Computer Methods and Programs in Biomedicine 211 (2021) 106424

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study

Wonju Seo[a,1], Sung-Woon Park[b,1], Namho Kim[a], Sang-Man Jin[c,**], Sung-Min Park[a,d,e,*]

[a] Department of Convergence IT Engineering, Pohang University of Science and Technology, Republic of Korea
[b] Division of Endocrinology and Metabolism, Department of Medicine, CHA Gangnam Medical Center, CHA University, Republic of Korea
[c] Division of Endocrinology and Metabolism, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Republic of Korea
[d] Department of Electrical Engineering, Pohang University of Science and Technology, Republic of Korea
[e] Institute of Convergence Science, Yonsei University, Republic of Korea

ARTICLE INFO

Article history:
Received 5 May 2021
Accepted 14 September 2021

Keywords:
Diabetes
Deep neural network
Continuous glucose monitoring
Data-driven approach
Blood glucose management

ABSTRACT

Background: The accurate prediction of blood glucose (BG) level is still a challenge for diabetes management. This is due to various factors such as diet, personal physiological characteristics, stress, and activities influence changes in BG level. To develop an accurate BG level predictive model, we propose a personalized model based on a convolutional neural network (CNN) with a fine-tuning strategy.
Methods: We utilized continuous glucose monitoring (CGM) datasets from 1052 professional CGM sessions and split them into three groups according to type 1, type 2, and gestational diabetes mellitus (T1DM, T2DM, and GDM, respectively). During the preprocessing, only CGM data points were utilized, and future BG levels of four different prediction horizons (PHs, 15, 30, 45, and 60 min) were used as output. In training, we trained a general CNN and a multi-output random forest regressor using a hold-out method for each group. Next, we developed two personalized models: (1) by fine-tuning the general CNN on partial sample points of each CGM dataset, and (2) by learning a CNN from scratch on the points.
Results: For all groups, the fine-tuned CNN showed the lowest average root mean squared error, average mean absolute percentage error, highest average time gain (PH = 15 and 60 min in T1DM) and highest percentage in region *A* of Clarke error grid analysis at all PHs. In the performance comparison between the fine-tuned CNN and other models, we found that the fine-tuned CNN improved the performance of the general CNN in most cases and outperformed the scratch CNN at all PHs in all groups, making the fine-tuning strategy was useful for accurate BG level prediction. We analyzed all cases of four predictive patterns in each group, and found that the input BG level trend and the BG level at the time of prediction were related to the future BG level trend.
Conclusions: We demonstrated the efficacy of the fine-tuning method in a large number of CGM datasets and analyzed the four predictive patterns. Therefore, we believe that the proposed method will significantly contribute to the development of an accurate personalized model and the analysis for its predictions.

1. Introduction

Glucose monitoring to evaluate individual response to treatment and assess whether their glycemic targets are being safely attained is an integral part of diabetes management. Monitoring the blood glucose (BG) level in daily life has been achieved using self-monitoring blood glucose (SMBG) or continuous glucose monitoring (CGM). In contrast to the inconvenience of the finger prick blood sampling and discrete measurement of SMBG, CGM offers glucose levels measured every one to five minutes through a glucose sensor inserted into the subcutaneous tissue. People with diabetes using real-time CGM can take action in a timely manner

* Corresponding author at: Department of Convergence IT Engineering, Pohang University of Science and Technology, Republic of Korea.
** Corresponding author at: Division of Endocrinology and Metabolism, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Republic of Korea.
E-mail addresses: tjdnjswn22@postech.ac.kr (W. Seo), skagh1597@postech.ac.kr (N. Kim), sangman.jin@samsung.com (S.-M. Jin), sungminpark@postech.ac.kr (S.-M. Park).
[1] These author contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2021.106424
0169-2607/

in response to changes in the measured BG level, and clinicians can provide appropriate care by analyzing their BG level trends. This benefit of CGM has been translated into improved clinical outcomes such as a significant reduction in HbA1c [1,2], and a significant decrease in hypoglycemia [3,4], which resulted in a rapidly expanding clinical use of real-time CGM in type 1 diabetes [5]. Furthermore, prediction of future BG level or hypoglycemia [6] using CGM data has already revolutionized the clinical management of diabetes through various types of devices such as real-time CGM devices equipped with predictive alarms, sensor-augmented insulin pumps with predictive suspend feature, and hybrid closed-loop devices. Improved accuracy of BG level prediction is one of the key elements for better clinical utility of such devices, and an increasing body of studies have been conducted to predict future BG levels with better accuracy [7,8].

With recent advancements in data-driven techniques, machine learning, and time series approaches, people have picked up a keen interest in BG level prediction. This is because a BG level prediction model can be developed from mainly CGM data points without complicated physiological modeling. In these approaches, time series model is fitted to CGM data points and then it predicts the future BG level using the fitted parameters. Sparacino et al. [9] proposed the first-order autoregressive (AR) and the first-order polynomial models to predict the future BG level. With a 30 min prediction horizon (PH), they presented that hypo/hyperglycemia BG threshold levels (i.e., 70 mg/dL and 180 mg/dL, respectively) could be predicted more than 20–25 min in advance. Recently, Yang et al. [10] proposed an autoregressive integrated moving-average (ARIMA) model with continuous update of its order and coefficients. The proposed model outperformed an adaptive univariate model and ARIMA model with a fixed order and showed 6.27% of relative absolute deviation (RAD) for patients with type 1 diabetes and 5.41% of RAD for patients with type 2 diabetes. Another method of the data-driven approach is to utilize machine learning models, particularly neural networks. Perez-Gandia et al. [11] proposed an artificial neural network (ANN) using five consecutive CGM values and the time of prediction, and the ANN outperformed an AR model. Some studies proposed ANNs with meal information [12], insulin information [13], or information on both meals and insulin [13] as inputs to improve their predictive performances. Furthermore, a Bayesian regularized neural network using glycemia-related features such as heart rate, and sleeping time was proposed to model the dynamics of BG [14].

Recently, attempts [15–18] have been made to use a deep neural network (DNN) because it can learn more complex nonlinearities than conventional machine learning models. Li et al. [15] proposed a convolutional recurrent neural network to predict a BG level with 30 min PH and showed 21.07 mg/dL of RMSE for using clinical datasets. The same group [18] proposed a dilated convolution neural network (D-CNN) which increases the size of the receptive field without heavy computational costs and slightly improved the performance to 19.28 mg/dL of RMSE. Both proposed models outperformed baseline models such as a neural network, a support vector regression, the latent variable model, and autoregressive model. Mhaskar et al. [16] proposed a DNN based on the domain knowledge of the prediction error-grid analysis (PRED-EGA) [19] and demonstrated the DNN's superiority over a shallow ANN on average PRED-EGA scores. Lastly, Aliberti et al. [17] proposed a non-linear AR neural network and a long-short term memory (LSTM) network as a multi-patient data-driven approach. While the non-linear AR network showed a good predictive performance within 30 min PH, the LSTM network showed the good performance with longer PH.

