IoMT-Enabled Real-time Blood Glucose Prediction with Deep Learning and Edge Computing

Taiyu Zhu, *Student Member, IEEE*, Lei Kuang, *Student Member, IEEE*, John Daniels, *Student Member, IEEE*, Pau Herrero, *Member, IEEE*, Kezhi Li, *Member, IEEE*, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*—Blood glucose (BG) prediction is essential to the success of glycemic control in type 1 diabetes (T1D) management. Empowered by the recent development of the Internet of Medical Things (IoMT), continuous glucose monitoring (CGM) and deep learning technologies have been demonstrated to achieve the state of the art in BG prediction. However, it is challenging to implement such algorithms in actual clinical settings to provide persistent decision support due to the high demand for computational resources, while smartphone-based implementations are limited by short battery life and require users to carry the device. In this work, we propose a new deep learning model using an attention-based evidential recurrent neural network and design an IoMT-enabled wearable device to implement the embedded model, which comprises a low-cost and low-power system on a chip to perform Bluetooth connectivity and edge computing for real-time BG prediction and predictive hypoglycemia detection. In addition, we developed a smartphone app to visualize BG trajectories and predictions, and desktop and cloud platforms to backup data and fine-tune models. The embedded model was evaluated on three clinical datasets including 47 T1D subjects. The proposed model achieved superior performance of root mean square error (RMSE), mean absolute error, and glucose-specific RMSE, and obtained the best accuracy for hypoglycemia detection when compared with a group of machine learning baseline methods. Moreover, we performed hardware-in-the-loop *in silico* trials with 10 virtual T1D adults to test the whole IoMT system with predictive low-glucose management, which significantly reduced hypoglycemia and improved BG control.**

***Index Terms*—Diabetes, deep learning, Internet of Things (IoT), edge computing, glucose prediction, artificial intelligence.**

I. Introduction

DIABETES is a chronic disease characterized by hyperglycemia, which affects just under half a billion people worldwide [1]. Due to autoimmune destruction of pancreatic β-cell resulting in an absolute insulin deficiency, people living with type 1 diabetes (T1D) require lifelong management to maintain the blood glucose (BG) levels in a safe range. To do so, they need to consistently adhere to a series of self-care behaviors, such as monitoring BG levels, administrating exogenous insulin, and carefully scheduling meals and exercise. Otherwise, the risk of hypoglycemia and hyperglycemia would increase, which may lead to various short and long-term complications. Hyperglycemia is a major responsible factor for the development of nephropathy, retinopathy, coronary heart diseases [2], but severe hypoglycemia is more dangerous and may cause coma, seizures, or even death [3]. In this regard, BG prediction is a crucial tool in T1D management to improve glycemic control, which allows for proactive interventions to reduce, or even prevent adverse glycemic events and diabetic complications. However, due to high inter- and, in the long term, intra-subject variability, developing an accurate BG prediction model is still challenging [4].

With the rapid development of the Internet of things (IoT), recent advances in continuous glucose monitoring (CGM) have been shown to enhance the treatment for people with T1D [5]. A CGM system comprises an implanted sensor to measure interstitial BG levels and a transmitter to send measurements to a receiver, such as a customized hardware box, smartphone, or smart watch with a fixed frequency (e.g., every five minutes). As a well-established paradigm of the Internet of Medical Things (IoMT) [6], CGM can also be combined with insulin pumps as sensor-augmented therapy, i.e., artificial pancreas (AP). In this context, BG prediction can be used in closed-loop AP systems with model predictive control [7] and enables predictive low-glucose management (PLGM) systems that have been proved to be effective for reducing hypoglycemia in clinical settings [8].

The widespread use of CGM has produced a large amount of data that offers the promise of developing artificial intelligence (AI) technologies in BG prediction, especially for machine learning algorithms [9]. In particular, deep learning-based models have recently achieved the state of the art in terms of accuracy [10]–[13]. Of note, by employing the latest deep learning technologies, the increasingly complex models rely on a huge number of parameters, neurons, and layers for model inference. Thus, how to implement these models in actual clinical settings to bring actual therapeutic benefits is under-researched, which can be problematic since on-device inference with a large number of model parameters requires intensive computational resources and memory consumption.

The existing methods to implement deep learning models for BG prediction are mainly based on customized smartphone apps [14]–[16]. However, several limitations exist in these methods including lack of wearability, battery constraints, and the dependency on mobile operating systems. It is inconvenient for T1D users to carry smartphones or other handheld devices

This work was supported by EPSRC EP/P00993X/1 and President's Ph.D. Scholarship at Imperial College London. T. Zhu and L. Kuang have equal contribution. *(Corresponding author: K. Li)*

T. Zhu, L. Kuang, J. Daniels, P. Herrero, P. Georgiou are with Centre for Bio-inspired Technology, Imperial College London, London, United Kingdom. (e-mail: {taiyu.zhu17, lei.kuang18, john.daniels11, pherrero, 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).

all the time, especially during high-intensity activities which would reduce the awareness of subsequent hypoglycemia in T1D [17]. In addition, the battery level of smartphones and smartwatches are significantly drained because the prediction algorithms continuously run in the background with Bluetooth connectivity [18]. As a result, the decision support system will be unavailable when the devices run out of power. Moreover, the smartphone implementation is highly dependent on mobile operating systems, such as Android and iOS, and deep learning libraries, such as PyTorch [14] and TensorFlow Lite [15], [16]. Many existing apps for diabetes management suffer from the frequent updates of mobile operating systems. T1D users, especially the elderly population, may need to purchase extra expensive smartphones if the implementation does not support their own devices. Cloud implementation could be a solution to this problem, but it is largely limited by Internet connectivity since there are many daily scenarios suffering from poor coverage of WiFi and mobile signals. Thus, a power-efficient and low-cost wearable device based on edge computing [19]–[21] is preferred in T1D management to provide real-time BG prediction and predictive hypoglycemia detection. The outcomes of this study also indicate the possibility of embedding deep learning algorithms into CGM devices (e.g., wearable transmitters).

In this work, we propose a new deep learning algorithm and develop a novel IoMT-enabled wearable device to implement the algorithm using a system on a chip (SoC) for Bluetooth low energy (BLE) connectivity and edge computing. In particular, a computationally efficient recurrent neural network (RNN) with the attention mechanism is introduced to obtain accurate BG predictions. We employ evidential regression to compute model uncertainty and improve the detection of impending hypoglycemia. Then the well-trained model was embedded into the SoC of the customized wearable device with an optimized circuitry to minimize energy consumption. Receiving the measurements from CGM, the wearable device performs real-time model inference to obtain BG predictions and hypoglycemia warning for decision support, which can be further integrated into AP systems. Finally, we evaluated the prediction accuracy of the embedded model, analyzed the power and edge computing performance, and tested the efficacy of the wearable device in the simulation of 10 virtual T1D adults with the FDA-accepted UVA/Padova T1D simulator [22]. The original contributions of this work can be summarized as follows.

- We propose a new attention-based lightweight RNN for real-time BG prediction and hypoglycemia detection with CGM input data on edge devices.
- We design an IoMT-enabled wearable device with a low-cost and power-efficient SoC that communicates with CGM and other devices of T1D management through BLE and performs edge computing for the model inference of the embedded deep learning algorithm. A cloud platform is developed for model training and data backup.
- The embedded model is evaluated by three clinical datasets and compared against a variety of machine learning and deep learning baseline methods.
- We analyze the power and memory footprint of the wearable device and perform *in silico* trials to validate the therapeutic efficacy of the PLGM system integrated with the wearable device.

The remainder of this paper is organized as follows. We first present an overview of related work in Section II. The details of the system design including the software and hardware are illustrated in Section III. In Section IV, we describe the clinical datasets and analyze the experiment results on clinical datasets and *in silico* trials. Finally, we conclude this article and discuss the future work in Section V.

II. Related Work

A. BG Prediction with Machine Learning and CGM Data

The forecasting of BG levels over short and long-term prediction horizons (PHs) plays an important role in T1D management. In general, the prediction algorithms reported in the literature can be mainly divided into physiological modeling, data-driven approaches, and hybrid methods [9]. However, due to the larger inter-subject variability, it is difficult to develop a generic physiological model that has proper parameter settings for each personal profile. Fortunately, with an increasing amount of CGM data, machine learning approaches have been shown to achieve superior prediction accuracy [9]. In this regard, a common strategy is to treat BG prediction as a supervised learning task that uses continuous CGM sequences and other relevant features (e.g., daily activities) as model input and future BG levels as the corresponding targets. Conventional machine learning solution to this task include the autoregressive integrated moving average (ARIMA) [23], random forests [24], artificial neural networks [25], and support vector machine (SVR) [26].

Particularly, empowered by various architectures of deep neural networks (DNNs), deep learning-based models have obtained superior performance on BG prediction and outperformed conventional machine learning baseline methods in recent studies. Instead of merely using a feed-forward structure, RNNs fetch the output at previous timesteps as a part of current input, which makes it a powerful tool in sequence processing and regression tasks. In addition, the long short-term memory (LSTM) and gated recurrent unit (GRU) are two classic RNN cells that solve the issues of gradient vanishing and exploding of vanilla RNNs [27], which have been widely applied in previous work on BG prediction. Martinsson *et al.* [28] proposed an LSTM-based model to learn physiological patterns of BG dynamics only using CGM input. In [29], a bidirectional LSTM (Bi-LSTM) was used to predict BG concentration and outperformed an ARIMA baseline.

In addition, temporal convolutional networks (TCNs) based on convolution neural networks (CNNs) and causal convolutions are comparable to RNNs in sequence modelling [30], [31]. Li *et al.* [15] proposed a TCN-based model to classify the BG changes between current and future values. A convolutional recurrent neural network (CRNN) was proposed in [16], which used CNN layers to extract feature maps and LSTM to obtain final predictive BG levels.

