable device, enabling real-time communication with CGMs and delivering decision support to a cohort of patients. The computational requirements for running the model on the edge are modest, necessitating only 14.7 KB of RAM space and 282.4KB of Flash space. Moreover, a single cycle of decision support consumes less than 2 mW of power on average, underscoring the efficiency of the system. In our prior study [18], we emphasized a patient-specific and single-horizon model utilizing GRUs, with the wearable device being powered by a coin cell battery. Based on feedback from T1D users, in the present study, we significantly downsized the device by introducing a new PCB layout powered by a Li-Poly battery. This compact form factor allows users to conveniently carry the device. The inclusion of a new vibration module proves advantageous for individuals with hearing impairments. Furthermore, the shift to a rechargeable Li-Poly battery also promotes environmental sustainability.

However, the distinctive nature of our approach, which combines population-specific and multi-horizon settings, poses challenges when attempting to draw direct comparisons with results presented in traditional personalized BG prediction. While population-specific and multi-horizon settings introduce additional complexities in model development, they were effectively addressed by leveraging the Transformer architecture and incorporating the gate mechanism to ensure model adaptability on demographic data. The practical implications of this model are profound, especially in real-world clinical scenarios. For example, the wearable we proposed can actively monitor and predict BG levels for a cohort of patients within a hospital setting. This application has gained significant traction, especially in the context of CGM utilization with inpatient settings, since the onset of the COVID-19 pandemic [54], [55].

Meanwhile, an observation from Fig. 4 is the limited feature importance of exogenous events on the prediction. This can possibly be attributed to the inherent inter-individual variability in these daily events when considered at the population level. Given the diverse patterns of meal consumption, insulin intake, and exercise routines among individuals, the consistency and reliability of these features could be compromised. The absence of these features offers several advantages. Firstly, it facilitates automatic closed-loop control without the need for manual input, streamlining the process and reducing potential human errors. Secondly, it optimizes the computational resources of the SoC, ensuring efficient performance and potentially extending the device's battery life.

In Table I and II, the RMSE differences between EPS-TFT and N-BEATS, the next best method, are minor for the 30-minute PH. To assess the significance of these differences, we performed paired t-tests, preceded by Shapiro-Wilk tests to confirm data normality. The results, indicating $p < 0.05$ for both datasets and PHs, confirm that these improvements are statistically significant. While the RMSE enhancement may seem modest in some clinical scenarios, it is crucial for bolus and basal insulin delivery systems, such as precision dosing and artificial pancreases. Such improvements can impact real-time insulin dosing decisions, thereby influencing overall insulin administration and enhancing glycemic control. While deep learning models significantly outperformed conventional machine learning in reducing RMSE, simpler methods such as linear regression, when used in conjunction with smartphone apps [48], are advantageous in specific scenarios. These include situations where real-time decision-making is less important, for users preferring an easy-to-use interface, and in environments where data privacy concerns limit the training of more complex algorithms.

A limitation of this study is its restricted capacity to make cross-population predictions - that is, training the model on one population and evaluating it on another. This is mainly due to the significant variability between populations, as illustrated in Fig. 5, and the differences in the datasets, such as the CGM resolutions (5 minutes vs 15 minutes). Despite this limitation, we have taken initial steps to assess the model general-

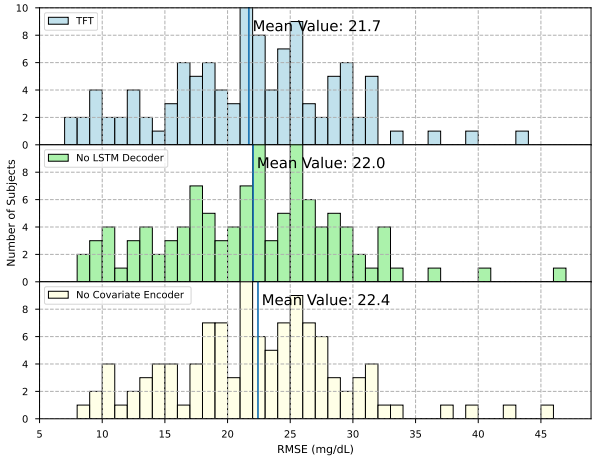
ization by conducting a five-fold cross-individual validation on the ShanghaiDM dataset, with stratification according to diabetes types. The model's performance on five distinct, previously unseen cohorts resulted in a mean RMSE of 14.7 and 23.5 mg/dL for 30 and 60-minute PHs, respectively. To further explore and improve upon cross-individual and cross-population prediction, our future work will incorporate domain generalization strategies, such as meta-learning [22], which are designed to mitigate domain shifts and enhance predictive accuracy across diverse training and evaluation datasets. While the present model was deployed on an external wristband, we intentionally selected a SoC analogous to those found in the majority of commercial CGM transmitters. Our future work involves a collaboration with hardware manufacturers to integrate the model directly into the CGM systems or insulin pumps, thereby offering on-device decision support. The decision support efficacy of our model and wearable device has been initially validated through a hardware-in-the-loop *in silico* trial, as detailed in Appendix B. To more comprehensively establish clinical utility and effectiveness, we are preparing to conduct real-world clinical trials, gather expert reviews, perform user studies, and apply decision matrix analysis. In managing potential missing CGM data, our wristband will incorporate a dual-step protocol. Initially, a real-time algorithm will monitor for CGM readings, triggering sound alerts if they are not received within a specified interval. Meanwhile, a retrospective algorithm will scan the historical data to assess sequences with partial data loss. For gaps of less than one hour, linear extrapolation will be utilized for imputation. For longer absences, the wristband will suspend predictions and issue both haptic and sound signals to alert the user, ensuring consistent CGM connection and prompt response to data discrepancies.