There are two types of prediction models: a general model that is trained from the general CGM datasets and a personalized model that is trained from personal CGM datasets. As the number of accumulated data points increases, the personalized model learns more precise individual-specific glucose patterns and physiological characteristics; however, the model is weak and easily over-fitted when the number of accumulated data points is small. On the contrary, the general model has a good generalization performance, but it is difficult to make predictions that fit different BG level patterns for each patient. To utilize both models' advantages, the general DNN is trained on large numbers of CGM datasets and then the DNN is tuned on a new CGM dataset. The method is called a fine-tuning and it is powerful because the model can learn a specific BG level pattern of the new CGM dataset while maintaining the learned general features.

This paper presents a method of developing a personalized model with a fine-tuning strategy. The study makes three main contributions. We demonstrated the personalized model on a large CGM datasets ($n = 894$), including three types of diabetes and results of each group were statistically analyzed and compared. Next, we developed the personalized model using only CGM data points. The model does not require any additional burdensome inputs such as carbohydrates in a meal and insulin doses. Finally, four predictive patterns were visualized and analyzed in each group. We confirmed that the input BG level trend and the BG level at the time of prediction were related to the future BG level trend.

2. Materials and methods

The protocol for this study was approved by the Institutional Review Board (IRB) of the Samsung Medical Center (IRB file No. 2020–01–018–001). The need for informed consent was waived by the board.

2.1. Materials

2.1.1. Clinical dataset acquisition

We screened all the patients ($n = 1114$) who visited the endocrinology clinic of an academic tertiary care facility (Samsung Medical Center, Seoul, Korea) and underwent professional CGM sessions for any purpose. The CGMS GOLD™ (Medtronic MiniMed, Northridge, CA, USA) was used for the three-day professional CGM sessions. Details of the methods used in the professional CGM sessions have been described elsewhere [20,21].

The CGMS GOLD™ (Medtronic MiniMed) is a retrospective, professional, and blinded CGM device, in which the CGM data points were calibrated by entered BG points obtained using SMBG, and then their noise was removed by a finite impulse response filter. The filtering algorithms of the CGMS GOLD™ (Medtronic MiniMed) were described by Keenan and associates [22]. To ensure the accuracy of the devices (median relative absolute difference ~10%) [23,24], all the study participants inserted the CGM sensors into the subcutaneous tissue in the abdomen, and were instructed to measure capillary BG with finger stick test more than three times per day for calibration. Each CGM dataset was derived from each professional CGM session by extracting CGM data points, which was unchanged from the raw BG data recorded by CGMS GOLD™ (Medtronic MiniMed).

From 1296 professional CGM sessions collected between 2009 and 2016 at the Samsung Medical Center, we selected CGM datasets derived from the 1052 professional CGM sessions obtained from people with diabetes, after exclusion of the CGM datasets meeting the following exclusion criteria: First, patients without type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM), and gestational diabetes mellitus (GDM) were excluded (1075 CGM sessions from 894 patients were left). Next, the CGM datasets in which at least 10% of data points included 1) an undefined number (i.e., NaN), 2) 2.2 mmol/L (i.e., lower limit), and 3) 22.2 mmol/L (i.e., upper limit) were excluded from further analyses (1053 CGM

Table 1
Baseline characteristics (n = 894).

Characteristic	Value
Age, yr	55.0 ± 13.4
Sex, male/female	508 (56.8%)/386 (43.2%)
BMI, kg/m^2	24.95 ± 8.87
Duration of diabetes, y (n = 878)	12.62 ± 8.75
Chronic kidney disease (eGFR <60 mL/min/1.73 m^2) (n = 874)	156 (17.8%)
HbA1c (n = 870),%	8.13 (2.47)
C-peptide (n = 842), ng/mL	1.87 (1.48)
Type of diabetes	
Type 1 diabetes mellitus	141 (15.77)
Type 2 diabetes mellitus	733 (81.99)
Gestational diabetes mellitus	20 (2.24)
Sessions per patient	1.18
Lipid medication	388 (43.40)
Hypertension medication	353 (39.49)

Values are presented as mean ± standard deviation or number (%).
BMI, body mass index; HbA1c, glycosylated hemoglobin; eGFR, estimated glomerular filtration rate.

sessions from 894 patients were left). Among two CGM datasets extracted from one patient, one CGM dataset contained the same CGM data points of the other CGM dataset. In this case, the short CGM dataset was excluded (1052 CGM sessions from 894 patients were left). We then used a cubic interpolation to substitute missing CGM points. If the interpolated CGM point exceeded the lower or the upper limit, it was clipped to the lower limit or the upper limit. The interpolated points were used in training and validation, but not in testing. Finally, we multiplied all CGM data points by 18.018 to convert mmol/L to mg/dL.

We collected clinical information to include age, sex, body mass index, duration of diabetes (years), type of diabetes (T1DM, T2DM, or GDM), and medication for dyslipidemia and hypertension. HbA1c, estimated glomerular filtration rate (eGFR) and c-peptide were measured within one month, three months and three years, respectively, before the CGM data were also collected. The characteristics of participants included in this study are summarized in Table 1. Among the total subjects, 56.8% were male, and the mean age and BMI were 55 years and 24.95 kg/m^2, respectively. In vast majority of the participants (801 of 894), CGM data were obtained in outpatient setting. The mean HbA1c measured within one month prior to CGM and mean duration of diabetes were 8.13% and 12.62 years, respectively, of the three types of diabetes included, T2DM was the most common with 81.99%, T1DM with 15.77%, and GDM with 2.24%. There were 43.40% and 39.49% of subjects taking lipid and hypertension medication, respectively.

