scientific reports

OPEN

A multimodal deep learning architecture for predicting interstitial glucose for effective type 2 diabetes management

Muhammad Salman Haleem[1,2]✉, Daphne Katsarou[3], Eleni I. Georga[3], George E. Dafoulas[4], Alexandra Bargiota[5], Laura Lopez-Perez[6], Miguel Rujas[6], Giuseppe Fico[6], Leandro Pecchia[7], Dimitrios Fotiadis[3] & Gatekeeper Consortium

The accurate prediction of blood glucose is critical for the effective management of diabetes. Modern continuous glucose monitoring (CGM) technology enables real-time acquisition of interstitial glucose concentrations, which can be calibrated against blood glucose measurements. However, a key challenge in the effective management of type 2 diabetes lies in forecasting critical events driven by glucose variability. While recent advances in deep learning enable modeling of temporal patterns in glucose fluctuations, most of the existing methods rely on unimodal inputs and fail to account for individual physiological differences that influence interstitial glucose dynamics. These limitations highlight the need for multimodal approaches that integrate additional personalized physiological information. One of the primary reasons for multimodal approaches not being widely studied in this field is the bottleneck associated with the availability of subjects' health records. In this paper, we propose a multimodal approach trained on sequences of CGM values and enriched with physiological context derived from health records of 40 individuals with type 2 diabetes. The CGM time series were processed using a stacked Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (BiLSTM) network followed by an attention mechanism. The BiLSTM learned long-term temporal dependencies, while the CNN captured local sequential features. Physiological heterogeneity was incorporated through a separate pipeline of neural networks that processed baseline health records and was later fused with the CGM modeling stream. To validate our model, we utilized CGM values of 30 min sampled with a moving window of 5 min to predict the CGM values with a prediction horizon of (a) 15 min, (b) 30 min, and (c) 60 min. We achieved the multimodal architecture prediction results with Mean Absolute Point Error (MAPE) between 14 and 24 mg/dL, 19–22 mg/dL, 25–26 mg/dL in case of Menarini sensor and 6–11 mg/dL, 9–14 mg/dL, 12–18 mg/dL in case of Abbot sensor for 15, 30 and 60 min prediction horizon respectively. The results suggested that the proposed multimodal model achieved higher prediction accuracy compared to unimodal approaches; with upto 96.7% prediction accuracy; supporting its potential as a generalizable solution for interstitial glucose prediction and personalized management in the type 2 diabetes population.

Keywords Multimodal AI, Deep learning, Interstitial glucose prediction, Time series modelling

Type 2 Diabetes Mellitus (T2DM) is characterized by insulin resistance, leading to elevated blood glucose levels, and it accounts for approximately 90% of all diagnosed cases of diabetes[1]. Individuals with T2DM are at 15% higher risk of mortality[2] as the International Diabetes Federation estimated 537 million people were affected, causing 6.7 million deaths in 2021[3] and projected to rise up to 783 million by 2045. This not only impacts

[1]School of Engineering, University of Warwick, Coventry CV4 7AL, UK. [2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK. [3]Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece. [4]Faculty of Medicine, University of Thessaly, Volos, Greece. [5]Department of Endocrinology and Metabolic Diseases, University Hospital of Larisa, Larissa, Greece. [6]Universidad Politécnica de Madrid-Life Supporting Technologies Research Group, ETSIT, Madrid, Spain. [7]Università Campus Bio-Medico, Via Álvaro del Portillo, 21, 00128 Roma, Italy. ✉email: salman.haleem@warwick.ac.uk; m.haleem@qmul.ac.uk

population health but also imposes a heavy financial strain on both individuals and the global healthcare system, as the American Diabetes Association reported that the total cost of diagnosed diabetes in the United States was 412.9 billion USD in 2022[4], with a 35% increase in medical costs over the past decade[5]. Efforts to mitigate these costs focus on early detection, effective management strategies, and preventive measures to reduce the incidence and severity of diabetes complications[5].

Multiple studies have shown that self-monitoring of blood glucose is effective in supporting diabetes management[6]. With the advent of modern wearables and technologies, one study identified high compliance with regular monitoring of blood glucose and other T2DM variables (e.g. diet, physical activity) among individuals using smartphones compared to those using paper diaries[7]. The regular self-monitoring of blood glucose can promote adherence to clinical guidelines for diet and physical activity, resulting in improvements in hemoglobin A1c (HbA1c) levels[8]. Traditionally, regular blood glucose monitoring requires a finger prick test, which is invasive and cumbersome[9]. In contrast, Continuous Glucose Monitoring (CGM) measures the concentration of glucose in the interstitial fluid at regular intervals[10]. While CGM is well established in type 1 diabetes care, its use in T2D is expanding. The ADA Standards of Care in Diabetes 2025 recommend CGM for adults with T2D on glucose-lowering therapies, reflecting its growing role in managing glycemic variability[11]. CGM use in T2D has been associated with reduced risk of severe hypoglycemia, diabetic ketoacidosis, and hospitalizations. Developing accurate glucose predictive models for this population is therefore timely and clinically relevant.