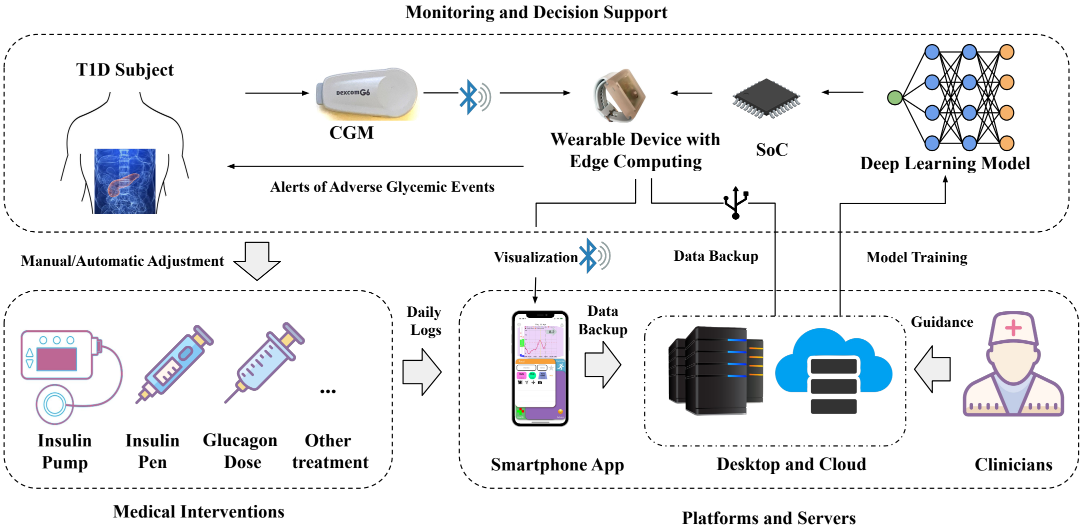


Fig. 1. System architecture of the T1D management system with the proposed wearable device, which contains three subsystems as follows: monitoring and decision support, medical interventions, and platforms and servers. The wearable device is a part of the monitoring and decision support subsystem that can provide real-time measurement, BG prediction, and hypoglycemia warning.

To further improve the prediction accuracy, a group of the latest advances has been applied to improve BG prediction but increased computational complexity, such as dilated connections [32], attention mechanism [33], ensemble learning [13], multi-task learning [14], and residual connections [15].

Although previous studies incorporated physiological measurements and daily activities as model input, such as carbohydrate intake [14], insulin injection [15], and exercise levels [26], prediction with CGM data only is a practical and valuable option in real-world scenarios [34]. On one hand, some physiological features require extra wearable devices (e.g., insulin pumps and wristbands) that are not widespread in T1D management systems [35] and would introduce artifact errors due to hardware issues, such as signal loss and drained battery. On the other hand, the manual data entries are burdensome and likely to cause human errors (e.g. missing meals).

B. IoMT Systems in Healthcare and T1D Management

The IoMT is defined as the connectivity of numerous medical devices to healthcare systems and care providers. Integrated with a variety of physiological sensors, communication modules, and recent AI technologies, IoMT provides significant clinical benefits and has the potential to have a major impact on the healthcare domain [36]. Firstly, the proliferation of wireless sensors and personal wearable devices enables IoMT to develop efficient continuous and remote monitoring systems for healthcare infrastructures. For instance, Catarinucci *et al.* [37] proposed the architecture of a smart hospital system with a specifically designed wireless sensor network, aiming to automatically monitor and track people and medical devices within hospitals. In [38], an IoMT-enabled low-power wearable system was developed to address the needs of long-term remote electrocardiogram monitoring. Secondly, IoMT offers promising solutions to enhance self-care and early diagnosis of various diseases [39]. Su *et al.* [40] integrated IoMT technologies and deep learning algorithms into a screening system to assess characteristic signals of patients with valvular heart disease. Similarly, Tuli *et al.* [41] combined IoMT, fog computing, and ensemble deep learning to develop an automatic system for heart disease analysis. In [39], the authors proposed SPHERE, an IoMT system to improve the wellbeing of the elderly population with chronic diseases. In recent work, IoMT was also employed with 5G cloud computing and deep learning for telemedicine diagnosis of epidemic diseases [42].

In particular, IoMT has opened a door to efficient and reliable BG monitoring and glycemic control to improve diabetes management [43], [44], leveraging various wearable devices and interconnections in AP systems, such as CGM, insulin pumps, insulin pens, glucagon pumps, and physiological wristbands for measuring signals (Fig. 1). In [45], the authors presented a smart diabetic healthcare system with the hardware implementation of a development board and a microcontroller unit (MCU) to control an insulin pump and transfer health records to cloud storage. They applied a hash algorithm to provide authenticity for individual data and improve the security of the IoMT system. Moreover, Herrero *et al.* [46] proposed the Bio-inspired Artificial Pancreas which comprises a customized handheld unit to implement glycemic control algorithms on an MCU and communicate with CGM, insulin pump, and a dedicated smartphone app via Bluetooth. The cloud services were provided in the app for remote monitoring. This system was demonstrated by the UVA/Padova T1D simulator and further validated in a clinical trial [47]. Similar IoMT-enabled AP systems with cloud services have been reported in the literature, such as the Bionic pancreas [48] and DiAs system [49]. Considering various security issues existing in implantable sensors and wireless interconnections, Astillo *et al.* [50] developed a misbehavior detection system to assess the trustworthiness of the wearable devices in AP systems, including CGM, controllers, and insulin pumps, and also evaluated the system in the UVA/Padova T1D simulator.

C. Edge AI in IoMT

Most existing solutions to implement decision-support algorithms and interact with wearable devices in IoMT-based T1D management systems would introduce several essential challenges, such as the battery limit of smartphone platforms, lack of wearability for customized handheld devices, and high-latency decision making with cloud platforms. Edge AI [51], as known as edge intelligence, is a promising solution to tackle these challenges, which allows edge computing to execute AI algorithms, e.g., deep learning. This technique is in an early stage [51] and emerging in recent research of IoMT [52]. In [53], the authors proposed a framework of edge computing and machine learning to predict early warning scores with vital signs, aiming at providing decision support for critical care interventions. Kong *et al*. [54] proposed a CNN-based deep learning model to detect mask-wearing to help prevent infection of COVID-19 with the edge implementation on the Intel Neural Compute Stick and Raspberry Pi 4. Olokodana *et al*. [55] introduced a machine learning model, ordinary kriging, to detect seizures with an edge device of Raspberry Pi 3B+.

However, there remain several challenges for edge computing to be widely adopted in IoMT healthcare systems [56]. Due to the computational constraints and memory limit, it is challenging to deploy complex algorithms with a high number of parameters and variables, e.g., deep learning models. The daily use of personal medical devices generates a considerable amount of data, which puts a huge burden on the storage of edge devices. Meanwhile, protecting the security and privacy of personal health data is an important consideration in data transmission and task offloading between IoMT devices [57]. Edge devices are also vulnerable to malicious attacks, which require systematic security approaches, such as trust management and defense mechanisms, to ensure trustworthiness in IoMT systems. Finally, due to the limited capacity of batteries, power management is essential for edge devices to consistently provide high-quality services and prevent any signal or data loss, especially for the devices that provide continuous long-term decision support in clinical settings.

III. System Design

In this section, we present the details of the framework to develop the proposed deep learning model for BG prediction and the corresponding implementation of the embedded system in the customized wearable device.

A. Framework Overview

Fig. 1 depicts an overview of the proposed system architecture in T1D management. There are three subsystems: 1) monitoring and decision support, 2) medical interventions, 3) platforms and servers, which are described in the subsequent sections. The IoMT-enabled wearable device is in the center of the monitoring and decision support system. Communicating with the CGM via Bluetooth connectivity, the wearable device empowers a T1D user with real-time BG prediction and hypoglycemia detection. Then the user can interact with the subsystem of medical interventions to adjust treatment. The data transmission between the wearable device and the platforms and servers aims at data visualization, data backup, and updating the embedded deep learning model.

1) Monitoring and Decision Support: As the core component of the proposed system, it contains a CGM sensor that measures BG levels every five minutes and transmits the real-time measurements to a specifically designed wearable wristband via BLE. The SoC of the wearable device performs the embedded deep learning algorithm to predict BG levels and detect forthcoming hypoglycemic events. The historical CGM measurements and DNN weights are stored in the Flash memory, which can be accessed and updated by the platforms and servers. This essential subsystem can run solely without interactions with other devices to guarantee persistent and reliable decision support throughout day and night. In addition, thanks to a power-efficient design of SoC, the battery life of the wearable device (six months) is longer than that of the CGM sensor (10 days) and transmitter (three months).

2) Medical Interventions: Automatic control with the same SoC that enables Bluetooth communication with insulin pumps has been validated in our previous work [46]. In this work, we consider manual control to fit different clinical scenarios since insulin pumps are not widely used by people with T1D. Once receiving predictions and warnings from the wearable device, a T1D subject is allowed to seek necessary interventions in advance and manually adjust existing medical treatment.

3) Platforms and Servers: A smartphone app can connect with the wristband through Bluetooth to visualize current CGM reading, predictions, and historical BG trajectories, while recording daily activities, such as meals, excise, and health conditions. A desktop platform with a specifically designed graphical user interface (GUI) (Fig. 8 in the Appendix) is employed to train the deep learning models and backup collected data. It communicates with the wristband through USB ports and can upload data to the Amazon cloud storage, i.e, a bucket of AWS S3. T1D users are allowed to perform these operations by themselves or with the guidance of healthcare providers or clinicians if needed. To facilitate users without a programming background, we deploy the deep learning models in the cloud using AWS SageMaker. Thus, the models can be automatically trained with newly uploaded data on the cloud platform and downloaded from the cloud storage to the wearable device.

B. Problem Formulation and Feature Engineering

Denoting a BG level measured by CGM at timestep t as G_t, the target of prediction is to estimate a future BG value of G_{t+p}, where p is a PH (e.g., 30 minutes) normalized by resolution of CGM. To extract hidden representations, the input data contains a sequence of retrospective data $\mathbf{X_t}$ with a length of L, i.e., $\mathbf{X}_t = [\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-\Delta}] \in \mathbb{R}^{d \times L}$, where d is the dimension of the input features; $\mathbf{x}_t \in \mathbb{R}^{d \times 1}$ denotes the input vector at the timestep t; and $\Delta = L-1$. Considering edge computing that delivers computation close to data sources, we derive all the input features from CGM measurements and corresponding timestamps in the monitoring and decision support system (Fig. 1). The timestamps for a 24-hour period are converted into two types of time index to map seasonal patterns: min-max normalization with a range of $[0, 1]$ [32] and sine-cosine encoding [13]. The BG change over the PH is used as the learning target y_t to reduce underlying bias [15], [32], i.e., $y_t = f_n(G_{t+p} - G_t)$, where f_n is the min-max normalization to scale each feature.