2.2. Methods

2.2.1. Data preprocessing

Each CGM dataset derived from each professional CGM session is represented by following equation:

$$CGM_{1:N} = [CGM_t | t = 1, \ldots, N] \quad (1)$$

where CGM_t is a CGM data point at time of 5 × t min and N is the total length of the CGM dataset. For training a model in a supervized learning scheme, inputs and outputs should be defined. The input was defined as a vector that consists of previous 16 consecutive *sCGM* data points [18]. $sCGM_t$ is a scaled CGM_t using a min-max scaler.

$$\mathrm{sCGM_t} = \frac{CGM_t - CGM_{min}}{CGM_{max} - CGM_{min}} \quad (2)$$

where CGM_{min} is a minimum BG level and CGM_{max} is a maximum BG level in a training set. The output was defined as the difference between $sCGM_t$ and $sCGM_{t+i}$, where i is either 3, 6, 9, or 12. The input and the output of each sample point are represented as follows:

$$x_t = [sCGM_{t-15},\ sCGM_{t-14}, \ldots, sCGM_{t-1}, sCGM_t],$$

$$\Delta sCGM_{t+i} = sCGM_{t+i} - sCGM_t,\ i = 3,\ 6,\ 9,\ 12 \quad (3)$$

After prediction, a future scaled CGM data point was restored as follows:

$$\widehat{sCGM}_{t+i} = sCGM_t + \widehat{\Delta sCGM}_{t+i},\ i = 3,\ 6,\ 9,\ 12 \quad (4)$$

where $\widehat{sCGM}_{t+i}$ is the estimated scaled CGM data point at t with (5 × i) min PH. Then, the predicted scaled CGM data point with Eq. (4) was converted to the predicted CGM data point by a reversed min-max scaler.

2.2.2. Network design

We selected a VGG-like CNN (Fig. 1) for the following reasons: (1) training and deciding with CNN are faster than with LSTM due to parallel computation; (2) a VGG structure [25] is well-known in the computer vision field and has the advantage of using a small number of parameters by increasing the receptive field by stacking several small-sized filters; and (3) a recent study showed the CNN's superiority than LSTM on BG level prediction [26]. Although more complex models can be implemented, performance may be negligible when complex models containing only CGM data points [27].

Convolution and max-pooling layers served to extract local features and dense layers modeled a regression model with the extracted features. The output dense layer had four units with a linear activation function, and the loss function was a mean squared error. As hyper-parameters, the number of filters (i.e., t), the unit of dense layers (i.e., d), the activation function, the learning rate, batch size, and epoch were considered. Adam was used as the optimizer. According to the activation function, He normal or Xavier normal initializer was used to initialize weights of each layer. Each unit represents the prediction with 15, 30, 45, and 60 min PH, respectively.

2.2.3. Training and validation

The advantage of using a DNN compared to conventional machine learning models is that the DNN can transfer its learned weights to other DNNs after pre-training. For example, in the field of medical imaging, an accurate model is developed by using weights learned from natural images [28]. The transfer learning strategy depends on the size of the new data and the similarity between the old task and the new task. A fine-tuning strategy in transfer learning is to tune the weights of all layers of a model to the new task. The goal of this study was to train the model trained with CGM datasets to fit a new CGM dataset. Thus, the tasks are the same, and the pre-trained weights are considered good weight initialization for the personalized model; hence, we reused the weights for the new CGM dataset. The entire training and evaluation processes are shown in Fig. 2.

We divided the entire CGM datasets into three groups (i.e., T1DM, T2DM, and GDM) according to the type of diabetes. The reason for grouping the dataset was that the dynamics of BG are different for each group. The pathophysiology underlying hyperglycemia such as altered insulin secretion and sensitivity is markedly different between the types of diabetes [29], and glucose response patterns substantially vary according to an individual's insulin secretion rate and resistance [30]. In each group, we used a hold-out method, in which the patients were divided into 80% for training and validation, and 20% for testing.

As shown in Fig. 2a, CGM datasets extracted from the training set were used for the general CNN training, and hyper-parameters

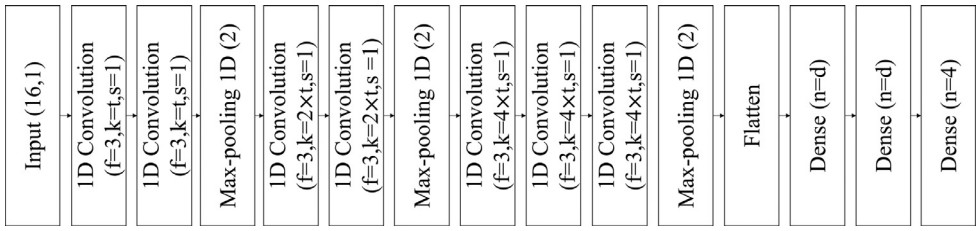

Fig. 1. The proposed CNN structure that is based on a VGG network. In the convolution layers, zero-padding was used. *f* means the filter size, *k* means the number of filters, *s* means the size of stride, and *n* is the unit of *a* dense layer. *t* and *d* are tunable hyper-parameters.