Regular acquisition of blood glucose data holds significant potential for predicting future glucose levels and improving glycaemic control[12]. It also enables the estimation of critical glycaemic events, such as hypoglycaemia (defined as blood glucose levels below 70 mg/dL) and hyperglycaemia (above 180 mg/dL)[13]. However, there are certain challenges associated with predicting blood glucose via CGM values only. Firstly, there is a proven 10-min sensor delay between interstitial fluid glucose and actual blood glucose as measured by CGM[14]. Secondly, CGM systems are susceptible to occasional sensor failure or signal loss, and therefore require reliable strategies to ensure continuity of glucose monitoring during these periods[15,16].

Advances in artificial intelligence (AI), including traditional machine learning and deep learning techniques, have enabled the development of models that predict interstitial glucose levels 15 to 60 min in advance based on historical CGM-derived glucose readings[17,18]. However, clinical studies suggest the patient-specific differences in glycaemic variability are possible due to underlying conditions (e.g., demographics, comorbidities, or diet plans)[19] and currently, predictive models do not inform CGM variations based on these underlying conditions.

To address the aforementioned challenge, in this study, glucose levels were monitored using a continuous glucose monitoring (CGM) device, which measures glucose concentration in the interstitial fluid via the subcutaneous tissue. We investigated a multimodal deep learning approach to estimate short-term interstitial glucose levels in individuals with T2D in real-life conditions by informing continuous glucose monitoring (CGM) data with baseline health conditions. The multimodal learning approach involves a sequential deep learning pipeline trained on CGM sequences and context information, while baseline health data serve as auxiliary knowledge to inform CGM variations, which are then combined via a multimodal fusion function. The overall workflow of our architecture has been presented in Fig. 1. The details are presented in the upcoming sections according to the TRIPOD statement as tabulated in Supplementary Table 1.

Results

Participants

Table 1 presents the set of input variables along with the characteristics examined in the present predictive modelling study for the 40 subjects. In particular, 15 out of 40 patients used the GlucoMen Day CGM Menarini® sensor (we call Sensor 1) for a monitoring period of 10–19 days, and the remaining 25 patients used the Libre Abbott® system (we call Sensor 2) for a monitoring period of 8–28 days. This has served as initial step for Ambulatory Glucose Profile (AGP) analysis.

Based on AGP analysis, the time spent in clinically defined glucose ranges varied across participants. On average, participants spent approximately 2.68% of the time in low range (glucose < 70 mg/dL) and 0.64% in very low range (glucose < 54 mg/dL), as shown in Fig. 2. Time spent in the high range (glucose > 180 mg/dL) was approximately 20%, and time spent in the very high range (glucose > 250 mg/dL) was approximately 5%. The mean interstitial glucose across the cohort was 146.1 ± 22.98 mg/dL (see Table 2). The results were extracted prior to data curation and preprocessing. Also, the Augmented Dickey-Fuller (ADF) confirmed that the CGM time series for each patient was stationary over the observation period.

CGM prediction

Table 3 reports the performance of the unimodal prediction pipeline for a 15-min prediction horizon, evaluated using Mean Absolute Percentage Error (MAPE)[20]. The results indicate that emphasizing local CGM features, via weighting mechanisms informed by high CGM contextual variability, enhances prediction accuracy. The Convolutional Neural Network (CNN) driven Long Short Term Memory (LSTM) i.e. CNN-LSTM model with attention achieved the lowest MAPE across both sensors, demonstrating the benefit of incorporating temporal dynamics and adaptive focus on high-variability regions in glucose trends. The statistical significance among difference in MAPE among different architectures of CGM pipeline has been performed via *T test*[21] and has been presented in Table 4. The results show significant improvement towards adding LSTM and attention mechanism. However, our experiments also suggest that adding more complex layers (e.g. adding multilayer convolutional layers) will add further complexity in the architecture; resulting in decreased performance of the model.

Comparison between multimodal and unimodal architectures

After the CGM pipeline development, we then developed the multimodal architecture by performing additive concatenation between the CGM pipeline and baseline pipeline trained via fully connected dense architecture.