Combining CGM sequences with time index, we perform feature selection during the validation phase. The best validation performance was obtained with the CGM time series $\mathbf{G}_t$ and min-max normalized timestamps $\mathbf{S}_t$, i.e., $\mathbf{X}_t = f_n([\mathbf{G}_t; \mathbf{S}_t])$. However, we notice that there is a large number of missing gaps in the historical CGM measurements, due to some inevitable reasons (e.g., sensor calibration and signal loss), which account for around 10% of the total length. Thus, we interpolate the missing CGM data in the middle of input sequences and extrapolates the missing CGM data at the tail to avoid involving future information in current predictions.

C. BG Prediction by Evidential RNN Models

Although RNN-based models have exhibited superior performance in BG prediction, a challenge of implementing such models in actual clinical settings is the lack of evaluating the uncertainty and confidence of predictions. It is essential to determine whether a prediction is reliable and confident when a deep learning model aims to provide critical decision support in a healthcare system. In the context of T1D management, a lower bound of prediction value is a useful indicator, as low glucose episodes, i.e., hypoglycemia, may lead to life-threatening events.

To this end, we propose an embedded edge evidential neural network (E3NN) model to compute the lower bounds (LBs) of each prediction. Fig. 2 shows the architecture of the proposed deep learning model, consisting of a base model with a stack of RNN layers, an attention layer, a dropout layer, a dense layer, and an evidential output layer. The input of E3NN is a multivariate time series with CGM and timestamps, while the output comprises the parameters of the evidential distribution to compute prediction values and LBs. The GRU cells are employed rather than LSTM because they achieved better validation performance with a smaller number of parameters [32]. The cell operations are denoted as

$$\begin{aligned}
\mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r), \\
\mathbf{z}_t &= \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z), \\
\hat{\mathbf{h}}_t &= \sigma(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h \mathbf{r}_t \odot \mathbf{h}' + \mathbf{b}_h), \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \hat{\mathbf{h}}_t, \qquad (1)
\end{aligned}$$

where $\mathbf{r}_t$, $\mathbf{z}_t$, $\hat{\mathbf{h}}_t$, and $\mathbf{h}_t$ denote reset gate vector, update gate vector, candidate activation, and cell output, respectively; $[\mathbf{W}_r, \mathbf{U}_r, \mathbf{b}_r]$, $[\mathbf{W}_z, \mathbf{U}_z, \mathbf{b}_z]$, and $[\mathbf{W}_h, \mathbf{U}_h, \mathbf{b}_h]$ denote the set of input weights, cell output weights, bias for reset gate, update gate, candidate activation, respectively. Dropout layers are used to prevent DNNs from overfitting and improve model generalization.

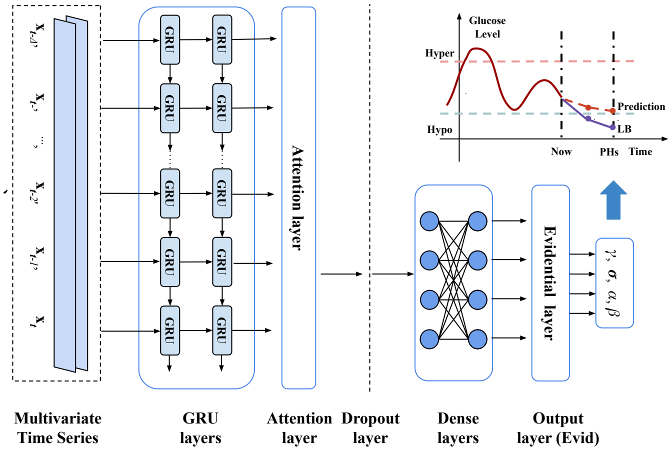


Fig. 2. Block diagram of the proposed E3NN. The model input is a multivariate time series. The output of E3NN includes the four parameters $p[\gamma, \sigma, \alpha, \beta]$ of the posterior distribution to compute BG predictions with corresponding LBs.

The output of the second GRU layer is fed into an attention layer to obtain a weighted context vector $\mathbf{c}_t$ as follows

$$\mathbf{c}_t = \sum_{i=t-L+1}^{t} a_{t,i} \mathbf{h}_i, \qquad (2)$$

where $a_{t,i}$ and $\mathbf{h}_i$ are attention weight and the hidden state at the i-th timestep, respectively. The attention weights are derived by alignment scores that indicate the relationship between the current cell output and retrospective information at previous timesteps. In the experiments, we explored a group of alignment functions, including additive [58], general [59], dot product [59], and location-based attention [59]. Here we use the general form, considering it achieved the largest improvement of the validation performance, which can be defined as follows

$$a_{t,i} = \frac{\exp(\mathbf{h}_i \mathbf{W}_a \mathbf{h}_t)}{\sum_{i=t+1-L}^{t} \exp(\mathbf{h}_i \mathbf{W}_a \mathbf{h}_t)}, \qquad (3)$$

where $\mathbf{W}_a$ computes the alignment scores, which is parametrized by a feed-forward network; and the attention weights are normalized by the Softmax function. The output of the attention layer is processed by a dense layer with ReLU activation to extract high-level features $\mathbf{h}_t^d$, which can be expressed as

$$\mathbf{h}_t^d = \mathtt{ReLU}(\mathbf{W}_d \mathbf{c}_t + \mathbf{b}_d) \qquad (4)$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are the weights and bias of the dense layer.

To compute model uncertainty and corresponding LBs, we assume the observed prediction targets are drawn from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown mean and variance, i.e., $\mu \sim \mathcal{N}(\gamma, \sigma^2/\lambda)$, $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, where Γ stands for the gamma function. We can estimate posterior distribution with an approximation of the Normal Inverse-Gamma (NIG) distribution with four parameters of $\gamma, \sigma, \alpha, \beta$ [60]. In this case, the objective of the deep learning model is obtaining the parameters of the NIG distribution rather than a single prediction value. Hence, a dense layer with four-dimensional output is used as the evidential layer (Evid) to compute these parameters as the final output, i,e., $\gamma, \sigma, \alpha, \beta = \texttt{Evid}(\mathbf{h}_t^d)$.

According to Bayesian probability theory, the likelihood of an observed prediction target can be obtained by applying marginalization to the parameters of the Gaussian distribution (μ, σ). It has been proven that, given the assumption of the NIG approximation, the Gaussian likelihood function can be solved by a form of the generalized Student-t distribution (St) [60]. Therefore, the model is trained by a negative log likelihood loss to fit the observations with uncertainty, which is defined as follows

$$\mathcal{L}_t = -\log(\texttt{St}(y_t|2\alpha, \gamma, \sqrt{\frac{\beta(1+\lambda)}{\lambda\alpha}})) \tag{5}$$

where 2α, γ, $\sqrt{\frac{\beta(1+\lambda)}{\lambda\alpha}}$ are the degrees of freedom, location parameter, scale parameter of the Student-t distribution. The prediction $\hat{y}_t$ and lower bound LB are derived as follows

$$\hat{y}_t = \gamma, \quad LB = \hat{y}_t - k\sqrt{\frac{\beta}{\lambda(\alpha-1)}}. \tag{6}$$

where k is a personalized hyperparameter to adjust the LBs for hypoglycemia detection, which is determined in the validation phase for each subject. Processed by the inverse function of the feature normalization, predictive BG levels can be restored by adding the predictive BG changes to the current BG levels.

D. Edge Computing

Compared with model implementation on the cloud, edge computing can offer more reliable real-time services on the wearable device with extremely low latency of decision making, which are not limited by Internet connectivity. Deep learning with edge inference is emerging research in the fast-growing areas of AI and IoT. Existing inference frameworks, such as TensorFlow Lite Micro [61] and CMSIS-NN [62], currently support a limited subset of operations and DNN layers. Therefore, we convert the E3NN TensorFlow models to C models based on the CMSIS-DSP library that offers high-performance APIs for math functions such as matrix operations, and the firmware development is based on the latest nRF5 SDK v17.0.2.

The BLE SoC is based on an ARM Cortex-M4 Core with a tight memory budget (512 KB Flash and 64 KB SRAM). However, the SoC not only communicates with the front-end CGM transmitter through the BLE protocol but also runs the trained RNN model on the edge. Thus, we optimize the SRAM usage of the model inference. In particular, assuming the input of the embedded model involves L CGM readouts associated with timestamps, the first GRU layer processes a two-dimensional data sample at each timestep and repeats for L rounds. Each round of operations is dependent on the output from the previous iteration and cannot be computed in parallel for acceleration. However, as the second GRU layer runs in the same way except for using the output of the first GRU layer as input data, the operations of these two GRU layers can be pipelined. Thus, the cells of the stacked RNN layers at the same timestep are performed in one round and iterated L times. This interleaving process reduces the SRAM utilization since only one output needs to be temporally stored instead of an output vector with a length of L. The dropout layer only applies in the training phase, which is disabled in model inference and thus not implemented on the SoC.

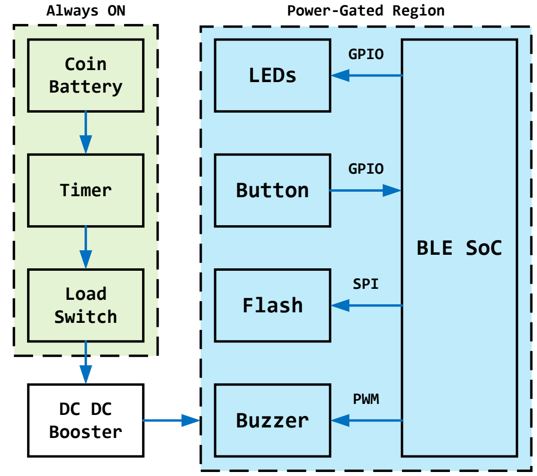


Fig. 3. Block diagram of the proposed IoMT system in T1D management, which is powered by a coin battery and embedded with LEDs, Buttons, and a Buzzer for user interactions as well as a NOR Flash for data storage. The system employs a hardware power-gating circuit including a timer and a load switch for energy saving, which can maintain ultra-low-power during the idle period.