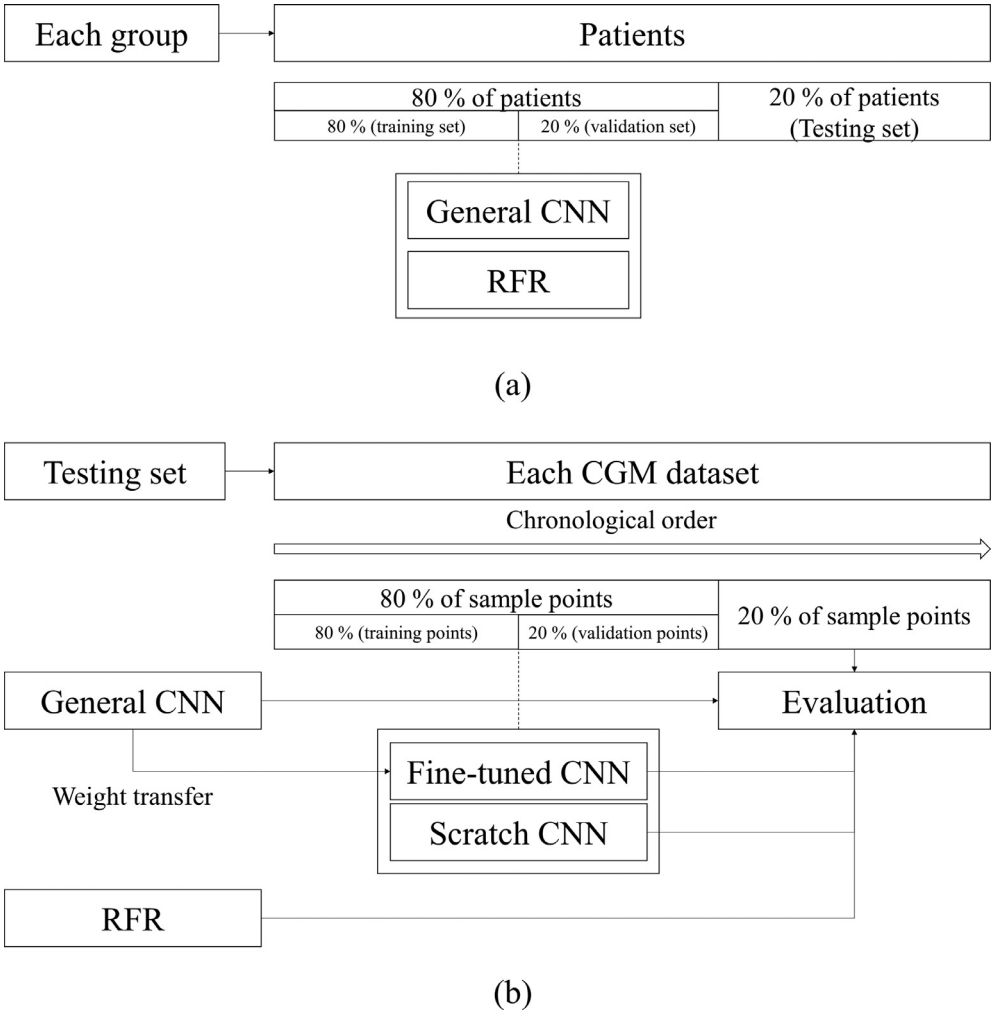

Fig. 2. The total training and evaluation process. (a) General CNN and RFR training process in each group. CGM datasets extracted from the training set were used for training and CGM datasets extracted from the validation set were used as validation for hyper-parameter tuning and calculating the validation loss. (b) In the chronological order of each CGM dataset in the testing set, the training sample points were used for training and the validation sample points were used as validation for early stopping and calculating the validation loss. The remaining sample points were used for evaluation.

were tuned using CGM datasets extracted from the validation set. We used the Bayesian optimizer of Keras tuner for up to 50 trials to tune hyper-parameters and early stopping with five patience steps during tuning was used to speed up the tuning process. After finding the optimized hyper-parameters, the optimized general CNN was trained on CGM datasets extracted from the training set without early stopping. As a baseline model, we developed a multi-output random forest regressor (RFR). The RFR was recently used for predicting BG levels [31] and was performed better than a baseline AR model [32]. We tuned hyper-parameters of RFR using Bayesian optimization in the same method as the general CNN.

Next, we fine-tuned the general CNN (fine-tuned CNN) for each CGM dataset and trained a CNN from scratch (scratch CNN). As shown in Fig. 2b, the training sample points extracted from each CGM dataset were used for training, and the validation sample points were used for early stopping and calculating the validation loss. During the training process, the epoch and the batch size were the same with the values optimized through the tuning process, and an early stop method was used with five steps of patience. The optimized learning rate was utilized when training the scratch CNN, while for fine-tuning the general CNN, a tenth of the optimal learning rate was used to prevent forgetting the learned

features. The dense layers were only fine-tuned because the convolutional layers served as local feature extractors.

2.2.4. Evaluation metrics

To evaluate the performance of the models, we used two main performance evaluators. The first is a numerical performance evaluator, including RMSE, mean absolute percentage error (MAPE), and time gain (TG):

$$RMSE = \sqrt{\frac{1}{M}\sum_{t=1}^{M}\left(y_t - \hat{y}_t\right)^2} \quad (5)$$

$$MAPE = \frac{1}{M}\sum_{t=1}^{M}\frac{\left|y_t - \hat{y}_t\right|}{y_t} \times 100 \quad (6)$$

$$TG_i = 5 \times \left(i - argmax_\tau\left(\hat{y}_{1+\tau:M-i+\tau} \star y_{1:M-i}\right)\right),\ i = 3, 6, 9, 12 \quad (7)$$

where M is the number of all predicted BG level, y_t is an actual BG level, and $\hat{y}_t$ is the predicted BG level matching y_t. RMSE and MAPE are numerical metrics related to the accuracy of prediction, and TG is the effective PH and it is calculated by obtaining a time shift having the highest cross correlation coefficient between the original series and the predicted series shifted by τ. These metrics were calculated by each CGM dataset. The second is a clinical evaluation performance known as the Clarke-Error grid analysis (CEGA) [33]. In the CEGA, regions A and B are considered as clinically acceptable zones and a high percentage of region A is preferable. This was calculated by using all predicted BG levels and all original BG levels. We calculated all metrics for each PH.

2.2.5. Analysis of prediction patterns

Instead of predicting a single future BG level, predicting multiple BG levels provides information about the future BG level trend. The trend of BG is shown as increasing or decreasing and determined by the sign of the slope of the linear regression model of continuous BG level. The combination of the input BG level trend and the future BG level trend has four predictive patterns. After visualizing four representative patterns by group, we analyzed each pattern based on the input BG level trend and the BG level at the time of prediction.

2.2.6. Tools

To develop all the models, we used Tensorflow (Version 2.1), Keras (Version 2.3.1), Keras tuner (Version 1.0.1), Scipy (Version 1.4.1), and Sklearn (Version 0.23.2) libraries using Python (Version 3.6.10). The CEGA was evaluated with R studio (Version 3.4.4) ('getClarkeZones' function). The Wilcoxon signed-rank test was used to compare the metrics of the fine-tuned CNN with those of other models. Mann–Whitney U test was used to compare statistics between groups or analyze predictive patterns. Statistical significance was considered when $p < 0.05$.

3. Results

3.1. Numerical performance

The average numerical results on the clinical datasets were calculated and summarized in Table 2. The metrics of other models were compared with the metrics of the fine-tuned CNN.

For T1DM, the fine-tuned CNN had the lowest average RMSE and MAPE at all PHs, and the highest average TG (PH = 15, and 60 min). The general CNN exhibited the highest average TG (PH = 15, 30, and 45 min). When comparing the fine-tuned CNN with the scratch CNN, there were significant differences for TG at all PHs (Wilcoxon signed rank test). When comparing the fine-tuned CNN with the RFR, there was significant difference for TG (PH = 30 min, Wilcoxon signed rank test). For T2DM, the fine-tuned CNN exhibited the lowest average RMSE, MAPE, and the highest average TG at all PHs. The general CNN showed the highest average TG (PH = 60 min). When comparing the fine-tuned CNN with the scratch CNN, there were significant differences for TG at all PHs, and for RMSE and MAPE (PH = 15 min, Wilcoxon signed-rank test). When comparing the fine-tuned CNN with the RFR, there were significant difference for TG at all PHs (Wilcoxon signed-rank test). For GDM, the fine-tuned CNN showed the lowest average RMSE, MAPE, and the highest average TG at all PHs. The general CNN showed the highest average TG (PH = 30 and 45 min).