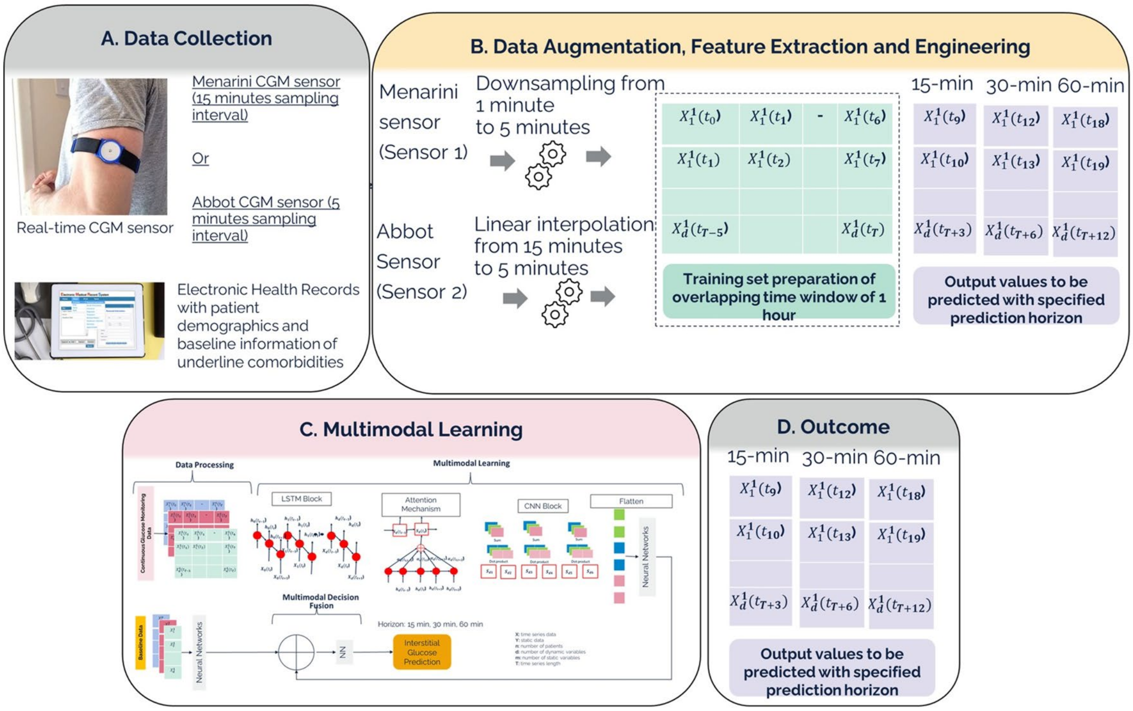

A. Data Collection
Menarini CGM sensor (15 minutes sampling interval)
Or
Abbot CGM sensor (5 minutes sampling interval)
Real-time CGM sensor
Electronic Health Records with patient demographics and baseline information of underline comorbidities
B. Data Augmentation, Feature Extraction and Engineering
Menarini sensor (Sensor 1)
Downsampling from 1 minute to 5 minutes
Abbot Sensor (Sensor 2)
Linear interpolation from 15 minutes to 5 minutes
$X_1^1(t_0)$ $X_1^1(t_1)$ - $X_1^1(t_6)$
$X_1^1(t_1)$ $X_1^1(t_2)$ $X_1^1(t_7)$
$X_d^1(t_{T-5})$ $X_d^1(t_T)$
Training set preparation of overlapping time window of 1 hour
15-min 30-min 60-min
$X_1^1(t_9)$ $X_1^1(t_{12})$ $X_1^1(t_{18})$
$X_1^1(t_{10})$ $X_1^1(t_{13})$ $X_1^1(t_{19})$
$X_d^1(t_{T+3})$ $X_d^1(t_{T+6})$ $X_d^1(t_{T+12})$
Output values to be predicted with specified prediction horizon
C. Multimodal Learning
Data Processing
Continuous Glucose Monitoring Data
Multimodal Learning
LSTM Block
Attention Mechanism
CNN Block
Flatten
Sum
Dot product
Neural Networks
Multimodal Decision Fusion
Baseline Data
Neural Networks
NN
Horizon: 15 min, 30 min, 60 min
Interstitial Glucose Prediction
X: time series data
Y: static data
n: number of patients
d: number of dynamic variables
m: number of static variables
T: time series length
D. Outcome
15-min 30-min 60-min
$X_1^1(t_9)$ $X_1^1(t_{12})$ $X_1^1(t_{18})$
$X_1^1(t_{10})$ $X_1^1(t_{13})$ $X_1^1(t_{19})$
$X_d^1(t_{T+3})$ $X_d^1(t_{T+6})$ $X_d^1(t_{T+12})$
Output values to be predicted with specified prediction horizon

Fig. 1. Overall architecture of predicting the continuous interstitial glucose via multimodal architecture.