We import RNN weights as 4-byte hex data as a raw format representing 32-bit floating-point numbers, aiming to maintain the prediction accuracy with less loss of precision compared with post-training quantization. These weights are fixed and stored in the Flash memory, which can be claimed as static constant variables.

E. Embedded System and Wearable Device Design

To meet the requirements raised for edge computation, the proposed system involves a lightweight and compact hardware design for a low-power and low-cost wearable device. It is embedded with four main peripherals including the LEDs, button and buzzer for essential user interactions as shown in Fig. 3, where a Nordic SoC (nRF52832) is employed as the system controller. The entire system can be divided into two parts. The first part is a hardware power-gating circuit that includes a timer and a load switch to control the on/off state of the system, while the second part inside the power-gated region aims to save energy during the idle period.

Once a prediction is made and an adverse BG event is detected, the user can be notified through either the light or sound, generated by the LED and buzzer, whereas a simple click on the button can stop the notification. In addition, to backup the CGM readouts for post-processing, a NOR Flash is employed that provides 16 MB memory capacity. Considering that each data sample transmitted by CGM every five minutes only contains 16 bytes, such Flash memory space can support long-term data storage for more than one year. The only drawback is that the Flash-type memory does not support random access, thus the writable address must be determined

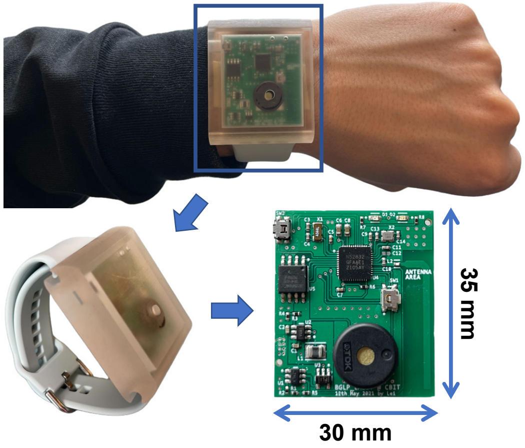


Fig. 4. IoMT-enabled wearable device consists of a PCB designed in a dimension of 35 mm x 30 mm and a transparent case manufactured by 3D printing.

at the start point. To solve this, a binary search algorithm is implemented on the SoC which significantly improves the efficiency compared with searching exhaustively.

The BLE SoC can enter a soft power-down mode for energy saving before starting the next CGM readout. However, its peripherals can still consume some power if they are connected to the main supply. As a result of this, the power gating technique is applied in the system to shut off the current to the BLE SoC and its peripherals during the idle period. This is realized through a timer integrated circuit which generates a periodic power-gated signal to control a load switch. In addition, due to the need for user notifications, such a process may take a different amount of time. Because of this, an extra signal driven by the BLE SoC is connected to the timer integrated circuit to enter the shutdown mode.

Due to the compactness of the proposed system, the finial hardware is populated onto a 2-layer printed circuit board (PCB) in a dimension of 35 mm x 30 mm as shown in Fig. 4. The PCB is inside a 3D printed case with a transparent appearance. The button for users to confirm hypoglycemic events is located at the top-left edge. The black cylinder that occupies a large area of the PCB is a buzzer. The size of the wearable device is close to a smartwatch (e.g., Apple Watch). It can be powered by a single coin battery (CR2302) with a lifespan of six months.

IV. Experiments

In this section, we first describe the clinical data used in this work and the process of model development. Then we present the performance of the proposed system, including the prediction accuracy of BG levels and hypoglycemia on three datasets, embedded deployment of the wearable device, and hardware-in-the-loop *in silico trials*.

A. Clinical Datasets

We developed and evaluated the algorithms using three datasets collected from a number of T1D subjects in clinical trials. The first one is the OhioT1DM dataset [11], which is publicly available and contains the eight-week data of 12 T1D subjects who wore Medtronic Enlite CGM that measures BG levels every five minutes. The other two, the ABC4D dataset and ARISES dataset, are proprietary datasets (Imperial College London, London, UK). The ABC4D dataset contains data of 25 T1D participants over a six-month clinical trial (NCT02053051), where the participants used Dexcom G5 CGM. The ARISES dataset was collected in a six-week clinical trial (NCT03643692) with 12 T1D subjects whose BG levels were measured by Dexcom G6 CGM.

B. Experiment Setup and Evaluation Metrics

The OhioT1DM dataset contains the training set and testing set of each T1D subject [11], which account for the data of around 40 days and 10 days, respectively. Similarly, each of the ABC4D and the ARISES datasets was divided into a training set that includes the first 80% data and a testing set with the last 20% data. For each training set, the last 25% data was used as a validation set for hyperparameter tuning. This setup can avoid introducing temporal dependencies into training and testing sets, which was commonly used in previous work [10]. The selected values of the hyperparameters are listed in Table VIII in the Appendix.

We developed a personalized model for each T1D subject with 30-minute and 60-minute PHs and compared the proposed E3NN model against a group of baseline methods in the literature. The ARIMA [23] and SVR [26] were selected as two classic machine learning baselines [10], while the TCN [15], CRNN [16], LSTM [28], and Bi-LSTM [29], were employed as deep learning baselines. All the considered models were implemented by Python 3.8 and used the same input features, except for the ARIMA that used CGM input only. We respectively applied statsmodels 0.12 and scikit-learn 0.23 libraries to build the ARIMA and SVR models. The deep learning models were developed by TensorFlow 2.2 and Keras 2.3. We trained them using an Adam optimizer and early stopping to mitigate overfitting, which was accelerated by NVIDIA GTX 1080 Ti GPU. Notably, to evaluate the performance of model inference on the wearable SoC, we sequentially fed input data to the embedded E3NN models through a universal asynchronous receiver-transmitter (UART) with the general serial port data transmission protocol.

We evaluated the accuracy of BG prediction using three classic metrics: the root mean square error (RMSE), mean absolute error (MAE), and glucose-specific RMSE in mg/dL, which can be expressed as

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{N}\sum_{t=1}^{N}(G_t-\hat{G}_t)^2}, \\ \mathrm{MAE} &= \frac{1}{N}\sum_{t=1}^{N}|G_t-\hat{G}_t|, \\ \mathrm{gRMSE} &= \sqrt{\frac{1}{N}\sum_{t=1}^{N}P(G_t)(G_t-\hat{G}_t)^2}, \end{aligned} \tag{7}$$

where N is the number of total data samples in the testing sets; $P(G_t) \geq 1$ is a penalty function that penalizes underestimation in hyperglycemia and overestimation in hypoglycemia, whose formulation is defined in [63]. An error score (ES) that sums up the RMSE and MAE for the 30-minute and 60-minute PHs was used as the indicator of the validation performance.

According to the international consensus [64], a hypoglycemic event is defined as three consecutive CGM measurements below 70 mg/dL. The Matthews correlation coefficient (MCC) is used to evaluate hypoglycemia detection. It is a preferred metric in binary classifications since high MCC scores can be obtained only if the classifier performs well in all the categories of confusion matrix [65], which can be denoted as

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{F})(\mathrm{TN}+\mathrm{FN})}} \quad (8)$$

where TP stands for the number of true positives, i,e., the hypoglycemic events that are correctly detected by the predictions; TN is the number of true negatives; FP is the number of false positives; and FN means the number of false negatives.

TABLE I
PERFORMANCE OF THE PREDICTION MODELS EVALUATED ON THE OHIOT1DM DATASET

PH	Method	RMSE (mg/dL)	MAE (mg/dL)	gRMSE (mg/dL)
30 minutes	E3NN	$\mathbf{18.92 \pm 2.12}$	$\mathbf{13.46 \pm 1.49}$	$\mathbf{23.40 \pm 2.86}$
	TCN	$20.23 \pm 2.35^{\ddagger}$	$14.59 \pm 1.66^{\ddagger}$	$25.04 \pm 2.90^{\ddagger}$
	CRNN	$21.48 \pm 2.63^{\ddagger}$	$15.80 \pm 2.03^{\ddagger}$	$27.25 \pm 3.28^{\ddagger}$
	LSTM	20.11 ± 2.48	$14.06 \pm 1.69^{\dagger}$	$24.84 \pm 2.88^{*}$
	Bi-LSTM	$20.15 \pm 2.25^{*}$	$14.16 \pm 1.63^{\ddagger}$	$25.01 \pm 2.73^{\ddagger}$
	SVR	$21.37 \pm 2.25^{\ddagger}$	$16.27 \pm 1.68^{\ddagger}$	$26.73 \pm 2.87^{\ddagger}$
	ARIMA	$20.43 \pm 2.19^{\ddagger}$	$14.42 \pm 1.41^{\ddagger}$	$24.51 \pm 2.65^{\ddagger}$
60 minutes	E3NN	$\mathbf{32.54 \pm 3.61}$	$\mathbf{24.05 \pm 2.94}$	$\mathbf{41.52 \pm 4.83}$
	TCN	$34.21 \pm 3.71^{\dagger}$	$25.29 \pm 2.99^{\ddagger}$	$44.26 \pm 4.79^{\ddagger}$
	CRNN	$34.05 \pm 4.26^{\ddagger}$	$25.57 \pm 3.60^{\ddagger}$	$44.22 \pm 5.57^{\ddagger}$
	LSTM	$33.10 \pm 3.84^{*}$	24.50 ± 3.08	$42.65 \pm 5.20^{*}$
	Bi-LSTM	$33.76 \pm 4.06^{\ddagger}$	$25.10 \pm 3.31^{\dagger}$	$43.87 \pm 5.20^{\ddagger}$
	SVR	$33.99 \pm 3.59^{\ddagger}$	$25.69 \pm 2.77^{\ddagger}$	$44.21 \pm 4.94^{\ddagger}$
	ARIMA	$35.51 \pm 3.72^{\ddagger}$	$26.03 \pm 2.69^{\ddagger}$	$43.89 \pm 4.65^{*}$

$^{*}p \leq 0.05$ $^{\dagger}p \leq 0.01$ $^{\ddagger}p \leq 0.005$.