When comparing the metrics of the fine-tuned CNN between groups, the average RMSE of T2DM was lower than that of T1DM at all PHs, and the average MAPE and TG of T2DM were better than those of T1DM (PH = 15, 30, and 45 min). The average RMSE of GDM was lower than that of T1DM (PH = 15 min), and the average MAPE and TG of GDM were better than those of T1DM at all PHs. Finally, the average RMSE of T2DM was lower than that of GDM at all PHs, and the average MAPE and TG of GDM were better than those of T2DM at all PHs.

3.2. Clinical performance

Next, the clinical performance of each model was evaluated and summarized in Table 3.

In all groups, the fine-tuned CNN showed the highest percentage in region A at all PHs and the general CNN showed the highest percentage in region A at PH = 15 min in GDM. When comparing the percentage in region A of the fine-tuned CNN between groups, the percentage in region A of T2DM was higher than that of T1DM (PH = 15 and 30 min). The percentage in region A of GDM was higher than that of T1DM (PH = 15 and 45 min). Finally, the percentage in region A of GDM was higher than that of T2DM (PH = 15, 45, and 60 min). All model showed the high percentage in regions A and B; > 96% for T1DM, > 96% for T2DM, and > 99% for GDM.

3.3. Analysis of prediction patterns

We visualized four predictive patterns made by the fine-tuned CNNs in each group (Fig. 3).

When the input BG level increased overall, the future BG level increased/decreased overall (Fig. 3a, b, e, f, i, j). In each group, all cases of these two predictive patterns were analyzed. In T1DM, the average BG level at the time of prediction was 140.0 (± 53.7) mg/dL for the increased future BG level and 237.7 (± 76.4) mg/dL for the decreased future BG level. In T2DM, the average BG level at the time of prediction was 137.3 (± 47.6) mg/dL for the increased future BG level and 218.7 (± 70.1) mg/dL for the decreased future BG level. In GDM, the average BG level at the time of prediction was 126.7 (± 38.7) mg/dL for the increased future BG level and 219.6 (± 79.1) mg/dL for the decreased future BG level. There were significant differences for the BG level at the time of prediction between the two predictive patterns in all groups (Mann–Whitney U test).

On the contrary, when the input BG level decreased overall, the future BG level decreased/increased (Fig. 3c, d, g, h, k, l). In each group, all cases of these two predictive patterns were analyzed. In T1DM, the average BG level at the time of prediction was 198.7 (± 64.7) mg/dL for the decreased future BG level and 119.8 (± 43.9) mg/dL for the increased future BG level. In T2DM, the average BG level at the time of prediction was 184.7 (± 63.1) mg/dL for the decreased future BG level and 113.0 (± 37.0) mg/dL for the increased future BG level. Finally, in GDM, the average BG

Table 2
Average (SD) of RMSE, MAPE, and TG on the testing set of each group.

Group	Models	Average RMSE (mg/dL)				Average MAPE (%)				Average TG* (min)			
		15 min	30 min	45 min	60 min	15 min	30 min	45 min	60 min	15 min	30 min	45 min	60 min
T1DM	Fine-tuned	10.9	17.8	23.3	28.1	5.3	8.7	11.6	14.2	2.1	4.2	5.3	6.8
($n = 29$)	CNN	(4.97)	(7.73)	(10.13)	(12.62)	(3.06)	(4.55)	(5.66)	(6.86)	(2.8)	(4.35)	(4.83)	(5.56)
	General	11.0	18.2	24.0	29.1	5.4	9.0	12.2	15.0	2.1	4.4	5.8	6.5
	CNN	(5.0)	(7.71)	(9.87)	(12.24)	(3.12)	(4.68)	(5.87)	(7.04)	(2.8)	(4.29)	(5.17)	(5.66)
	Scratch	11.7	18.7	24.1	28.6	5.7	9.3	12.2	14.8	0.3†	0.8†	1.8†	2.3†
	CNN	(5.16)	(8.31)	(11.01)	(13.46)	(2.96)	(4.53)	(5.92)	(7.31)	(1.21)	(2.21)	(3.5)	(4.52)
	RFR	11.4	18.6	24.3	29.2	5.6	9.2	12.2	14.9	1.2	2.4†	3.0	4.1
		(5.32)	(8.37)	(10.86)	(13.22)	(3.22)	(5.01)	(6.2)	(7.32)	(2.18)	(3.34)	(4.13)	(4.91)
T2DM	Fine-tuned	10.5	17.2	22.9	27.7	5.2	8.6	11.5	14.2	2.6	4.6	5.7	6.7
($n = 147$)	CNN	(4.98)	(7.59)	(10.0)	(12.22)	(2.51)	(4.03)	(5.47)	(7.08)	(2.93)	(4.31)	(4.99)	(5.79)
	General	10.5	17.4	23.3	28.3	5.2	8.8	12.0	14.9	2.6	4.5	5.7	6.7
	CNN	(4.96)	(7.53)	(9.89)	(12.01)	(2.58)	(4.24)	(5.74)	(7.45)	(2.88)	(4.36)	(4.82)	(5.69)
	Scratch	11.7†	18.7	24.4	29.1	5.8†	9.3	12.2	14.8	0.3†	1.1†	1.6†	2.5†
	CNN	(5.22)	(8.43)	(11.18)	(13.66)	(2.59)	(4.27)	(5.74)	(7.42)	(1.27)	(2.42)	(3.41)	(4.63)
	RFR	10.8	17.7	23.6	28.6	5.4	9.0	12.2	15.1	1.6†	3.2†	4.3†	5.1†
		(5.02)	(7.67)	(10.1)	(12.31)	(2.61)	(4.29)	(5.81)	(7.41)	(2.36)	(3.41)	(4.08)	(4.59)
GDM	Fine-tuned	10.8	18.6	24.7	29.7	4.6	7.9	10.7	12.9	3.8	5.0	6.3	7.5
($n = 4$)	CNN	(4.64)	(6.76)	(7.95)	(9.21)	(1.1)	(2.1)	(3.32)	(4.44)	(2.5)	(4.08)	(4.79)	(6.45)
	General	11.1	19.1	25.7	31.1	4.9	8.4	11.6	14.7	2.5	5.0	6.3	6.3
	CNN	(4.6)	(6.77)	(8.08)	(9.34)	(0.67)	(1.54)	(2.26)	(3.01)	(2.89)	(4.08)	(4.79)	(4.79)
	Scratch	12.4	20.3	26.0	31.5	5.7	9.5	12.2	15.4	0.0	2.5	2.5	3.8
	CNN	(4.31)	(6.09)	(6.72)	(8.61)	(2.33)	(3.95)	(5.78)	(7.98)	(0.0)	(2.89)	(2.89)	(4.79)
	RFR	11.8	20.0	26.8	32.6	5.3	9.4	13.3	16.7	0.0	1.3	2.5	3.8
		(4.16)	(5.92)	(7.33)	(9.08)	(0.81)	(1.78)	(3.03)	(3.95)	(0.0)	(2.5)	(2.89)	(2.5)

SD, standard deviation; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; GDM, gestational diabetes mellitus, RMSE, root mean squared error; MAPE, mean absolute percentage error; TG, time gain; CNN, convolutional neural networks; RFR, random forest regressor
†$p < 0.05$, when compared to the fine-tuned CNN.
* TG was calculated after filling the missing values in the predicted time series with a cubic interpolation.