V. Conclusion

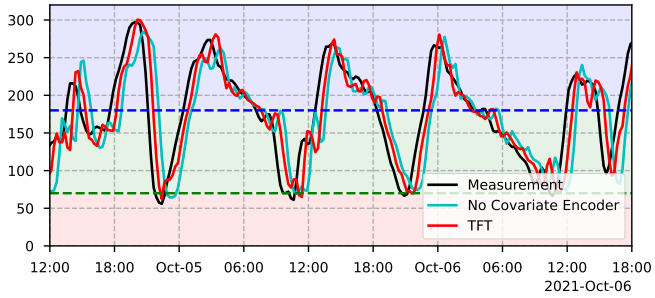
In this work, we propose EPS-TFT, a deep learning model to provide population-specific BG prediction for a diverse cohort of individuals living with diabetes. The model was developed and evaluated on multifaceted clinical datasets, encompassing both T1D and T2D subjects. When compared with five machine learning baseline methods, EPS-TFT achieved smallest RMSE, MAE, MAPE, and gRMSE for both 30 and 60-minute PHs, which stands as a promising tool in the accurate prediction of adverse glycemic events. When deploying EPS-TFT on our customized wearable device using edge computing, it consumes a mere 14.7 KB of RAM and 2 mW of power. This efficient setup allows for continuous decision support spanning from 51 days to 915 days on a single charge of the Li-Poly battery.

This research pioneers a population-level approach to BG prediction and has established a robust framework for the actual implementation of these models in digital health systems for real-world diabetes management.

Appendix



(a) RMSE distribution for the ShanghaiDM dataset



(b) Influence of the covariate encoder on prediction for a T1D individual

Fig. 7: Outcomes of the ablation study that excludes key submodules of the TFT model.

A. Ablation Study

In our ablation study, we evaluated the impact of excluding key submodules of the TFT model, including the covariate encoder and LSTM decoder. Fig. 7 presents the RMSE distribution for the ShanghaiDM dataset and illustrates the influence of the covariate encoder on a T1D individual's predictions. Notably, removing the covariate encoder results in degraded accuracy, particularly in critical scenarios such as hypoglycemia and hyperglycemia, highlighting its importance in our model's overall performance.

B. In Silico Trial

To assess the decision support capabilities of our wearable device, we executed a 3-month hardware-in-the-loop *in silico* trial, employing the UVA/Padova T1D simulator [56]. This trial incorporated 10 virtual adult subjects to account for intra- and inter-subject variability [57]. The simulator communicated CGM readings to the wearable, which then transmitted 60-minute BG predictions in a USB mode. Utilizing the predictive low-glucose management (PLGM) algorithms, the insulin pump suspended basal insulin delivery when predicted values were equal to or dropped below the hypoglycemia threshold of 70 mg/dL. Notably, when compared with open-loop systems, the decision support with the wearable device led to a significant reduction in the time spent below the glucose range (BG<70 mg/dL), decreasing from 5.3% to 1.9%, and a