TABLE II
PERFORMANCE OF THE PREDICTION MODELS EVALUATED ON THE ABC4D DATASET

PH	Method	RMSE (mg/dL)	MAE (mg/dL)	gRMSE (mg/dL)
30 minutes	E3NN	$\mathbf{20.11 \pm 2.54}$	$\mathbf{14.34 \pm 1.78}$	$\mathbf{24.90 \pm 3.39}$
	TCN	21.86 ± 5.52	$15.06 \pm 1.89^{\ddagger}$	27.05 ± 6.33
	CRNN	$22.96 \pm 3.28^{\ddagger}$	$16.61 \pm 2.21^{\ddagger}$	$29.11 \pm 4.35^{\ddagger}$
	LSTM	$20.26 \pm 2.58^{\ddagger}$	$14.53 \pm 1.84^{\ddagger}$	$25.16 \pm 3.37^{\dagger}$
	Bi-LSTM	$20.36 \pm 2.56^{\ddagger}$	$14.64 \pm 1.84^{\ddagger}$	$25.38 \pm 3.33^{\ddagger}$
	SVR	$21.89 \pm 2.52^{\ddagger}$	$16.64 \pm 1.99^{\ddagger}$	$27.74 \pm 3.56^{\ddagger}$
	ARIMA	$22.15 \pm 2.59^{\ddagger}$	$15.61 \pm 1.90^{\ddagger}$	$26.48 \pm 3.56^{\ddagger}$
60 minutes	E3NN	$\mathbf{33.88 \pm 4.81}$	$\mathbf{24.98 \pm 3.56}$	$\mathbf{43.77 \pm 6.44}$
	TCN	40.56 ± 17.10	$26.17 \pm 3.88^{\ddagger}$	51.30 ± 19.48
	CRNN	38.23 ± 13.61	$26.97 \pm 4.19^{\ddagger}$	49.14 ± 15.07
	LSTM	34.31 ± 4.94	$25.36 \pm 3.67^{\ddagger}$	44.32 ± 6.80
	Bi-LSTM	$34.38 \pm 5.15^{*}$	$25.43 \pm 3.79^{\ddagger}$	$44.48 \pm 7.11^{*}$
	SVR	$34.90 \pm 4.76^{\ddagger}$	$26.46 \pm 3.61^{\ddagger}$	$45.43 \pm 6.55^{\ddagger}$
	ARIMA	$38.59 \pm 5.12^{\ddagger}$	$28.02 \pm 3.74^{\ddagger}$	$47.95 \pm 7.14^{\ddagger}$

$^{*}p \leq 0.05$ $^{\dagger}p \leq 0.01$ $^{\ddagger}p \leq 0.005$.

C. Prediction Performance

1) BG Level Prediction: Table I, II, and III respectively present the results of BG level prediction for the OhioT1DM, the ABC4D and the ARISES datasets over 30-minute and 60-minute PHs. To indicate the statistical significance with respect to the considered baselines, we confirmed the normality of data distribution with ShapiroWilk test and employed paired t-test to compute p values. It is worth noting that the E3NN achieved the best RMSE, MAE, and gRMSE for all three datasets and obtained significant improvement, compared with the considered baseline methods. Particularly, the improvement of the E3NN methods on the OhioT1DM dataset is more significant than that on the other two datasets, which is possibly due to the high quality of the dataset with the smallest portion of missing CGM samples. We observe that the RMSE for the 60-minute PH is much higher than that for the 30-minute PH, because external events, such as meal intake and exercise, and internal changes in a T1D subject are more likely to occur within a longer period, which would have an impact on glucose dynamics.

Overall, the deep learning methods performed better than the classic machine learning baselines, except for CRNN. The RNN-based models, including LSTM and Bi-LSTM, exhibited better performance than TCN and CRNN that use CNN layers for feature extraction, where the LSTM performed best among the baseline model. In addition, it is noted that the performance of the ARIMA is good for the 30-minute PH but degraded with a longer PH. A possible explanation is that the ARIMA method uses a linear equation, for which it is difficult to capture non-linear long-term temporal dependencies. The trajectories of the CGM measurements and the predictive results of the E3NN, LSTM, TCN, and ARIMA methods are shown in Fig. 5. The dashed green and cyan lines indicate the thresholds of hypoglycemia and hyperglycemia, respectively. When compared with the LSTM and TCN methods, the E3NN method obtained less underestimation in hyperglycemic regions and less overestimation in hypoglycemic regions. However, it is observed that deep learning methods lack sensitivity for BG changes at the troughs of the plotted curves, as highlighted by the black ellipses. This may cause missed detection of severe hypoglycemia and lead to life-threatening events in clinical settings. Therefore, we introduced the corresponding LBs to address this challenge, which is detailed in Section IV-C3.

2) Comparison among Deep Learning Methods: As an edge AI application implemented on a hardware platform with limited computational resources, the memory footprint and operations per inference, as well as prediction accuracy, are important considerations during the selection of deep learning models. The deep learning models were developed by the TensorFlow library. Thus, we converted them into a TensorFlow Lite compressed format that supports on-device inference for many mobile and IoT devices, to analyze the hardware requirements. Table IV summarizes the number of parameters (Param) and floating-point operations per second (FLOPS), peak SRAM, Flash, and the ES for each DNN architecture. It is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3143375, IEEE Internet of Things Journal

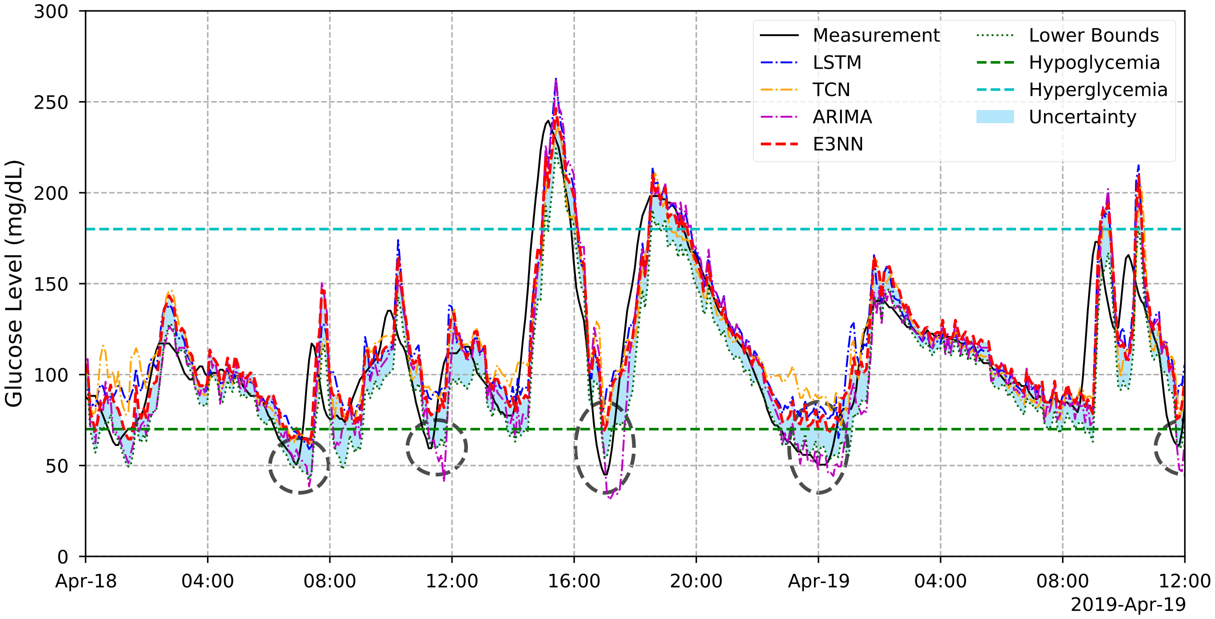


Fig. 5. 1.5-day prediction performance of the considered methods on the ARISES dataset with a 30-minute PH. The solid black line indicates the actual CGM measurements, while the dash-dotted blue, orange, and magenta lines are the results of LSTM, TCN, ARIMA methods. The dashed red line indicates the results of the E3NN method, where the LBs and uncertainty are represented by the dotted green line and shaded blue area, respectively. The black ellipses highlight the hypoglycemic events detected by the LBs.

TABLE III
PERFORMANCE OF THE PREDICTION MODELS EVALUATED ON THE ARISES DATASET

PH	Method	RMSE (mg/dL)	MAE (mg/dL)	gRMSE (mg/dL)
30 minutes	E3NN	$\mathbf{20.45 \pm 3.81}$	$\mathbf{14.78 \pm 2.62}$	$\mathbf{25.31 \pm 5.09}$
	TCN	$22.01 \pm 4.19^{\ddagger}$	$15.98 \pm 2.91^{\ddagger}$	$28.03 \pm 5.80^{\ddagger}$
	CRNN	$24.43 \pm 5.19^{\ddagger}$	$17.93 \pm 3.76^{\ddagger}$	$31.67 \pm 7.28^{\ddagger}$
	LSTM	20.74 ± 3.66	15.03 ± 2.55	$26.09 \pm 4.89^{*}$
	Bi-LSTM	$20.86 \pm 3.78^{*}$	$15.19 \pm 2.73^{\ddagger}$	$26.30 \pm 5.12^{\dagger}$
	SVR	$22.87 \pm 3.99^{\ddagger}$	$17.25 \pm 2.99^{\ddagger}$	$29.10 \pm 5.49^{\ddagger}$
	ARIMA	$21.76 \pm 4.73^{\ddagger}$	$15.59 \pm 2.71^{\ddagger}$	$26.20 \pm 5.20^{\ddagger}$
60 minutes	E3NN	$\mathbf{35.55 \pm 7.24}$	$\mathbf{26.22 \pm 5.28}$	$\mathbf{46.37 \pm 10.11}$
	TCN	$37.01 \pm 7.72^{\ddagger}$	$27.64 \pm 5.66^{\ddagger}$	$48.65 \pm 10.76^{\ddagger}$
	CRNN	$38.07 \pm 8.20^{\ddagger}$	$28.49 \pm 6.15^{\ddagger}$	$50.17 \pm 11.37^{\ddagger}$
	LSTM	$36.68 \pm 6.97^{\ddagger}$	$27.02 \pm 5.12^{*}$	$48.80 \pm 9.83^{\ddagger}$
	Bi-LSTM	$37.14 \pm 7.38^{\ddagger}$	$27.59 \pm 5.45^{\ddagger}$	$49.00 \pm 10.55^{\ddagger}$
	SVR	$37.08 \pm 7.48^{\ddagger}$	$27.75 \pm 5.44^{\ddagger}$	$48.79 \pm 10.42^{\ddagger}$
	ARIMA	$39.51 \pm 8.16^{\ddagger}$	$28.75 \pm 5.65^{\ddagger}$	$49.73 \pm 11.00^{\ddagger}$

$^{*}p \leq 0.05$ $^{\dagger}p \leq 0.01$ $^{\ddagger}p \leq 0.005$.