Table 3
The percentage of regions *A* and *B* in CEGA on the testing set.

Group	Models	*A* zone (%)				*B* zone (%)			
		15 min	30 min	45 min	60 min	15 min	30 min	45 min	60 min
T1DM	Fine-tuned CNN	96.7	90.6	84.1	78.2	2.8	8.2	13.9	19.4
($n = 29$)	General CNN	96.4	89.7	83.0	76.6	2.9	8.8	14.7	20.3
	Scratch CNN	96.1	89.5	82.1	76.5	3.2	9.1	16.0	21.0
	RFR	96.1	89.1	82.6	76.5	3.1	9.2	15.1	20.5
T2DM	Fine-tuned CNN	97.2	90.9	83.8	77.6	2.5	8.0	14.2	19.3
($n = 147$)	General CNN	97.1	90.3	82.8	75.8	2.5	8.4	14.8	20.6
	Scratch CNN	96.8	89.4	81.7	75.3	3.0	9.6	16.3	22.0
	RFR	97.0	89.9	82.3	75.4	2.7	8.8	15.2	21.0
GDM	Fine-tuned CNN	98.1	90.4	84.2	78.1	1.9	9.6	15.5	21.8
($n = 4$)	General CNN	98.1	89.5	82.1	76.9	1.9	10.5	17.6	22.3
	Scratch CNN	96.0	86.7	76.6	68.6	4.0	13.3	23.4	31.2
	RFR	97.4	89.0	79.1	69.3	2.6	11.0	20.6	30.2

T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; GDM, gestational diabetes mellitus, CEGA, Clarke's error grid analysis; CNN, convolutional neural networks; RFR, random forest regressor.

level at the time of prediction was 162.2 (± 73.3) mg/dL for the decreased future BG level and 129.9 (± 48.1) mg/dL for the increased future BG level. There were significant differences for the BG level at the time of prediction between the two predictive patterns in all groups (Mann–Whitney U test).

4. Discussion

The accurate and timely prediction of a future BG level is undoubtedly needed for further advancement of the diabetes management technologies. In this study, we proposed a method of developing a personalized model with the fine-tuning strategy. We validated the developed model using 1052 CGM datasets ($n = 894$) from professional CGM sessions. In summary, our results showed: (1) the fine-tuned CNN showed the lowest average RMSE, MAPE, and highest TG (PH = 15 and 60 min in T1DM) and the highest percentage in region *A* at all PHs in all groups. In most cases, the fine-tuning strategy improved the metrics of the general CNN; (2) the fine-tuned CNN performed on average better than the scratch CNN and showed significance differences for RMSE, MAPE (PH = 15 min in T2DM), and TG (all PHs in T1DM and T2DM); (3) after visualizing the four predictive patterns and analyzing all cases of the predictive patterns in each group, and we found that the input BG level trend and the BG level at the time of prediction were important in determining the future BG level trend.

In most cases, the general CNN showed a better average numerical performance than the RFR. In T1DM, the RFR showed the lower average MAPE (PH = 60 min). In the clinical performance, the general CNN showed a higher percentage in region *A* then the RFR at all PHs in all groups. From this comparison, we concluded that the general CNN is better than the baseline machine learning model. When comparing the fine-tuned CNN with the scratch CNN, the fine-tuned CNN showed the lower average RMSE, MAPE, and the higher average TG than the scratch CNN at all PHs in all groups. In the clinical performance, the fine-tuned CNN showed the higher percentage in region *A* than the scratch CNN at all PHs in all groups. Based on this comparison, we concluded that the fine-tuning strategy, which utilized the general features of the gen-

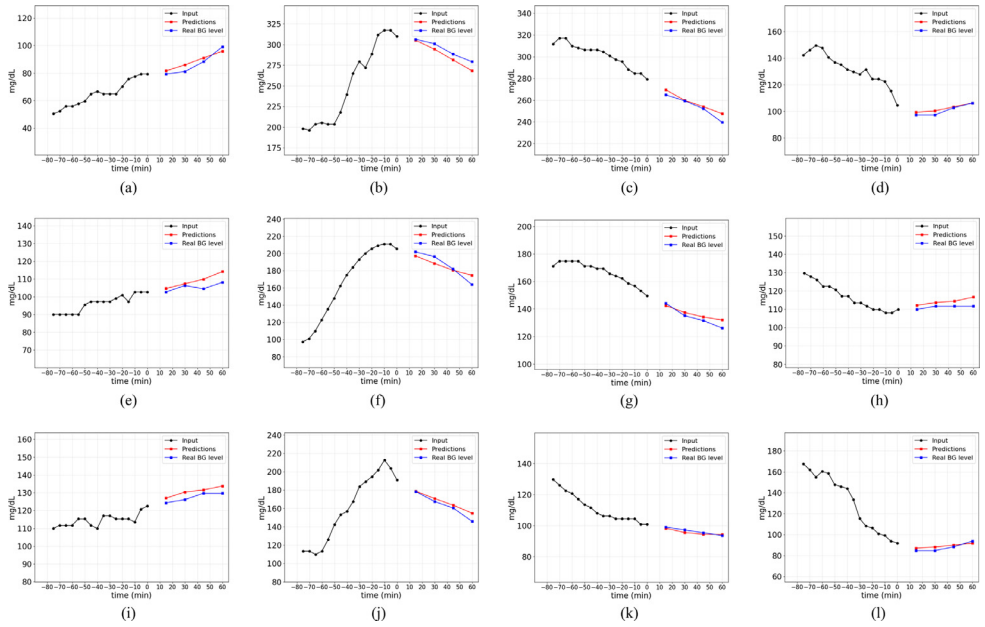

Fig. 3. Representative future BG level trends in each group. The circled black line indicates input CGM data points, the squared red line indicates its prediction, and the squared blue line indicates real BG level. The first row's figures were obtained from T1DM, the second row's figures from T2DM and the third row's figures from GDM. When the input BG level increased overall, the future BG level increased/decreased overall (a, b, e, f, i, j). When the input BG level decreased overall, the future BG level decreased/increased overall (c, d, g, h, k, l).

eral CNN, was superior to learning from scratch. When comparing the fine-tuned CNN with the general CNN, the fine-tuned CNN improved the numerical performance of the general CNN in most cases. In the clinical performance, the fine-tuned CNN showed the higher or similar percentage in region *A* than the general CNN at all PHs in all groups. However, there was no significant difference in the numerical metrics between the general CNN and the fine-tuned CNN. We guess that the use of only CGM data points for fine-tuning has a limitation. Therefore, to increase the effectiveness of the fine-tuning, future studies using additional inputs are necessary.