	Feature	Mean ± std
Demographics	Gender (%)	Male: 45 (18)
		Female: 55 (22)
	Age (years)	67 ± 9
Anthropometrics	Weight (kg)	80.42 ± 28.46
	Height (m)	1.63 ± 0.12
	Waist circumference (cm)	104.13 ± 16.5
Biochemical tests	Baseline Blood Glucose (mg/dL)	136.6 ± 45.19
	Baseline HbA1c (%)	7.42 ± 1.11
	Creatinine (mg/dL)	1.99 ± 1.52
	Urea Level (mg/dL)	49.44 ± 30.65
	Total cholesterol (mg/dL)	144.82 ± 33.78
	LDL cholesterol (mg/dL)	67.17 ± 32.11
	HDL cholesterol (mg/dL)	46.86 ± 10.83
	Triglycerides (mg/dL)	203.71 ± 247.43
	White blood cell count (10^3/μL)	7.29 ± 1.83
	Red blood cell count (10^3/μL)	25.26 ± 37.27
	Haematocrit (%)	39.76 ± 8.86
	Plt (× 1000/μL)	206.31 ± 80.53
	SGOT (IU/L)	35.42 ± 32.22
	SGPT (IU/L)	25.99 ± 14.12
	K (mmol/L)	4.57 ± 0.48
	Na (mmol/L)	125.52 ± 42.24

Table 1. Descriptive characteristics of Central Greece pilot study. Plt, Platelet Count; SGOT, Serum Glutamic-Oxaloacetic Transaminase; SGPT, Serum Glutamic-Pyruvic Transaminase; HbA1c, Haemoglobin A1c.

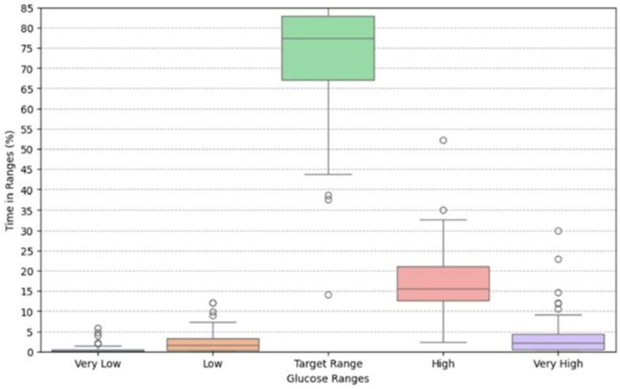


Fig. 2. The time in ranges based on AGP report.

AGP report variables	Mean ± standard deviation
Average glucose ($mgdL^{-1}$)	146.1 ± 22.98%
Glucose management indicator (%)	6.8 ± 0.55%
Glucose variability (%)	31.85 ± 7.36%

Table 2. The glucose statistics based on AGP report across all patients using CGM data over the entire period of the study (Data are presented as mean ± standard deviation).

	Menarini sensor MAPE	Abbott sensor MAPE
CNN	14.61 ± 18.98	7.22 ± 8.89
LSTM	14.62 ± 19.91	6.89 ± 9.15
CNN + LSTM	14.51 ± 20.06	7.04 ± 9.25
CNN + LSTM + Attention	14.24 ± 19.42	6.80 ± 9.31

Table 3. CGM population model comparison for predicting interstitial glucose with 15-min prediction horizon.

	CNN	LSTM	CNN + LSTM	CNN + LSTM + Attention
Menarini sensor (Sensor 1) statistical significance of difference				
CNN				*
LSTM			*	*
CNN + LSTM		*		*
CNN + LSTM + Attention	*	*	*	
Abbot sensor (Sensor 2) statistical significance of difference				
CNN		***	***	***
LSTM	***		***	***
CNN + LSTM	***	***		***
CNN + LSTM + Attention	***	***	***	

Table 4. The statistical significance of difference among different CGM pipelines mentioned in Table 3. *p_value < 0.05; **p_value < 0.01; ***p_value < 0.001.

The reason for opting for simpler architecture is because the performance of the complex multimodal architecture was constrained by the availability of baseline variables. Including more baseline variables reduced the number of subjects available for training, which in turn impacted model performance. Through our experiments, we have achieved 7 sets of baseline variables which are present with respective number of patient numbers acquired through both Menarini and Abbot sensors. The results have been presented in Table 5.

The comparison between unimodal and multimodal architectures has been presented in terms of MAPE as shown in Figs. 3 and 4. We have also tabulated these results not only in terms of overall MAPE, but also