TABLE IV
COMPARISON BETWEEN THE PROPOSED E3NN AND CONSIDERED DEEP LEARNING BASELINE METHODS.

Method	Param	FLOPs	SRAM	Flash	ES (mg/dL)
TCN	124K	248K	13.7KB	499KB	94.44
CRNN	52K	136K	8.1KB	227KB	96.78
LSTM	53k	1577K	7.3KB	2096KB	91.87
Bi-LSTM	141K	412K	13.2KB	624KB	93.17
E3NN	**32K**	**93K**	13.8KB	**171KB**	**88.97**

noteworthy that the E3NN model obtained the best prediction performance (the lowest ES) with the smallest numbers of parameters, FLOPs, and Flash. Although the E3NN consumes relatively high peak SRAM, this amount is much smaller than the available capacity of most commercial MCUs, as well as the target SoC in this work (64KB). We observe that the LSTM model achieved the second-best ES at the cost of a large number of parameters and Flash requirement that is likely to exceed the memory constraint.

3) Hypoglycemia Detection: A widespread application of BG prediction in T1D management systems is to prevent hypoglycemic episodes that would lead to fatal complications. We detected impending hypoglycemia using the LBs of E3NN predictions and the prediction values of the considered baseline methods at the same PHs. Table V presents the MCC scores evaluated on the three clinical datasets. Although the TCN-based model obtained higher RMSE, MAE, and gRMSE results than the LSTM and Bi-LSTM in Table I, II, and III, it is worth noting that the TCN achieved the best performance of hypoglycemia detection among all the considered deep learning baseline methods. A possible explanation is that the TCN-based models have longer effective memory than canonical RNNs with the same capacity, as suggested in [31]. Therefore, the TCN model could better understand the patterns of hypoglycemia caused by external events that occurred hours ago, such as postprandial hypoglycemia. In our previous work [15], we also noticed that the TCN model exhibited a short prediction time lag, indicating good sensitivity to the changes in the troughs of glucose trajectories, i.e., hypoglycemia regions. Meanwhile, it is noted that the ARIMA outperformed the SVR model with a higher MCC score. Therefore, we compared the E3NN with the TCN model

TABLE V
MCC SCORES OF THE HYPOGLYCEMIA PREDICTION EVALUATED ON THE THREE DATASETS

PH	Method	OhioT1DM	ABC4D	ARISES
30 min	E3NN	$\mathbf{0.70 \pm 0.09}$	$\mathbf{0.68 \pm 0.09}$	$\mathbf{0.70 \pm 0.12}$
	TCN	$0.55 \pm 0.10^{\ddagger}$	$0.49 \pm 0.20^{\ddagger}$	$0.40 \pm 0.12^{\ddagger}$
	ARIMA	$0.65 \pm 0.09^{*}$	0.59 ± 0.07	0.61 ± 0.10
60 min	E3NN	$\mathbf{0.57 \pm 0.09}$	$\mathbf{0.54 \pm 0.11}$	$\mathbf{0.49 \pm 0.14}$
	TCN	$0.38 \pm 0.11^{\ddagger}$	$0.37 \pm 0.18^{\ddagger}$	$0.30 \pm 0.15^{\ddagger}$
	ARIMA	0.49 ± 0.06	$0.48 \pm 0.06^{\ddagger}$	0.45 ± 0.12

$^{*}p \leq 0.05$ $^{\dagger}p \leq 0.01$ $^{\ddagger}p \leq 0.005$.

TABLE VI
DETAILS OF FLASH AND SRAM MEMORY FOOTPRINT

Layer	Input Shape	Flash (B)	SRAM (B)	Time
Input	(2, 12)	0	96	0
GRU 1 **	(1, 2)	52,224	1,536	22.58 ms
GRU 2 **	(1, 64)	37,632	768	16.16 ms
Attention	(12, 32)	20,480	2,096	28.92 ms
Dense	(1, 64)	16,640	256	6.22 ms
Evidential	(1, 64)	1,040	16	0.32 ms
Output	(1, 4)	0	16	0

** This layer is repeatedly executed for 12 times.

and ARIMA in Table V.

Notably, the E3NN model achieved the highest MCC scores for each dataset in both 30-minute and 60-minute PHs. It is interesting to note that the MCC scores of the ARIMA method are significantly higher than the TCN and the other deep learning methods, while the improvement of the E3NN on the ABC4D and ARISES datasets is not significant when compared with the ARIMA. In Fig 5, we see that the ARIMA predictions can identify more hypoglycemic events than the TCN with a time-shifted delay on the curve, which, however, degrades the RMSE performance (Table I). It is reasonable since the weights of the DNN models were optimized by the regression loss that aims to enhance RMSE performance, instead of the accuracy of hypoglycemia detection. It is observed that the LBs of the E3NN curve successfully detected five hypoglycemic events circled by the black ellipse, which are likely to be missed if we use the prediction values only. In particular, the use of LBs increased average MCC scores for the three datasets by 0.13 ($p < 0.005$) and 0.17 ($p < 0.005$) for the 30-minute and 60-minute PHs, respectively. Hence, these results suggest that evidential regression is an important improvement in BG prediction methods based on deep learning.

D. Edge Implementation

The proposed RNN predictor was implemented on the BLE SoC for edge computing. By means of utilizing the optimized CMSIS-DSP library that is pre-compiled and included in the latest NRF52 SDK, the SoC is able to accept high-throughput data while performing rapid computation of matrix operations that typically involve single-cycle multiplication and accumulation. This enables efficient data processing with minimal overhead and the real-time execution of computation-intensive algorithms.

Table VI presents the detailed utilization of Flash and SRAM memory in Byte (B) for the implementation of the proposed RNN model. As the RAM memory was allocated dynamically during the run time, the implementation of this RNN model only led to an increase of 2.48% on the SRAM utilization compared with that without the edge computation. Whereas the capacity of the Flash memory is the main bottleneck that limits the size of the RNN predictor, occupying 66.13% of the total flash utilization. In addition, the computation time was empirically estimated by executing each layer for 100 rounds and averaging the run time through the UART timestamp. The result shows an average computation time of approximately 500 ms. Moreover, compared with the implementation by TensorFlow Lite Micro in Table IV, our implementation significantly reduced Flash from 171 KB to 125 KB and peak SRAM from 13.8 KB to 3.5 KB, mainly because it computed outcomes using low-level CMSIS-DSP APIs without interpreting the network graph. For each considered T1D subject, the RMSE between the testing results of Python model and those of the edge model is less than 10^{-5} mg/dL.

The final firmware for the BLE SoC utilizes 189.03 KB Flash and 16.12 KB SRAM memory, which provides the following six functionalities: 1) CGM sensor connectivity and readout, 2) input data pre-processing, 3) edge computing of the RNN predictor, 4) external flash memory management, 5) basic user interactions, and 6) designer mode for data readout and parameter update.

E. Power Analysis

Power estimation was conducted by using a source meter Keithley 2606A, which supplied 3 volts and monitored the power in real time. Fig. 6 presents the power monitoring of a typical cycle that lasts for 13 seconds, which consumes an average run-time power of 3.78 mW. The power spikes at the initial stage indicate the Bluetooth scanning process, which involves a tunable window and interval. The highest power occurring in the middle indicates the edge computing for the embedded predictor. During the pulse at the end, the system polls the power-gating circuit to enter shutdown mode.

At the very beginning, the device keeps scanning the target sensor and involves an on-off current switching with a peak value around 6.5 mA. To reduce the power consumption of this process, the BLE scanning window is shortened into a duty cycle of 10%, resulting in an average power of 5 mW. Once the target sensor is connected, the BLE SoC will start the authentication and bonding process, which typically lasts for 2 seconds. After the success of bonding, the device is able to request the glucose data and start prediction. Notably, the running of the RNN predictor is the most energy-hungry process as it utilizes the on-chip digital signal processing for the computation of floating-point arithmetic operations, but only takes a short operating time of around 500 ms. Depending on the predicted blood glucose level, the notifications of the low excursion with intermittent alarms are generated through the LED and buzzer. During this period, the system waits for user response but maintains a low power that is less than 1 mW.

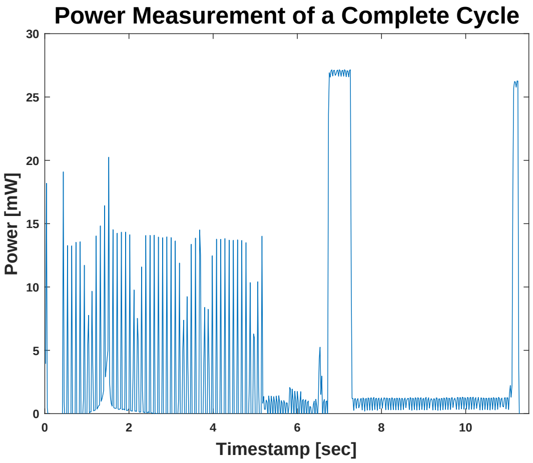


Fig. 6. Power measurement of a complete run cycle that involves BLE scanning, authentication and bonding, edge AI computing, and hypoglycemia notification. Among these processes, the two main contributors, including BLE scanning and edge AI computation, lasted for 5 and 0.52 seconds, which consumed an average power of 5 mW and 27 mW, respectively.

If the button is pressed, i.e., the warning of a hypoglycemic event is confirmed, the system will enter the shutdown mode by triggering the on-board timer for power gating.

In the real application, the device can be powered by a single coin battery, e.g., CR2032, which typically has a capacity of 240 mAh. This capacity enables the device to operate for six months, assuming that each CGM readout is processed every five minutes. The commercial CGM sensors and transmitters in the market typically require a replacement every 10 days and 3 months, respectively. Thus, the achieved battery life of our wearable device is long enough to cover these periods.