Compared with recent studies using DNNs [15,17,18], we used a much larger CGM dataset, and our study included people with T1DM, T2DM, and GDM (Table 1). This means that our model may be appropriate for use in real-world diabetes management. Some models [15,18] used various inputs, but our method relied on only CGM data points; so, our method can relieve the patient of burdens such as calculating carbohydrate in a meal or entering insulin doses. Finally, the fine-tuned CNN showed better RMSE and MAPE on real patients with T1DM than other models [15,17,18] and the fine-tuned CNN showed higher percentages in regions *A* [17] in the first condition, so we concluded that the fine-tuned CNN is competitive. However, it showed much higher TG than the fine-tuned CNN. This is because they used various inputs such as insulin bolus and meal data [13]. These factors provide information about the rise or fall of BG levels, allowing the model to make better prediction.

We visualized the four representative predictive patterns and analyzed all cases of the predictive patterns in each group (Fig. 3). Remarkably, the same patterns were identified in all three groups, indicating that there exist shared characteristics. When the input BG level increased/decreased overall, the future BG level increased/decreased overall (Fig. 3a, c, e, g, i, k); this result can be estimated by a linear regression. However, there were opposite cases not explained by the linear regression (Fig. 3b, d, f, h, j, l). The pattern that the input BG level increased/decreased overall and the future BG level decreased/increased overall appeared at the high/low BG level at the time of prediction. These predictions can improve RMSE, MAPE, and TG around the peak/nadir, helping to respond quickly to the decrease/increase of BG level after BG level peak/nadir.

This study has several limitations. First, we analyzed data obtained from the professional CGM sessions in a single tertiary care facility. The results of the current study should be cautiously extrapolated to primary care settings. Second, the professional CGM devices used in this study, CGMS GOLD^{TM} (Medtronic MiniMed), retrospectively calibrated and filtered BG level at the end of the monitoring [22], which contrasts with real-time CGM. Although the accuracy of the devices in hyperglycemia (median relative absolute difference ~10%) has been reported to be comparable to current CGM devices, their accuracy in hypoglycemia has been reported to be lower than the latest real-time CGM devices [23,24]. Third, the number of samples of GDM ($n = 20$) was smaller than T1DM and T2DM. Finally, the CGM data used in this study may not represent real-life conditions including an extreme glycemic fluctuation such as meal perturbations, vigorous activities, and hypoglycemic events. Although the vast majority (~90%) of the CGM data were obtained from an outpatient setting with the study participants blinded to their CGM data, the study participants may not represent the general diabetes population in terms of degree of diabetes education and adherence to lifestyle instructions. Also, the 3 day CGM sessions might not have represented their long-term real-life conditions. Therefore, the proposed method should be validated by further clinical studies in a setting closer to real-world environments.

5. Conclusion

We proposed a method for developing a personalized model with fine-tuning and demonstrated its efficacy on large CGM datasets to include three types of diabetes. Our results showed that the fine-tuned CNN showed the lowest average RMSE, MAPE, highest TG (PH = 15 and 60 min in T1DM), and the highest percentage in region *A* at all PHs in all groups. In the performance comparison between the fine-tuned CNN and other models, the fine-tuned CNN improved the performance of the general CNN in most cases

and outperformed the scratch CNN at all PHs in all groups. From these results, we concluded that the fine-tuning strategy was useful for accurate BG level prediction. We analyzed all cases of the four predictive patterns in each group and found that the input BG level trend and the BG level at the time of prediction were important in determining the future BG level trend. We believe that our method and results will be useful for building the personalized model and interpreting its prediction.

Declaration of Competing Interest

The authors declare that there are no financial interests or financial conflicts with the subject matter discussed in this study.

Ethics statement

The protocol of this study was approved by the Institutional Review Board (IRB) of the Samsung Medical Center (IRB file No. 2020–01–018–001). The need for informed consent was waived by the board.

Data availability

Availability of data and materials. The data that support the findings of this study are available from Samsung Medical Center but restrictions apply to the availability of these data. The data were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Samsung Medical Center.

Acknowledgments

This study was supported by the Ministry of Science and ICT (MSIT), Korea under the ICT Consilience Creative program (IITP-2020–2011–1–00783) supervised by the Institute for Information and communications Technology Promotion (IITP), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A5A1015596), the Technology Innovation Program (or Industrial Strategic Technology Development Program, 20001841, Development of System for Intelligent ContextAware Wearable Service based on Machine Learning) funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea), and a National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIT) (No. 2020R1A2C2005385).

References

[1] R.W. Beck, T. Riddlesworth, K. Ruedy, A. Ahmann, R. Bergenstal, S. Haller, C. Kollman, D. Kruger, J.B. McGill, W. Polonsky, E. Toschi, H. Wolpert, D. Price, Effect of continuous glucose monitoring on glycemic control in adults with type 1 diabetes using insulin injections: the DIAMOND randomized clinical trial, JAMA 317 (2017) 371–378.

[2] M. Lind, W. Polonsky, I.B. Hirsch, T. Heise, J. Bolinder, S. Dahlqvist, E. Schwarz, A.F. Ólafsdóttir, A. Frid, H. Wedel, E. Ahlén, T. Nyström, J. Hellman, Continuous glucose monitoring vs conventional therapy for glycemic control in adults with type 1 diabetes treated with multiple daily insulin injections: the GOLD randomized clinical trial, JAMA 317 (2017) 379–387.

[3] J. Bolinder, R. Antuna, P. Geelhoed-Duijvestijn, J. Kröger, R. Weitgasser, Novel glucose-sensing technology and hypoglycaemia in type 1 diabetes: a multicenter, non-masked, randomized controlled trial, Lancet 388 (2016) 2254–2263.

[4] N. Hermanns, B. Schumann, B. Kulzer, T. Haak, The impact of continuous glucose monitoring on low interstitial glucose values and low blood glucose values assessed by point-of-care blood glucose meters: results of a crossover trial, J. Diabetes Sci. Technol. 8 (2014) 516–522.

