Edge-Based Temporal Fusion Transformer for Multi-Horizon Blood Glucose Prediction

Taiyu Zhu[†], Tianrui Chen[†], Lei Kuang[†], Kezhi Li[†★] and Pantelis Georgiou[†]
† Centre for Bio-inspired Technology, Imperial College London, London, United Kingdom
Email: {taiyu.zhu17, tianrui.chen21, lei.kuang18, pantelis}@imperial.ac.uk
★ Institute of Health Informatics, University College London, London, United Kingdom
Email: ken.li@ucl.ac.uk

Abstract—Deep learning models have achieved the state of the art in blood glucose (BG) prediction, which has been shown to improve type 1 diabetes (T1D) management. However, the models in most existing studies can only provide single-horizon prediction and face a variety of real-world challenges, such as lacking hardware implementation and interpretability. In this work, we introduce a new deep learning framework, the edge-based temporal fusion transformer (E-TFT), for multi-horizon BG prediction, and implement the trained model on a customized wristband with a system on a chip (Nordic nRF52832) for edge computing. E-TFT employs a self-attention mechanism to extract long-term temporal dependencies and enables post-hoc explanation for feature selection. On a clinical dataset with 12 T1D subjects, it achieved a mean root mean square error of 19.09 ± 2.47 mg/dL and 32.31 ± 3.79 mg/dL for 30 and 60-minute prediction horizons, respectively, and outperformed all the considered baseline methods, such as N-BEATS and N-HiTS. The proposed model is effective for multi-horizon BG prediction and can be deployed on wearable devices to enhance T1D management in clinical settings.

Index Terms—Diabetes, deep learning, multi-horizon prediction, edge computing, transformer.

I. INTRODUCTION

Diabetes mellitus is a global health challenge that affects more than half a billion people and becomes an increasingly heavy economic burden to many countries [1]. Type 1 diabetes (T1D) occurs when the immune system attacks and destroys the β-cells of the pancreas, and people with T1D require the administration of exogenous insulin and long-term blood glucose (BG) management [2]. In this case, BG level prediction is a viable method to enable proactive interventions and thereby reduce the short and long-term complications caused by hyperglycemia or hypoglycemia, such as retinopathy, nephropathy, and cardiovascular disease [3].

In recent decades, continuous glucose monitoring (CGM) systems have been widely adopted in T1D management [4], which generally measure BG concentration every five minutes and thus provide vital data for achieving accurate BG prediction. Daily events, such as carbohydrate intake, insulin bolus delivery, and exercise, can significantly influence BG levels, which are commonly recorded via smartphone apps, as shown in Fig. 1, and employed as additional data sources to develop data-driven decision support algorithms. In this regard, artificial intelligence technology, especially deep learning, has

shown excellent performance in BG prediction for various T1D datasets [5]. Recent advances in self-attention mechanisms [6] have achieved the state-of-the-art in natural language processing [7] and boosted the development of transformer-based models for time series forecasting, such as Informer [8] and Autoformer [9]. Among these, temporal fusion transformers (TFT) [10] is a specifically-designed multi-horizon transformer that employs gating mechanisms to simultaneously process various data for real-world applications, including exogenous time series, future known information, and static metadata. This feature is extremely useful in BG prediction with heterogeneous data sources (Fig. 1).

To bring the models to practical use in T1D management, edge computing-enabled deep learning models for wearable medical devices are in urgent need [11]. However, increasingly complex models rely on a large number of parameters and intensive computational resources. It is challenging to implement these models on a system on a chip (SoC) that interacts with CGM and provides real-time BG prediction at the same time. In addition, model interpretability plays an important role in actual clinical settings, which is the key to feature selection and pattern analysis in time series forecasting but is still under research [5], [10].