F. In Silico Trial

To evaluate the performance of the whole system with the wearable device, we performed a 3-month hardware-in-the-loop *in silico* trial using the UVA/Padova T1D simulator, which is a common experimental setup of pre-clinical trials in T1D management systems. In particular, we employed 10 virtual adult subjects with additional intra- and inter-subject variability [66] and used the carbohydrate of meal protocol as follows: 70 g (breakfast, 7 am), 110 g (lunch, 2 pm), and 90 g (dinner, 9 pm), with the variability of mealtime (STD = 30%) and meal size (CV = 10%). The simulator sent CGM values to the wearable device and received 30-minute predictions through a debug mode with the UART and USB ports. We performed the PLGM algorithm with the settings in [67], where the pump suspended basal insulin when the predictions were at or below the threshold of hypoglycemia, i.e., 70 mg/dL.

Table VII presents the outcomes of the PLGM and a control group (i.e., no suspension) as a baseline, evaluated by time in range (TIR) of [70, 180] mg/dL, time below range (TBR) (BG< 70 mg/dL), time of severe hypoglycemia (TSH) (BG< 54 mg/dL), low blood glucose risk index (LBGI). It is noted that integrating the wearable device with the PLGM significantly reduced the LBGI and percent time of hypoglycemia and severe hypoglycemia without a decrease of TIR.

TABLE VII
GLYCEMIC OUTCOMES OF THE IN SILICO TRIAL

Method	TIR (%)	TBR (%)	TSH (%)	LBGI
Control	74.26 ± 7.62	$5.44 \pm 3.38^{\ddagger}$	$2.00 \pm 1.45^{\ddagger}$	$1.50 \pm 0.81^{\ddagger}$
PLGM	74.83 ± 9.02	2.02 ± 1.14	0.47 ± 0.38	0.65 ± 0.28

$^{*}p \leq 0.05$ $^{\dagger}p \leq 0.01$ $^{\ddagger}p \leq 0.005$.

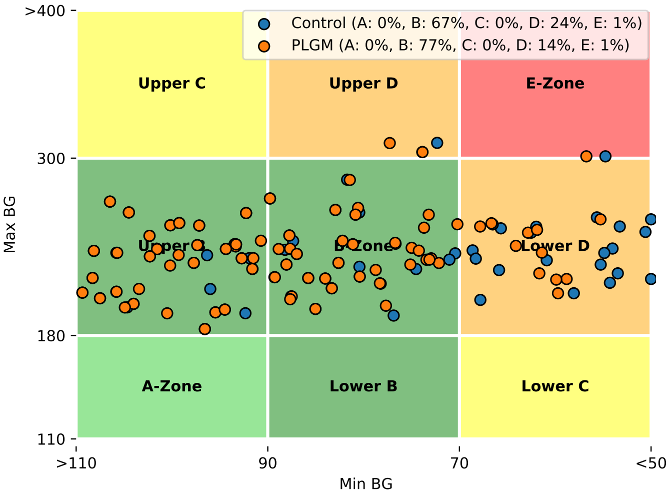


Fig. 7. CVGA plot comparing PLGM (orange dots) against the control group (blue dots) for a virtual adult subject in the trial. Each dot stands for the extreme values of BG trajectories over 24 hours.

Fig. 7 depicts the outcomes of control-variability grid analysis (CVGA) for a virtual adult subject. CVGA is a common method to visualize the efficacy of glucose regulation [68], [69], where each dot on the plot indicates the minimum and maximum BG values in a 24-hour period. We observe that, compared with the control group, more of the PLGM dots are located in the left bottom zones. Specifically, the PLGM improved the percentage of the A+B zone from 67% to 77% and reduced 10% of the dots in the D+E zone, indicating good BG control. Besides the PLGM, other interventions, such as glucagon delivery and rescue carbohydrate recommendations, could also be performed in clinical settings to further reduce the incidence of hypoglycemia, based on the real-time BG predictions of the wearable devices.

V. CONCLUSION

In this article, we proposed a GRU-based RNN model, the E3NN, with attention mechanism and evidential regression and developed a novel IoMT-enabled wearable device to implement the deep learning algorithm for real-time BG prediction and hypoglycemia warning with edge computing on the SoC. When evaluated on the three clinical datasets, the proposed model obtained the best prediction accuracy for both future BG level and impending hypoglycemic events with the smallest number of model parameters and FLOPs, compared with the considered deep learning baseline methods. Moreover, the optimized hardware design of the wearable device enables extremely low energy consumption for edge inference and BLE connectivity, which can run 24/7 operations

TABLE VIII
LIST OF HYPERPARAMETERS

Parameter	Value
Hidden units of GRU layers	[64,32]
Hidden units of the attention layer	64
Hidden units of the dense layer	64
Dropout rate	0.1
Learning rate	1×10^{-3}
Length of input sequences	12
Batch size	32
Number of epochs	300
Early stopping patience	30

over six months. The results of *in silico* trials demonstrated that integrating the wearable device into the T1D management system notably improved glycemic outcomes of BG control.

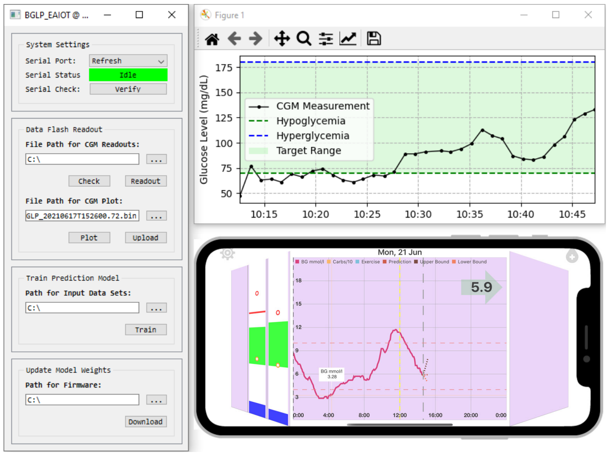
In future work, the wearable device with the proposed algorithm will be evaluated in actual clinical trials to further investigate the performance of software and hardware in real-world settings and modify the functions and GUIs according to user feedback. Considering that the edge computing for the E3NN is based on a tiny MCU of the SoC, it is possible to deploy the prediction algorithm in other T1D IoMT wearable devices with the collaboration of manufactures, such as CGM and insulin pumps, to provide on-device decision support.

Appendix A
Hyperparameters

Table VIII listed the hyperparameters used in the E3NN model, which were determined by the Hyperband algorithm [70] with the Keras Tuner.

Appendix B
GUIs of Smartphone and Desktop Platforms

Fig. 8 depicts the GUIs of the iOS app and the desktop platform developed by Swift 4.2 and PyQt5 5.15, respectively. The desktop platform consists of multiple panels, including system settings, data readout, model training and update, and visualization of historical CGM data. Besides the historical trajectories, the smartphone platform also supports the visualization of the current CGM value and trend as shown by the green arrow.



BGLP_EAIOT @ ...

System Settings
Serial Port: Refresh
Serial Status: Idle

Data Flash Readout
File Path for CGM Readouts: C:\
File Path for CGM Plot: GLP_20210617T152600.72.bin

Train Prediction Model
Path for Input Data Sets: C:\

Update Model Weights
Path for Firmware: C:\

Fig. 8. Overview of the smartphone and desktop GUIs.

Acknowledgment

This research has been funded by Engineering and Physical Sciences Research Council (EPSRC EP/P00993X/1) and the President's PhD Scholarship at Imperial College London (UK). We would like to thank Prof. Nick Oliver, Dr. Monika Reddy, Dr. Chukwuma Uduku, and Narvada Jugnee for their contribution in obtaining the ABC4D and ARISES datasets. We would also like to thank Prof. Robert Spence for his contribution in designing the smartphone GUI.

References

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, and Others, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas," *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, 2019.

[2] 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.

[3] J.-F. Yale, B. Paty, and P. A. Senior, "Hypoglycemia," *Canadian Journal of Diabetes*, vol. 42, pp. S104—-S108, 2018.

[4] I. Contreras, S. Oviedo, M. Vettoretti, R. Visentin, and J. Vehí, "Personalized blood glucose prediction: A hybrid approach using grammatical evolution and physiological models," *PloS One*, vol. 12, no. 11, p. e0187754, 2017.

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

[6] M. Swan, "Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0," *Journal of Sensor and Actuator Networks*, vol. 1, no. 3, pp. 217–253, 2012.

[7] J. E. Pinsker, A. J. Laguna Sanz, J. B. Lee, M. M. Church, C. Andre, L. E. Lindsey, F. J. Doyle III, and E. Dassau, "Evaluation of an artificial pancreas with enhanced model predictive control and a glucose prediction trust index with unannounced exercise," *Diabetes Technology & Therapeutics*, vol. 20, no. 7, pp. 455–464, 2018.

[8] T. Battelino, R. Nimri, K. Dovc, M. Phillip, and N. Bratina, "Prevention of hypoglycemia with predictive low glucose insulin suspension in children with type 1 diabetes: a randomized controlled trial," *Diabetes Care*, vol. 40, no. 6, pp. 764–770, 2017.

[9] A. Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, and G. Hartvigsen, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artificial Intelligence in Medicine*, vol. 98, pp. 109–134, 2019.

[10] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[11] 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.

[12] J. Chen, K. Li, P. Herrero, T. Zhu, and P. Georgiou, "Dilated recurrent neural network for short-time prediction of glucose concentration." in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, 2018, pp. 69–73.

[13] 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*, 2020, pp. 105–109.

[14] 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.

[15] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[16] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[17] H. M. Rooijackers, E. C. Wiegers, M. van der Graaf, D. H. Thijssen, R. P. Kessels, C. J. Tack, and B. E. de Galan, "A single bout of high-intensity interval training reduces awareness of subsequent hypoglycemia in patients with type 1 diabetes," *Diabetes*, vol. 66, no. 7, pp. 1990–1998, 2017.

[18] A. Trifan, M. Oliveira, and J. L. Oliveira, "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.

[19] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, 2016.

[20] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[21] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, "Scaling for edge inference of deep neural networks," *Nature Electronics*, vol. 1, no. 4, pp. 216–222, 2018.

[22] 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.

[23] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, "A machine learning approach to predicting blood glucose levels for diabetes management," in *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence*, 2014.