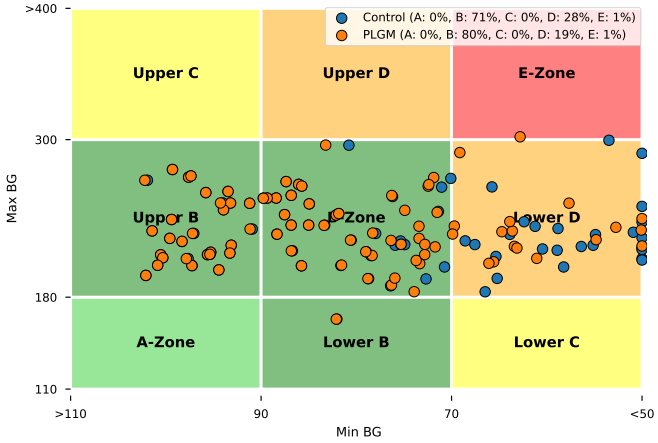


Fig. 8: Control-variability grid analysis to evaluate glucose regulation effectiveness, where each dot represents the daily minimum and maximum BG values.

modest improvement in the time within the target range (70-180 mg/dL), increasing from 74.6% to 75.2%. Fig. 8 illustrates the control-variability grid analysis results for a virtual adult. The comparative analysis suggested that the PLGM regimen resulted in a higher proportion of dots within the desirable A+B zone, increasing from 71% to 80%, and a 9% reduction in the D+E zone, indicating enhanced glucose control.

VI. Acknowledgement

This research was funded by Engineering and Physical Sciences Research Council (EPSRC EP/P00993X/1 and EPSRC EP/S021612/1). Taiyu Zhu was supported by President's Ph.D. Scholarship at Imperial College London.

References

[1] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya *et al.*, "IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, p. 109119, 2022.

[2] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B. J. Anderson, L. M. Jacobsen, D. A. Schatz, and Å. Lernmark, "Type 1 diabetes mellitus," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–17, 2017.

[3] M. Tauschmann and R. Hovorka, "Technology in the management of type 1 diabetes mellitus—current status and future prospects," *Nature Reviews Endocrinology*, vol. 14, no. 8, pp. 464–475, 2018.

[4] R. A. DeFronzo, E. Ferrannini, L. Groop, R. R. Henry, W. H. Herman, J. J. Holst, F. B. Hu, C. R. Kahn, I. Raz, G. I. Shulman *et al.*, "Type 2 diabetes mellitus," *Nature Reviews Disease Primers*, vol. 1, no. 1, pp. 1–22, 2015.

[5] J. C. Pickup, S. C. Freeman, and A. J. Sutton, "Glycaemic control in type 1 diabetes during real time continuous glucose monitoring compared with self monitoring of blood glucose: meta-analysis of randomised controlled trials using individual patient data," *BMJ*, vol. 343, 2011.

[6] 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.

[7] 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.

[8] R. W. Beck, T. D. Riddlesworth, K. Ruedy, A. Ahmann, S. Haller, D. Kruger, J. B. McGill, W. Polonsky, D. Price, S. Aronoff *et al.*, "Continuous glucose monitoring versus usual care in patients with type 2 diabetes receiving multiple daily insulin injections: a randomized trial," *Annals of Internal Medicine*, vol. 167, no. 6, pp. 365–374, 2017.

[9] 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.

[10] D. P. Zaharieva, K. Turksoy, S. M. McGaugh, R. Pooni, T. Vienneau, T. Ly, and M. C. Riddell, "Lag time remains with newer real-time continuous glucose monitoring technology during aerobic exercise in adults living with type 1 diabetes," *Diabetes Technology & Therapeutics*, vol. 21, no. 6, pp. 313–321, 2019.

[11] D. M. Maahs, P. Calhoun, B. A. Buckingham, H. P. Chase, I. Hramiak, J. Lum, F. Cameron, B. W. Bequette, T. Aye, T. Paul *et al.*, "A randomized trial of a home system to reduce nocturnal hypoglycemia in type 1 diabetes," *Diabetes Care*, vol. 37, no. 7, pp. 1885–1891, 2014.

[12] 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.

[13] 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.

[14] T. Zhu, L. Kuang, K. Li, J. Zeng, P. Herrero, and P. Georgiou, "Blood glucose prediction in type 1 diabetes using deep learning on the edge," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[15] M. F. Rabby, Y. Tu, M. I. Hossen, I. Lee, A. S. Maida, and X. Hei, "Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–15, 2021.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] 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.

[18] 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.

[19] M. He, W. Gu, Y. Kong, L. Zhang, C. J. Spanos, and K. M. Mosalam, "CausalBG: Causal recurrent neural network for the blood glucose inference with IoT platform," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 598–610, 2019.

[20] 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.

[21] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "LSTMs and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 706–712.

[22] 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.

[23] S. Langarica, M. Rodriguez-Fernandez, F. Nunez, and F. J. Doyle III, "A meta-learning approach to personalized blood glucose prediction in type 1 diabetes," *Control Engineering Practice*, vol. 135, p. 105498, 2023.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[25] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 2019, pp. 4171–4186.

[26] 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.