Set	Variables	Menarini subjects (Sensor 1)	Abbot subjects (Sensor 2)
0	[‘Age’, ‘Gender’]	15	25
1	[‘Age’, ‘Gender’, 'HbA1c (%)']	8	17
2	[‘Age’, ‘Gender’, 'HbA1c (%)', 'Weight (kg)', 'Height (m)']	6	13
3	[‘Age’, ‘Gender’, 'HbA1c (%)', 'Weight (kg)', 'Height (m)', 'HDL cholesterol (mg/dL)', 'Total cholesterol (mg/dL)']	5	9
4	[‘Age’, ‘Gender’, 'HbA1c (%)', 'Weight (kg)', 'Height (m)', 'HDL cholesterol (mg/dL)', 'Total cholesterol (mg/dL)', 'Blood Glucose (mg/dL)', 'Urea level (mg/dL)']	5	7
5	[‘Age’, ‘Gender’, 'HbA1c (%)', 'Weight (kg)', 'Height (m)', 'HDL cholesterol (mg/dL)', 'Total cholesterol (mg/dL)', 'Blood Glucose (mg/dL)', 'Urea level (mg/dL)', 'K (mmol/L)', 'Haematocrit (%)', 'LDL cholesterol (mg/dL)']	3	5
6	[‘Age’, ‘Gender’, 'HbA1c (%)', 'Weight (kg)', 'Height (m)', 'HDL cholesterol (mg/dL)', 'Total cholesterol (mg/dL)', 'Blood Glucose (mg/dL)', 'Urea level (mg/dL)', 'K (mmol/L)', 'Haematocrit (%)', 'LDL cholesterol (mg/dL)', 'SGOT (IU/L)', 'White blood cell count (10^3/μL)']	2	4

Table 5. List and number of baseline and demographic variables present for both Menarini and Abbot subjects.

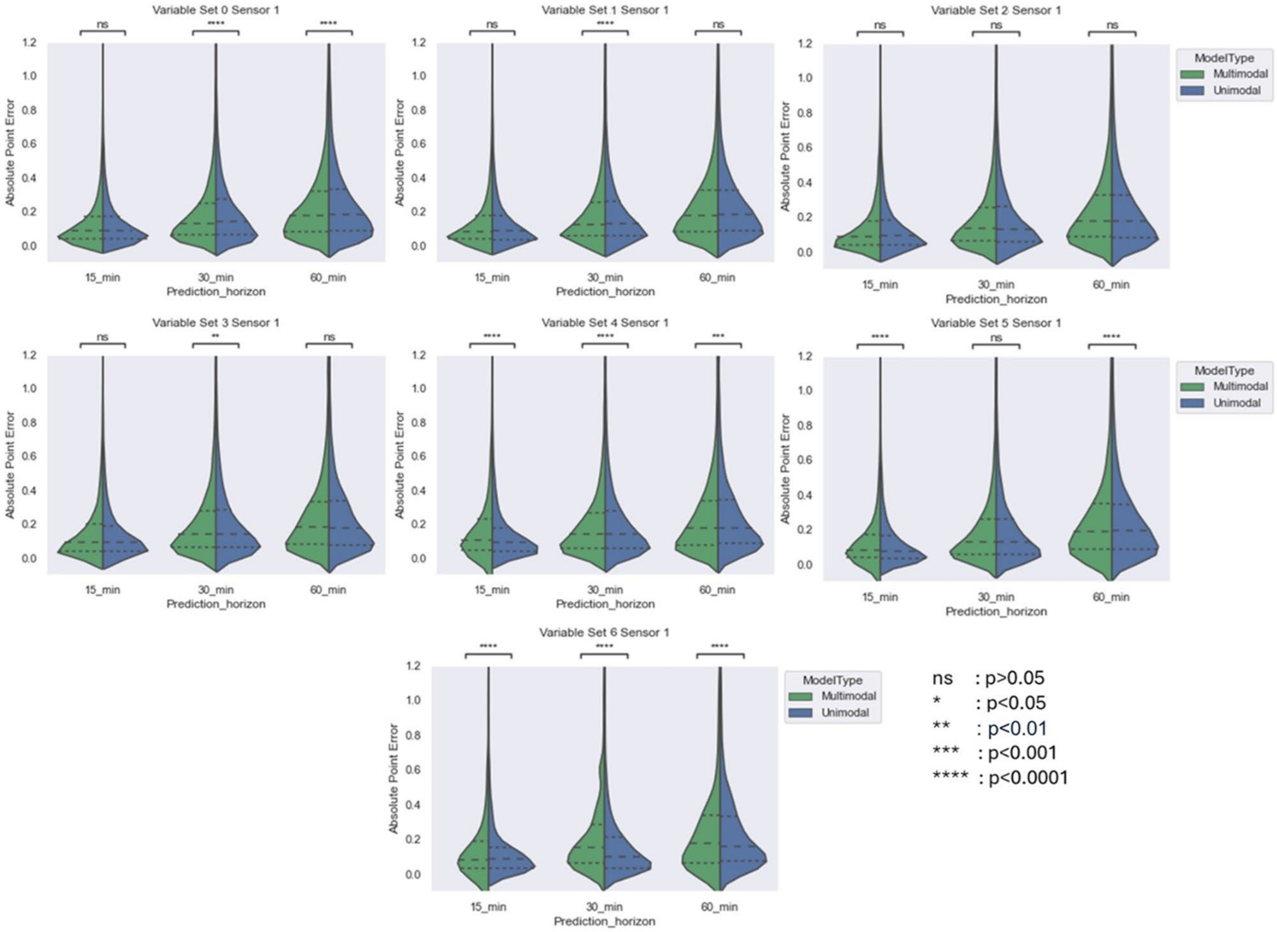


Fig. 3. Comparing violin plot of absolute point error for multimodal and unimodal architectures developed for Menarini sensor across different variable sets at three prediction horizon. The violin plot shows the distribution of absolute point error 25%, 50% and 75% quartile via dashed line.