II. RELATED WORK

Leveraging powerful deep neural networks, a wide range of deep learning models has been applied to BG prediction [5] and outperformed conventional machine learning algorithms, such as support vector regression (SVR) [12]. In particular, many existing studies are based on recurrent neural networks (RNNs), including long short-term memory (LSTM) [13]–[16] and gated recurrent units (GRUs) [11], [17]–[19], which employ hidden states to extract temporal features from input sequences. Recently, attention mechanisms were investigated and combined with LSTM [14] and GRU [19] to assign weights to hidden states and further enhance BG prediction. In [19], the authors further adopted the encoder of the original transformer [6] as a baseline method, which exhibited comparable performance to RNN models for BG prediction on three clinical datasets. However, most existing studies used many-to-one RNNs for single-horizon prediction [13]–[20]. These models require multiple sets of model weights for two or more
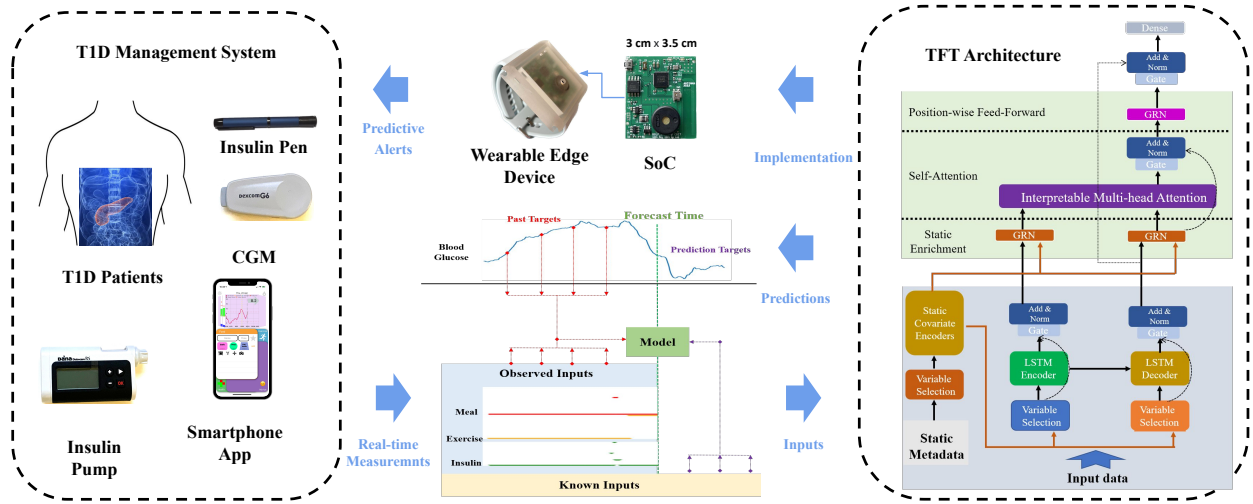
Fig. 1. Framework of a T1D management system, wearable edge device, and the architecture of the proposed TFT model.

prediction horizons (PHs), e.g., 30 and 60 minutes, which are inefficient for hardware implementation, especially when storage memory is limited. Multi-horizon models can output completed sequences but are difficult to train. Smartphone apps, based on iOS [18], [19] or Android [17], [21] operating systems, are the most common platform to deploy deep learning-based glucose predictors. Nevertheless, smartphone platforms suffer from the lack of wearability, battery issues, and frequent updates of operating systems [11]. To enable predictors to be operated on wearable and edge devices, pioneering work has implemented RNN models on a micro-controller unit [15] and a SoC [11]. In this work, we propose an edge-based TFT (E-TFT) model for BG prediction, as shown in Fig 1. After selecting essential input features through interpretability analysis, a TFT model is trained and further deployed on the SoC of a customized wearable edge device, which can fetch real-time CGM measurements and provide predictive warning of hypoglycemia and hyperglycemia, which has the great potential to be incorporated in other wearable devices, such as CGM transmitters and insulin pumps.

III. METHODOLOGY

A. Data Preparation

To enhance reproducibility and compare model performance, we use OhioT1DM dataset [22], a publicly available dataset that contains a variety of data fields collected in an eight-week clinical trial with 12 T1D adults. This dataset also served as a benchmark dataset in Blood Glucose Level Prediction (BGLP) Challenges [22]. The training and testing sets are provided separately, and the last 25% data of each training set were employed as a validation set. There are many missing values of CGM measurements, due to many practical reasons, such as signal loss and sensor calibration. To fill up the gaps without involving any future information, we apply linear extrapolation and clip values to the sensor operating range of 40-400 mg/dL.