[24] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "A predictive model of subcutaneous glucose concentration in type 1 diabetes based on random forests," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2012, pp. 2889–2892.

[25] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gómez, M. Rigla, A. de Leiva, and M. Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes Technology & Therapeutics*, vol. 12, no. 1, pp. 81–88, 2010.

[26] 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, 2012.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[28] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, "Blood glucose prediction with variance estimation using recurrent neural networks," *Journal of Healthcare Informatics Research*, vol. 4, no. 1, pp. 1–18, 2020.

[29] A. Mohebbi, A. R. Johansen, N. Hansen, P. E. Christensen, J. M. Tarp, M. L. Jensen, H. Bengtsson, and M. Mørup, "Short term blood glucose prediction based on continuous glucose monitoring data," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5140–5145.

[30] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction," in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, 2018, pp. 64–78.

[31] S. Bai, J. Z. Kolter, and V. Koltun, "Convolutional sequence modeling revisited," in *Workshop Track - Sixth International Conference on Learning Representations (ICLR)*, 2018.

[32] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *Journal of Healthcare Informatics Research*, pp. 1–17, 2020.

[33] 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 *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 706–712.

[34] F. Prendin, S. Del Favero, M. Vettoretti, G. Sparacino, and A. Facchinetti, "Forecasting of glucose levels and hypoglycemic events: Head-to-head comparison of linear and nonlinear data-driven algorithms based on continuous glucose monitoring data only," *Sensors*, vol. 21, no. 5, p. 1647, 2021.

[35] N. Allen and A. Gupta, "Current diabetes technology: striving for the artificial pancreas," *Diagnostics*, vol. 9, no. 1, p. 31, 2019.

[36] S. B. Baker, W. Xiang, and I. Atkinson, "Internet of things for smart healthcare: Technologies, challenges, and opportunities," *IEEE Access*, vol. 5, pp. 26 521–26 544, 2017.

[37] L. Catarinucci, D. De Donno, L. Mainetti, L. Palano, L. Patrono, M. L. Stefanizzi, and L. Tarricone, "An IoT-aware architecture for smart healthcare systems," *IEEE Internet of Things Journal*, vol. 2, no. 6, pp. 515–526, 2015.

[38] E. Spanò, S. Di Pascoli, and G. Iannaccone, "Low-power wearable ecg monitoring system for multiple-patient remote monitoring," *IEEE Sensors Journal*, vol. 16, no. 13, pp. 5452–5462, 2016.

[39] A. Gatouillat, Y. Badr, B. Massot, and E. Sejdić, "Internet of medical things: A review of recent contributions dealing with cyber-physical systems in medicine," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3810–3822, 2018.

[40] Y.-S. Su, T.-J. Ding, and M.-Y. Chen, "Deep learning methods in internet of medical things for valvular heart disease screening system," *IEEE Internet of Things Journal*, 2021.

[41] S. Tuli, N. Basumatary, S. S. Gill, M. Kahani, R. C. Arya, G. S. Wander, and R. Buyya, "Healthfog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments," *Future Generation Computer Systems*, vol. 104, pp. 187–200, 2020.

[42] Y. Tai, B. Gao, Q. Li, Z. Yu, C. Zhu, and V. Chang, "Trustworthy and intelligent covid-19 diagnostic iomt through xr and deep learning-based clinic data access," *IEEE Internet of Things Journal*, 2021.

[43] O. AlShorman, B. AlShorman, M. Alkhassaweneh, and F. Alkahtani, "A review of internet of medical things (iomt)–based remote health monitoring through wearable sensors: A case study for diabetic patients," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 1, pp. 414–422, 2020.

[44] A. Chakrabarty, S. Zavitsanou, T. Sowrirajan, F. J. Doyle III, and E. Dassau, "Getting IoT-ready: The face of next generation artificial pancreas systems," in *The Artificial Pancreas*. Elsevier, 2019, pp. 29–57.

[45] Z. A. Al-Odat, S. K. Srinivasan, E. M. Al-Qtiemat, and S. Shuja, "A reliable IoT-based embedded health care system for diabetic patients," *International Journal on Advances in Internet Technology Volume 12, Number 1 & 2, 2019*, 2019.

[46] P. Herrero, M. El-Sharkawy, J. Daniels, N. Jugnee, C. N. Uduku, M. Reddy, N. Oliver, and P. Georgiou, "The bio-inspired artificial pancreas for type 1 diabetes control in the home: system architecture and preliminary results," *Journal of Diabetes Science and Technology*, vol. 13, no. 6, pp. 1017–1025, 2019.

[47] P. Herrero, M. El Sharkawy, P. Pesl, M. Reddy, N. Oliver, D. Johnston, C. Toumazou, and P. Georgiou, "Live demonstration: A handheld bio-inspired artificial pancreas for treatment of diabetes," in *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*. IEEE, 2014, pp. 172–172.

[48] S. J. Russell, F. H. El-Khatib, M. Sinha, K. L. Magyar, K. McKeon, L. G. Goergen, C. Balliro, M. A. Hillard, D. M. Nathan, and E. R. Damiano, "Outpatient glycemic control with a bionic pancreas in type 1 diabetes," *New England Journal of Medicine*, vol. 371, no. 4, pp. 313–325, 2014.

[49] M. D. Breton, D. R. Cherñavvsky, G. P. Forlenza, M. D. DeBoer, J. Robic, R. P. Wadwa, L. H. Messer, B. P. Kovatchev, and D. M. Maahs, "Closed-loop control during intense prolonged outdoor exercise in adolescents with type 1 diabetes: the artificial pancreas ski study," *Diabetes Care*, vol. 40, no. 12, pp. 1644–1650, 2017.

[50] P. V. Astillo, G. Choudhary, D. G. Duguma, J. Kim, and I. You, "Trmaps: Trust management in specification-based misbehavior detection system for imd-enabled artificial pancreas system," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[51] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: the confluence of edge computing and artificial intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020.

[52] F. Alshehri and G. Muhammad, "A comprehensive survey of the internet of things (IoT) and ai-based smart healthcare," *IEEE ACCESS*, vol. 9, pp. 3660–3678, 2021.

[53] A. Pazienza, R. Anglani, G. Mallardi, C. Fasciano, P. Noviello, C. Tatulli, and F. Vitulano, "Adaptive critical care intervention in the internet of medical things," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. IEEE, 2020, pp. 1–8.

[54] X. Kong, K. Wang, S. Wang, X. Wang, X. Jiang, Y. Guo, G. Shen, X. Chen, and Q. Ni, "Real-time mask identification for covid-19: an edge computing-based deep learning framework," *IEEE Internet of Things Journal*, 2021.

[55] I. L. Olokodana, S. P. Mohanty, E. Kougianos, and O. O. Olokodana, "Real-time automatic seizure detection using ordinary kriging method in an edge-iomt computing paradigm," *SN Computer Science*, vol. 1, no. 5, pp. 1–15, 2020.

[56] S. Hamdan, M. Ayyash, and S. Almajali, “Edge-computing architectures for internet of things applications: A survey,” *Sensors*, vol. 20, no. 22, p. 6441, 2020.

[57] X. Xu, C. He, Z. Xu, L. Qi, S. Wan, and M. Z. A. Bhuiyan, “Joint optimization of offloading utility and privacy for edge computing enabled IoT,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2622–2629, 2019.

[58] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015.

[59] M. T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015.

[60] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, “Deep evidential regression,” in *Advances in Neural Information Processing Systems*, 2020.

[61] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[62] L. Lai, N. Suda, and V. Chandra, “CMSIS-NN: Efficient neural network kernels for arm cortex-m cpus,” *arXiv preprint arXiv:1801.06601*, 2018.

[63] 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.

[64] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, S. Garg, L. Heinemann, I. Hirsch, S. A. Amiel *et al.*, “International consensus on use of continuous glucose monitoring,” *Diabetes Care*, vol. 40, no. 12, pp. 1631–1640, 2017.

[65] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[66] 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.

[67] C. Liu, P. Avari, Y. Leal, M. Wos, K. Sivasithamparam, P. Georgiou, M. Reddy, J. M. Fernández-Real, C. Martin, M. Fernández-Balsells *et al.*, “A modular safety system for an insulin dose recommender: a feasibility study,” *Journal of Diabetes Science and Technology*, vol. 14, no. 1, pp. 87–96, 2020.

[68] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Basal glucose control in type 1 diabetes using deep reinforcement learning: an in silico validation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1223–1232, 2020.

[69] T. Zhu, K. Li, L. Kuang, P. Herrero, and P. Georgiou, “An insulin bolus advisor for type 1 diabetes using deep reinforcement learning,” *Sensors*, vol. 20, no. 18, p. 5058, 2020.

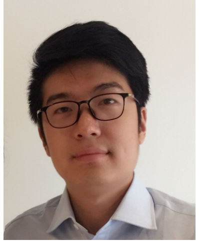
[70] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

Taiyu Zhu (Student Member, IEEE) received the B.Eng. (Hons.) degree in electrical and electronic engineering from the Australian National University, Canberra, Australia, and the M.Sc. degree from Imperial College London, London, U.K., in 2017 and 2018, respectively. He is currently working toward the Ph.D. degree with the Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London. His research interests include biomedical signal processing, machine learning and deep learning in diabetes technology, and artificial intelligence in healthcare. He was the recipient of the President’s Ph.D. Scholarship at Imperial College London.

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.

John Daniels (Student Member, IEEE) is a PhD student at the Centre for Bio-Inspired Technology, Imperial College London. He received the M.Eng degree in Electrical & Electronic Engineering with Management from Imperial College London, U.K., in 2015. His broad research interests lie at the intersection of machine learning, embedded systems, and diabetes technology.

Pau Herrero (Member, IEEE) received the MSc degree in Information Technologies from University of Girona (Spain) in 2002 and the Ph.D. degree in Automation from University of Angers (France) in 2007. Over that past 12 years, he has worked as a Research Fellow within the department of Electrical and Electronic Engineering at Imperial College London (UK). His main research interest lies in the field of diabetes technology and antimicrobial resistance. In particular, he has been involved in the development and clinical validation of closed-loop drug delivery systems and advanced clinical decision support systems for diabetes management and antimicrobial prescription. He holds five patents and has been involved in several technology transfer activities with the biotech industry.



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 (AM05M08SM13) 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.