in terms of Hyperglycaemic MAPE (where acquired interstitial glucose was greater than 180 mg/dL) as well as Hypoglycaemic MAPE (where acquired interstitial glucose was less than 70 mg/dL). The results have been presented in Table 6. Due to low number of Hypoglycaemic events in Type 2 diabetic subjects acquired from Abbot sensors, the Hypoglycaemic MAPE for these subjects have not been calculated. Besides, due to smaller number of participants for set 5 and onwards, their MAPE results were not stable and therefore not included in the results.

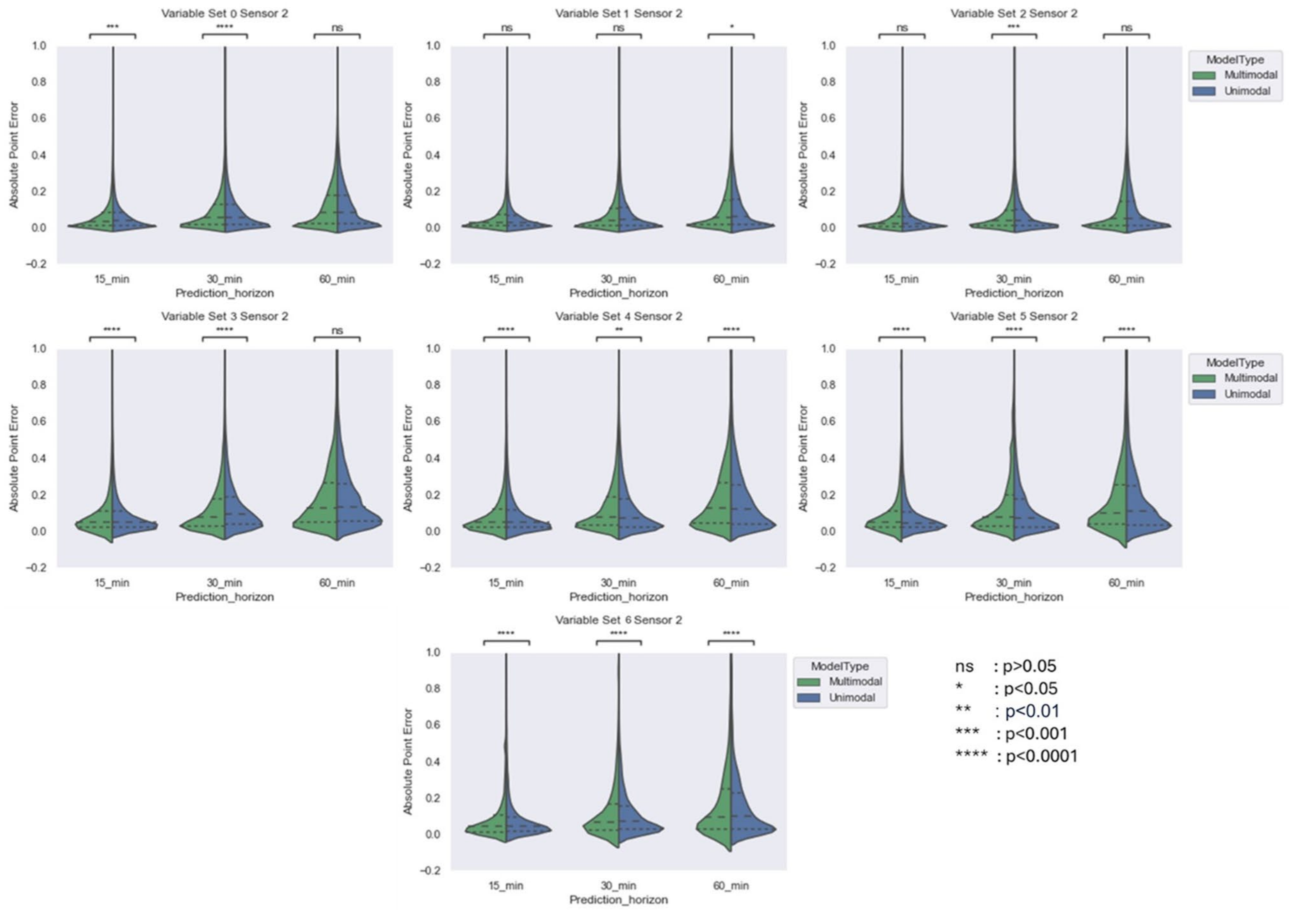


Fig. 4. Comparing violin plot of absolute point error for multimodal and unimodal architectures developed for Abbot sensor across different variable sets at three prediction horizon. The violin plot shows the distribution of absolute point error 25%, 50% and 75% quartile via dashed line.