B. Temporal Fusion Transformers

TFT can deal with mixed types of input data, as shown in the right part of Fig 1. In particular, CGM measurements are the target time series, while carbohydrates of meal intake, insulin bolus, and exercise are treated as exogenous time series inputs. We further incorporate the gender information as a static covariate and timestamps as future known time series. Overall, aiming to obtain multi-horizon prediction, TFT employs an encoder-decoder structure with the gated residual network (GRN) [10] as a basic block to determine the relationship between target time series and exogenous inputs. GRNs use exponential linear unit activation [23] and gated linear units [24] to control the nonlinear contributions of each input feature. The variable selection network (VSN) combines GRNs with a Softmax layer to assign variable selection weights for processed input features. Following the VSNs, LSTM layers are used in both the encoder and the decoder to extract the local patterns from past inputs and future known inputs, respectively.

Then the extracted temporal features are fed to upper modules (green area). First, a static enrichment layer uses GRNs to populate the impact of the static features. Then, a specifically designed multi-head attention layer, based on the self-attention mechanism [6], learns long-term dependencies while preserving global temporal interpretability, which is denoted as follows.

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^\mathsf{T}}{\sqrt{D_k}})\mathbf{V}, \qquad (1)$$

$$\mathbf{H}_i = Attention(\mathbf{Q}\mathbf{W}^Q, \mathbf{K}\mathbf{W}^K, \mathbf{V}\mathbf{W}^V), \qquad (2)$$

$$MultiHeadAttn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i]\mathbf{W}^O, \qquad (3)$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ denote queries, keys, values, respectively, associated with head-specific weights $\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V$; $\mathbf{H}_i$ stands for the output of i-th head out of a total of N heads;

D_k is the dimensions of keys, and $\mathbf{W}^O$ is a projection weight. Finally, a position-wise feed-forward network (PWFFN) is applied to obtain the output sequences, using a gated residual connection that skips all the upper modules.

C. Model Training and Edge Implementation

The weights of the TFT model are updated through back-propagation with an Adam optimizer to minimize a quantile loss [25]. We perform an early stopping technique to avoid overfitting, where a patience of 20 is applied to a total of 200 training epochs. To obtain mini-batches with a batch size of 128, we use a look-back sliding window to embed past 120-minute observations to predict future 60-minute BG sequences. Each input feature was scaled by standard normalization. After training a TFT model with all the available exogenous features, we perform feature selection using the variable selection weights of the encoder VSN as a post-hoc explanation method. Then, we obtain the weights of E-TFT by retraining TFT with the selected features, aiming to optimize the computational resources of the SoC.

To deploy E-TFT on the wristband for edge computing, we first decompose the model into matrix operations that explicitly describe the computations and logic of each layer. Then these matrix operations are translated from Python into Embedded C with optimized memory utilization, such as iterating matrix operations column by column, and pipelining interleaved processes. The involved SoC is based on Nordic nRF52832 that offers not only Bluetooth connectivity to acquire data from the CGM, but also efficient computation capability and memory space to perform real-time model inference. The device benefits from instant decision-making with very low latency since all processes are carried on locally. The performance of E-TFT is evaluated on the device through UART and USB ports.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

1) Baseline Methods: We consider a group of baseline methods in the experiments to evaluate the prediction performance. Classic machine learning predictors, including SVR [12] and LSTM [15], and advanced deep learning models, including dilated RNN (DRNN) [20], N-BEATS [26], and N-HiTS [27] are employed. Particularly, DRNN and N-BEATS achieved the best performance in the 2018 and 2020 BGLP challenges, respectively. Empowered by dilated connections, DRNN was designed to have a larger receptive field with fewer parameters and higher computational efficiency compared with standard RNNs. N-BEATS uses a deep stack of fully connected layers to process univariate inputs and showed promising performance in various time-series prediction problems. Based on N-BEATS, N-HiTS further introduces multi-rate sampling and hierarchical interpolation and achieved better accuracy than N-BEATS [27]. Among these, LSTM, DRNN, N-BEATS, and N-HiTS are adopted as multi-horizon predictors by modifying the output layers, while two single-horizon SVR models are trained for two PHs.
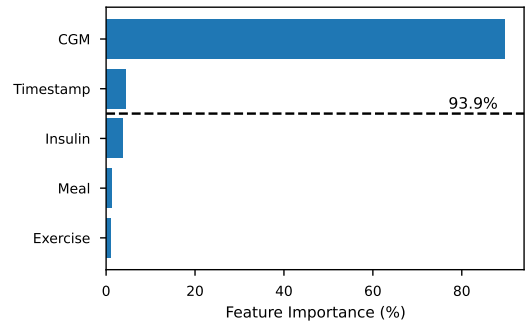

Fig. 2. Feature importance of the encoder input.