[27] 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," in *Proceedings of the 2020 Conference on Empirical methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[28] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with GPT-4," *CoRR*, vol. abs/2303.12712, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.12712

[29] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*, 2023, pp. 6778–6786.

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

[31] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 419–22 430, 2021.

[32] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The Eleventh International Conference on Learning Representations*, 2022.

[33] 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.

[34] A. K. Yetisen, J. Martinez-Hurtado, F. da Cruz Vasconcellos, M. E. Simsekler, M. S. Akram, and C. R. Lowe, "The regulation of mobile medical applications," *Lab on a Chip*, vol. 14, no. 5, pp. 833–840, 2014.

[35] A. Trifan, M. Oliveira, J. L. Oliveira *et al.*, "Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations," *JMIR mHealth and uHealth*, vol. 7, no. 8, p. e12649, 2019.

[36] P.-Y. Benhamou, S. Franc, Y. Reznik, C. Thivolet, P. Schaepelynck, E. Renard, B. Guerci, L. Chaillous, C. Lukas-Croisier, N. Jeandidier *et al.*, "Closed-loop insulin delivery in adults with type 1 diabetes in real-life conditions: a 12-week multicentre, open-label randomised controlled crossover trial," *The Lancet Digital Health*, vol. 1, no. 1, pp. e17–e25, 2019.

[37] M. Reddy, P. Herrero, M. E. Sharkawy, P. Pesl, N. Jugnee, D. Pavitt, I. F. Godsland, G. Alberti, C. Toumazou, D. G. Johnston *et al.*, "Metabolic control with the bio-inspired artificial pancreas in adults with type 1 diabetes: a 24-hour randomized controlled crossover study," *Journal of Diabetes Science and Technology*, vol. 10, no. 2, pp. 405–413, 2016.

[38] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 3, pp. 535–544, 2020.

[39] A. Burrello, D. J. Pagliari, M. Risso, S. Benatti, E. Macii, L. Benini, and M. Poncino, "Q-PPG: Energy-efficient PPG-based heart rate monitoring on wearable devices," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 6, pp. 1196–1209, 2021.

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

[41] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*, 2017, pp. 933–941.

[42] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, "A multi-horizon quantile recurrent forecaster," in *Time Series Workshop in Neural Information Processing Systems*, 2017.

[43] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *The 5th International Workshop on Knowledge Discovery in Healthcare Data in the 24th ECAI*, 2020, pp. 71–74.

[44] 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.

[45] 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.

[46] E. I. Georga, V. C. Protopappas, D. Ardigò, 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, 2013.

[47] J. Jeon, P. J. Leimbigler, G. Baruah, M. H. Li, Y. Fossat, and A. J. Whitehead, "Predicting glycaemia in type 1 diabetes patients: experiments in feature engineering and data imputation," *Journal of Healthcare Informatics Research*, vol. 4, no. 1, pp. 71–90, 2020.

[48] S. Kriventsov, A. Lindsey, A. Hayeri *et al.*, "The diabits app for smartphone-assisted predictive monitoring of glycemia in patients with diabetes: retrospective observational study," *JMIR Diabetes*, vol. 5, no. 3, p. e18660, 2020.

[49] 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, ECAI*, 2020, pp. 105–109.

[50] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski, "N-HiTS: Neural hierarchical interpolation for time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, 2023, pp. 6989–6997.

[51] S. Del Favero, A. Facchinetti, and C. Cobelli, "A glucose-specific metric to assess predictors and identify models," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1281–1290, 2012.

[52] 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.

[53] 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.

[54] N. Ehrhardt and I. B. Hirsch, "The impact of COVID-19 on cgm use in the hospital." *Diabetes Care*, vol. 43, no. 11, pp. 2628–2630, 2020.

[55] A. L. Fortmann, S. R. Spierling Bagsic, L. Talavera, I. M. Garcia, H. Sandoval, A. Hottinger, and A. Philis-Tsimikas, "Glucose as the fifth vital sign: a randomized controlled trial of continuous glucose monitoring in a non-ICU hospital setting," *Diabetes Care*, vol. 43, no. 11, pp. 2873–2877, 2020.

[56] C. Dalla Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator: new features," *Journal of Diabetes Science and Technology*, vol. 8, no. 1, pp. 26–34, 2014.

[57] P. Herrero, J. Bondia, O. Adewuyi, P. Pesl, M. El-Sharkawy, M. Reddy, C. Toumazou, N. Oliver, and P. Georgiou, "Enhancing automatic closed-loop glucose control in type 1 diabetes with an adaptive meal bolus calculator–in silico evaluation under intra-day variability," *Computer Methods and Programs in Biomedicine*, vol. 146, pp. 125–131, 2017.