The results reveals that increasing the prediction horizon negatively impacted cross-validation accuracy. As expected, MAPE increased with longer prediction horizons, with wider APE distributions observed at the 60-min horizon. Despite of constrained availability of baseline variables, the multimodal architecture significantly outperformed the unimodal model for the first four baseline variable sets at both 30-min and 60-min horizons. For the 15-min horizon, the difference in cross-validated MAPE between architectures was not statistically significant for the Menarini sensor (Sensor 1) with baseline sets 3 and 4 due to a smaller sample size (see Table 5), but was significant for the Abbott sensor (Sensor 2), where more subjects were available. In summary, while the unimodal and multimodal models performed comparably at the 15-min horizon, the multimodal architecture showed significantly better performance at 30 and 60 min, likely due to the incorporation of baseline variables that helped inform CGM trends over longer horizons.

Clinical explainability of prediction performance

We further evaluated the prediction performance of the multimodal architecture in a clinical context using Parkes Error Grid analysis[22]. The Parkes Grid Error was developed to present performance zones for blood/interstitial glucose prediction performance for type 2 diabetic subjects. It has 5 zones ranging from zone A – E with zone A defines "clinically accurate measurements with no impact on clinical actions" and zone B as "altered clinical action, little or no effect on clinical outcome". The results are tabulated in Table 7 as well as shown in Figs. 5 and 6 for some baseline variable sets. For each prediction horizon visualization, we selected the variable set where the multimodal architecture significantly outperformed the unimodal model (as shown in Figs. 3,4). The results demonstrate that, in all significant cases, multimodal predictions had a higher concentration of values within Zone A of Parkes' Error Grid for earlier baseline variable sets, across all horizons and for both sensors, indicating greater clinical accuracy. The variable sets with low number of patients do present better performance of unimodal architecture. Moreover, the multimodal models demonstrated improved performance in clinically critical ranges, accurately predicting glucose values as low as 70 mg/dL (hypoglycemia) and as high as 180 mg/dL (hyperglycemia). These findings suggest that incorporating personalized baseline information not only enhances statistical performance, but also improves the clinical reliability of CGM prediction models.

		Sensor 1								
		15 min			**30 min**			**60 min**		
	Set	**MS 1**	**UM 1**	***p*_value**	**MS 1**	**UM 1**	***p*_value**	**MS 1**	**UM 1**	***p*_value**
Overall MAPE	0	14.1	14.2		19.6	21.5	***	25.6	26.5	***
	1	14.8	15.1	*	19.9	20.9	***	25.6	25.5	
	2	14.9	15.2	*	20.3	20.6		26.0	26.2	
	3	16.5	16.4		21.9	22.8	***	26.3	26.8	
	4	24.3	15.3	***	21.4	22.8	***	26.2	27.4	**
Hyperglycaemia MAPE	0	12.5	12.5		15.8	17.6	***	21.8	21.3	
	1	11.3	10.4	***	16.0	16.5	*	23.1	24.6	***
	2	11.5	11.7		15.7	17.4	***	21.7	21.6	
	3	11.2	11.1		14.2	15.4	***	21.2	20.9	
	4	13.2	11.7		14.7	16.0	***	20.9	21.3	*
Hypoglycaemia MAPE	0	45.9	45.6		63.8	81.6	***	99.9	105.6	***
	1	52.1	53.3	*	74.7	82.3	***	102.2	97.5	**
	2	47.2	48.8	*	65.5	71.8	***	96.9	101.3	*
	3	55.4	50.3	***	70.8	82.6	***	93.3	95.7	
	4	76.3	42.7	***	69.4	81.6	***	93.8	101.7	***
	Sensor 2									
	Set	**MS 2**	**UM 2**	***p*_value**	**MS 2**	**UM 2**	***p*_value**	**MS 2**	**UM 2**	***p*_value**
Overall MAPE	0	6.7	6.8	***	9.2	9.4	***	12.0	12.0	
	1	5.9	5.8		7.9	8.0	*	10.2	10.4	*
	2	5.3	5.4	*	7.5	7.7	***	9.7	9.7	
	3	11.7	9.5	***	13.6	14.5	***	18.7	18.9	
	4	10.7	9.8	***	13.8	13.3	***	18.0	18.9	***
Hyperglycaemia MAPE	0	10.8	11.8	***	18.6	16.1	***	23.5	23.8	
	1	10.2	10.5		16.3	18.8	***	24.5	24.5	
	2	12.6	14.0	***	17.1	19.4	***	26.6	27.7	***
	3	12.6	11.5	***	16.6	18.1	***	23.6	25.9	***
	4	13.3	12.2	***	18.3	19.0	*	25.6	25.3	