TABLE I
30-MINUTE PREDICTION PERFORMANCE OF THE CONSIDERED MODELS FOR THE OHIOT1DM DATASET

Model	RMSE (mg/dL)	MAE (mg/dL)	gRMSE (mg/dL)
SVR	23.35 ± 3.80	15.70 ± 2.82	29.05 ± 4.98
LSTM	20.22 ± 2.63	14.32 ± 1.67	24.61 ± 3.26
DRNN	19.58 ± 3.51	13.32 ± 2.25	24.03 ± 4.39
N-BEATS	19.43 ± 2.27	13.43 ± 1.56	23.84 ± 2.72
N-HiTS	19.63 ± 2.28	13.90 ± 1.56	24.47 ± 2.78
E-TFT	$\mathbf{19.09 \pm 2.47}$	$\mathbf{13.07 \pm 1.59}$	$\mathbf{23.25 \pm 3.09}$

2) Evaluation Metrics: In BG prediction, the most common statistical metrics are root mean square error (RMSE) and mean absolute error (MAE). To analyze the clinical effects, we incorporate glucose-specific RMSE (gRMSE) [28], which penalizes the predictions that would lead to adverse BG events, and Clark error grid (CEG) analysis [29].

B. Feature Importance

Fig. 2 shows the histogram of the feature importance for the encoder. It is noted that the CGM time series plays the most significant role in the input features of the encoder, and CGM and timestamps account for a total of 93.9% contribution. Therefore, we only use these two selected features and gender information when training the E-TFT model. In this instance, the number of weights is largely reduced, and thus much less Flash and SRAM are required on SOC. When compared with the model with exogenous time series, the mean RMSE of E-TFT was improved by 0.3 mg/dL for the validation sets. E-TFT can also avoid manual data entry errors, such as misestimation of carbohydrates.

C. Prediction Performance

Table I and II present the results (Mean±STD) of BG prediction. It is worth noting that E-TFT achieved the smallest RMSE, MAE, and gRMSE for both 30 and 60-minute PHs, while DRNN, N-BEATS, and N-HiTS also showed significant improvement when compared with conventional methods of SVR and LSTM. Fig. 3 depicts two-day ambulatory glucose profiles of the ground truth of BG levels and trajectories of 60-minute TFT predictions. The dotted blue and green lines show the upper and lowered bounds derived from the 25th percentile and 75th percentile of quantile forecasts, and the

| Model	RMSE (mg/dL)	MAE (mg/dL)	gRMSE (mg/dL)
SVR	35.47 ± 5.20	25.29 ± 4.35	45.71 ± 6.95
LSTM	34.35 ± 4.21	25.44 ± 3.14	42.68 ± 5.40
DRNN	33.24 ± 5.38	23.62 ± 4.11	41.07 ± 6.72
N-BEATS	33.82 ± 3.93	24.43 ± 2.98	42.97 ± 4.89
N-HiTS	33.00 ± 3.66	24.50 ± 2.85	42.46 ± 4.63
E-TFT	$\mathbf{32.31 \pm 3.79}$	$\mathbf{23.24 \pm 2.84}$	$\mathbf{40.90 \pm 4.93}$

shaded blue area refers to the interquartile range. The RMSE of subject 563 is close to the mean RMSE of the whole cohort, while subject 570 has the largest RMSE. Thus, we choose these subjects for visualization. We observe that the deviation between predictions and actual CGM measurements is small, and most adverse glycemic events were predicted. We successfully detected severe hypoglycemia and hyperglycemia that were missed by the point predictions (i.e., 50th percentile, the red curve), using the quantile upper and lower bounds. We also present the corresponding CEG analysis in Fig. 3. In particular, we achieved a percentage of 90.4%+8.8% for subject 563 and 87.1%+11.4% for subject 575 in the A+B zones. The predictions in these zones would not lead to inappropriate interventions.