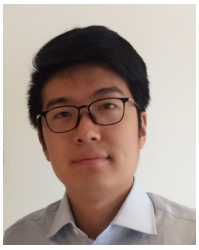
Taiyu Zhu (Member, IEEE) received the B.Eng. (Hons.) degree from the Australian National University, Canberra, Australia in 2017. He received the M.Sc. degree in 2018 and the Ph.D. degree in 2022, both in Electrical and Electronic Engineering, from Imperial College London, London, U.K. His doctoral research, supported by President's Ph.D. Scholarship at Imperial College London, focused on deep learning in diabetes management.

He is currently a Research Fellow with Department of Psychiatry, University of Oxford. His research, funded by the Novo Nordisk - Oxford Postdoctoral Research Fellowship, is pioneering deep learning technologies alongside the UK Biobank data to further improve the health outcomes for individuals with cardiometabolic conditions. His main research interest lies in advanced artificial intelligence technologies and their applications to healthcare.

Lei Kuang (Student Member, IEEE) received the B.Sc. degree in industrial electronics and control engineering from the Liverpool John Moores University in 2016, and M.Sc. degrees in embedded systems, analog and digital integrated circuit design from the University of Southampton and Imperial College London in 2017 and 2019 respectively. He is currently a PhD student at the Centre for Bio-Inspired Technology, developing comprehensive biomedical devices for the point-of-care diagnosis of infectious diseases. His research interests include high-speed Lab-on-Chip platform, compression and machine learning algorithms for CMOS imagers, digital IC design and real-time processing system for biomedical applications.

Chengzhe Piao (Student Member, IEEE) is currently pursuing Ph.D. studies at the Institute of Health Informatics, University College London, supervised by Dr. Ken (Kezhi) Li. He completed his M.Sc. and B.Sc. at the Beijing Institute of Technology. His research focuses on AI in healthcare, deep learning, reinforcement learning, and federated learning. Piao has contributed to various publications, including the IEEE Internet of Things Journal and IEEE ICDE.

Junming Zeng (Member, IEEE) received the Bachelor degree in Electronic Engineering from the University of Southampton, UK in 2016, and the Master and PhD degree from Imperial College London, UK in 2017 and 2022. He is currently a research associate with Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London UK. His research interests include Analogue and Mixed Signal IC design for CMOS Lab-on-Chip ion imaging platform. He is the recipient of the Department PhD Scholarship from Imperial College London, and the awardee of the Best Student Paper Award 1st Prize at ISCAS 2018.



Kezhi (Kenneth) Li (Member, IEEE) is a Lecturer (Assistant Professor) at Institute of Health Informatics (IHI), University College London (UCL). He received the PhD degree at Imperial College London (ICL) and B.Eng. degree at University of Science and Technology of China (USTC). His research interests lie in biomedical signal processing, machine learning and their applications in healthcare. Prior to joining UCL, he was a senior research associate at ICL, University of Cambridge, a research fellow at Royal Institute of Technology (KTH) in Stockholm and a research assistant at Microsoft Research Asia (MSRA) and USTC. He was the recipient of several best paper awards, including the best paper in WNIP workshop of Neurips 2017 and the winner of BGLP Challenge at IJCAI-ECAI 2018.

Pantelis Georgiou (Senior Member, IEEE) received the M.Eng. degree in electrical and electronic engineering and the Ph.D. degree from Imperial College London (ICL), London, U.K., in 2004 and 2008, respectively.

He is currently a Professor of Biomedical Electronics with the Department of Electrical and Electronic Engineering, ICL, where he is also the Head of the Bio- Inspired Metabolic Technology Laboratory, Centre for Bio-Inspired Technology. His research includes bio-inspired circuits and systems, CMOS based Lab-on-Chip technologies, and application of microelectronic technology to create novel medical devices. He has made significant contributions to integrated chemical-sensing systems in CMOS, conducting pioneering work on the development of ISFET sensors, which has enabled applications, such as point-of-care diagnostics and semiconductor genetic sequencing and has also developed the first bio-inspired artificial pancreas for treatment of Type I diabetes using the silicon-beta cell. He received the IET Mike Sergeant Medal of Outstanding Contribution to Engineering in 2013. In 2017, he was also awarded the IEEE Sensors Council Technical Achievement award. He is a member of the IET and serves on the BioCAS and Sensory Systems technical committees of the IEEE CAS Society. He is also on the IEEE Sensors council as a member at large and an IEEE Distinguished Lecturer. He is Co-founder of ProtonDx, commercialising technologies for rapid diagnostics for infectious diseases.