[5] N.C. Foster, R.W. Beck, K.M. Miller, M.A. Clements, M.R. Rickels, L.A. DiMeglio, D.M. Maahs, W.V. Tamborlane, R. Bergenstal, E. Smith, B.A. Olson, S.K. Garg, State of type 1 diabetes management and outcomes from the T1D exchange in 2016-2018, Diabetes Technol. Ther. 21 (2019) 66–72.

[6] W. Seo, Y.B. Lee, S. Lee, S.M. Jin, S.M. Park, A machine-learning approach to predict postprandial hypoglycemia, BMC Med. Inf. Decis. Mak. 19 (2019) 210.

[7] S. Oviedo, J. Vehí, R. Calm, J. Armengol, A review of personalized blood glucose prediction strategies for T1DM patients, Int. J Numer. Method Biomed. Eng. 33 (2017) e2833.

[8] A.Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, G. Hartvigsen, Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes, Artif. Intell. Med. (2019).

[9] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, C. Cobelli, Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series, IEEE Trans. Biomed. Eng. 54 (2007) 931–937.

[10] J. Yang, L. Li, Y. Shi, X. Xie, An ARIMA model with adaptive orders for predicting blood glucose concentrations and hypoglycemia, IEEE J. Biomed. Health Inf. 23 (2018) 1251–1260.

[11] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gómez, M. Rigla, A. de Leiva, M. Hernando, Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring, Diabetes Technol. Ther. 12 (2010) 81–88.

[12] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, C. Cobelli, Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration, IEEE Trans. Biomed. Eng. 59 (2012) 1550–1560.

[13] C. Zecchin, A. Facchinetti, G. Sparacino, C. Cobelli, How much is short-term glucose prediction in type 1 diabetes improved by adding insulin delivery and meal content information to CGM data? A proof-of-concept study, J. Diabetes Sci. Technol. 10 (2016) 1149–1160.

[14] I. Rodríguez-Rodríguez, J.V. Rodríguez, J.M. Molina-García-Pardo, M.Á. Zamora-Izquierdo, M.T.M.I.I. Martínez-Inglés, A comparison of different models of glycemia dynamics for improved type 1 diabetes mellitus management with advanced intelligent analysis in an internet of things context, Appl. Sci. 10 (2020) 4381.

[15] K. Li, J. Daniels, C. Liu, P. Herrero, P. Georgiou, Convolutional recurrent neural networks for glucose prediction, IEEE J. Biomed. Health Inf. 24 (2019) 603–613.

[16] H.N. Mhaskar, S.V. Pereverzyev, M.D. van der Walt, A deep learning approach to diabetic blood glucose prediction, Front. App. Math. Stat. 3 (2017) 14.

[17] A. Aliberti, I. Pupillo, S. Terna, E. Macii, S. Di Cataldo, E. Patti, A. Acquaviva, A multi-patient data-driven approach to blood glucose prediction, IEEE Access 7 (2019) 69311–69325.

[18] K. Li, C. Liu, T. Zhu, P. Herrero, P. Georgiou, GluNet: a deep learning framework for accurate glucose forecasting, IEEE J. Biomed. Health Inf. 24 (2019) 414–423.

[19] M. Kiviniemi, R. Hermann, J. Nurmi, A.G. Ziegler, M. Knip, O. Simell, R. Veijola, T. Lövgren, J. Ilonen, T.S. Group, A high-throughput population screening system for the estimation of genetic risk for type 1 diabetes: an application for the TEDDY (the environmental determinants of diabetes in the young) study, Diabetes Technol. Ther. 9 (2007) 460–472.

[20] J.E. Jun, S.E. Lee, Y.B. Lee, J.Y. Ahn, G. Kim, K.Y. Hur, M.K. Lee, S.M. Jin, J.H. Kim, Continuous glucose monitoring defined glucose variability is associated with cardiovascular autonomic neuropathy in type 1 diabetes, Diabetes Metab. Res. Rev. 35 (2019) e3092.

[21] J.H. Yoo, M.S. Choi, J. Ahn, S.W. Park, Y. Kim, K.Y. Hur, S.M. Jin, G. Kim, J.H. Kim, Association between continuous glucose monitoring-derived time in range, other core metrics, and albuminuria in type 2 diabetes, Diabetes Technol. Ther. 22 (2020) 768–776.

[22] D.B. Keenan, J.J. Mastrototaro, G. Voskanyan, G.M. Steil, Delays in minimally invasive continuous glucose monitoring devices: a review of current technology, J. Diabetes Sci. Technol. 3 (2009) 1207–1214.

[23] D.R.i.C.N.S. Group, The accuracy of the CGMS™ in children with type 1 diabetes: results of the Diabetes Research in Children Network (DirecNet) accuracy study, Diabetes Technol. Ther. 5 (2003) 781.

[24] M.J. Tansey, R.W. Beck, B.A. Buckingham, N. Mauras, R. Fiallo-Scharer, D. Xing, C. Killman, W.V. Tamborlane, K.J. Ruedy, Accuracy of the modified Continuous Glucose Monitoring System (CGMS) sensor in an outpatient setting: results from a diabetes research in children network (DirecNet) study, Diabetes Technol. Ther. 7 (2005) 109–114.

[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, (2014).

[26] T. El Idrissi, A. Idri, Deep learning for blood glucose prediction: CNN vs LSTM, in: Proceedings of the International Conference on Computational Science and Its Applications, Springer, 2020, pp. 379–393.

[27] N.S. Tyler, P.G. Jacobs, Artificial intelligence in decision support systems for type 1 diabetes, Sensors 20 (2020) 3214.

[28] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.

[29] F. Zaccardi, D.R. Webb, T. Yates, M.J. Davies, Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective, Postgrad. Med. J. 92 (2016) 63–69.

[30] H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin, M. Snyder, Glucotypes reveal new patterns of glucose dysregulation, PLoS Biol. 16 (2018) e2005143.

[31] I. Rodríguez-Rodríguez, I. Chatzigiannakis, J.V. Rodríguez, M. Maranghi, M. Gentili, M.Á. Zamora-Izquierdo, Utility of big data in predicting short-term blood glucose levels in type 1 diabetes mellitus through machine learning techniques, Sensors 19 (2019) 4482.

[32] Y. Amar, S. Shilo, T. Oron, E. Amar, M. Phillip, E. Segal, Clinically accurate prediction of glucose levels in patients with type 1 diabetes, Diabetes Technol. Ther. 22 (2020) 562–569.

[33] W.L. Clarke, D. Cox, L.A. Gonder-Frederick, W. Carter, S.L. Pohl, Evaluating clinical accuracy of systems for self-monitoring of blood glucose, Diabetes Care 10 (1987) 622–628.