D. Hardware Footprint

The final E-TFT model was implemented and deployed onto the target Bluetooth SoC, which occupies a total ROM space of 278.86 KB and requires 14.63 KB RAM memory for processing. The input of the encoder consists of past CGM and timestamps with a length of 24, while the decoder processes future timestamps with a length of 12 (i.e., 60-minute PH). The output with a dimension of seven refers to the 2nd, 10th, 25th, 50th, 75th, 90th, and 98th percentiles of quantile forecasts at each timestep. It is noted that the gated residual connection of the LSTM layer requires a considerable RAM space, 11.25 KB, to hold the data. Thus, it is recommended to statically allocate the memory space in advance to improve the system stability. Thanks to the on-chip DSPs that can perform high-performance floating-point arithmetic calculations, the ported E-TFT model can be computed within 1.9 seconds with the detailed computation time listed in Table III. In addition, by means of representing all weights in hexadecimal format represented by IEEE 754, the computation accuracy can be significantly improved. The RMSE between the outputs of the Python model and those of the C model is less than 1.5×10^{-5}.

V. DISCUSSION AND CONCLUSION

In this work, we propose E-TFT, a transformer-based deep learning model, for multi-horizon BG prediction, and implement it on a customized wristband to provide real-time decision support using edge computing. E-TFT exhibited excellent performance on the OhioT1DM dataset. It obtained the lowest RMSE, MAE, and gRMSE for the 30 and 60-minute PHs
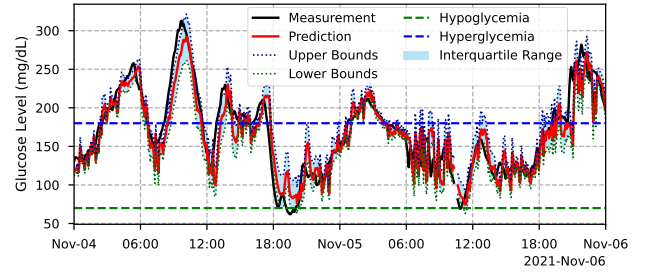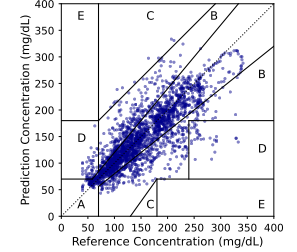

(a) Subject 563
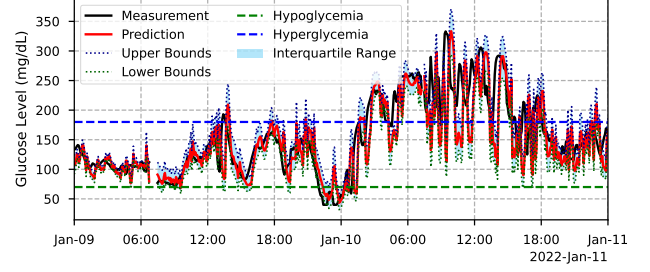
(b) Subject 575
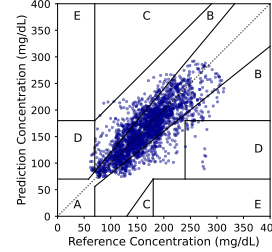
(c) Subject 563

(d) Subject 575

Fig. 3. Two-day visualization of CGM measurements and TFT prediction and CEG plots with the 60-minute PH.

Layer	Input Shape	Flash (B)	SRAM (B)	Time (ms)
Input	(2, 36)	0	240	N/A
Encoder	(2, 24)	11,448	14,976	142.87
Decoder	(1, 12)	5,592	3,384	17.29
LSTM	(36, 48)	169,728	12,104	1267.20
Attention	(36, 48)	21,952	10,760	227.99
PWFFN	(12, 48)	76,032	7,296	205.11
Output	(12, 7)	1,380	1,156	3.68

among all the considered methods. E-TFT has great potential to be directly implemented on CGM transmitters in the future collaboration with the manufacturer, since CGM can provide all the required data for real-time prediction. Future work also includes validating the wristband with the embedded E-TFT model in actual clinical trials or in T1D simulators to investigate clinical efficacy.

VI. ACKNOWLEDGEMENT

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] 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.

[3] D. Daneman, "Type 1 diabetes," *The Lancet*, vol. 367, no. 9513, pp. 847–858, 2006.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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*, pp. 1–1, 2022.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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*, pp. 1–12, 2022.

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

[21] 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*, vol. 24, no. 2, pp. 414–423, 2020.

[22] 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.

[23] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *International Conference on Learning Representations*, 2016.

[24] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[25] 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.

[26] 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.

[27] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza, M. Mergenthaler, and A. Dubrawski, "N-HiTS: Neural hierarchical interpolation for time series forecasting," *arXiv preprint arXiv:2201.12886*, 2022.

[28] 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.

[29] 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.