Table 6. Comparison of Multimodal architecture performance with Unimodal architecture in terms of Mean Absolute Point Error (MAPE), Hyperglycaemic (interstitial glucose > 180 mg/dL) and Hypoglycaemic (interstitial glucose < 70 mg/dL). MAPE between acquired interstitial glucose and predicted interstitial glucose. We have include first 5 variable sets from Table 5. Due to low number of Hypoglycaemic events from patients acquired by Abbot sensor, the performance in terms of Hypoglycaemic MAPE has not been included for these patients. MS 1 = Multimodal Sensor 1; UM 1 = Unimodal Sensor 1; MS 2 = Multimodal Sensor 2; UM 2 = Unimodal Sensor 2. Sensor 1: Menarini Sensor; Sensor 2: Abbot Sensor. *p_value < 0.05; **p_value < 0.01; ***p_value < 0.001.

Discussion

Interpretation

In this study, we presented a novel multimodal architecture for predicting interstitial glucose with prediction horizon of 15, 30 and 60 min based on 30-min CGM variation sampled at 5 min along with baseline information of Type 2 subjects representing physiological status. Our novel multimodal architecture addresses several key questions, and, to the best of our knowledge, this is the first study to develop a multimodal architecture for predicting interstitial glucose of Type 2 diabetic subjects with personalized prior information. This study has the potential to estimate the chronic events such as hyperglycaemia and hypoglycaemia in real time based on personalized information with high clinical reliability. The technique was tested on a specific dataset of CGM values collected via two different sensors (Menarini sensor 1 with 1-min sampling downsampled to 5 min and Abbot sensor 2 with 15-min sampling which were upsampled to 5 min for consistency) from an elderly population with Type 2 diabetes whose baseline information representing their physiological status was also provided. The use of this original dataset highlights the robustness of the method in handling the complexities and challenges inherent in cohort interstitial glucose prediction and its accuracy at the personalized level.

Initially, we developed the unimodal architecture in which we developed the training pipeline based solely on CGM values. We first utilized basic deep learning blocks (such as convolutional neural networks (CNN), and long short-term memory (LSTM)) which was followed by adding the attention mechanism in order to train CGM sequential features while highlighting them based on temporal context. This improved the interstitial glucose prediction performance at a prediction horizon of 15 min in terms of Mean Absolute Percentage Error (MAPE) as separate architectures were developed for CGM values acquired from two different sensors; thus compared separately.

	15 min					30 min					60 min				
Set	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
Multimodal Sensor 1															
0	**90.2%**	8.6%	1.1%	0.1%	0.0%	**84.2%**	13.4%	2.2%	0.2%	0.0%	**76.8%**	19.4%	3.6%	0.2%	0.0%
1	**89.9%**	8.8%	1.2%	0.1%	0.0%	**83.6%**	14.0%	2.2%	0.2%	0.0%	**75.8%**	20.6%	3.5%	0.2%	0.0%
2	88.4%	10.2%	1.3%	0.1%	0.0%	**82.5%**	15.1%	2.3%	0.2%	0.0%	**76.1%**	19.6%	4.0%	0.3%	0.0%
3	**88.4%**	9.9%	1.6%	0.2%	0.0%	**81.5%**	15.4%	2.9%	0.3%	0.0%	74.0%	21.2%	4.6%	0.2%	0.0%
4	84.5%	10.4%	4.4%	0.7%	0.0%	**81.0%**	15.7%	3.0%	0.3%	0.0%	73.6%	21.6%	4.6%	0.2%	0.0%
Unimodal Sensor 1															
0	89.7%	9.1%	1.0%	0.1%	0.0%	82.6%	15.4%	1.8%	0.1%	0.0%	76.8%	19.2%	3.7%	0.3%	0.0%
1	89.7%	8.9%	1.3%	0.1%	0.0%	83.5%	14.2%	2.1%	0.2%	0.0%	74.1%	22.2%	3.5%	0.2%	0.0%
2	**89.4%**	9.1%	1.3%	0.1%	0.0%	81.8%	15.9%	2.2%	0.1%	0.0%	75.1%	20.6%	4.2%	0.1%	0.0%
3	87.4%	10.8%	1.6%	0.2%	0.0%	80.7%	16.6%	2.4%	0.2%	0.0%	**74.3%**	20.8%	4.6%	0.3%	0.0%
4	**86.6%**	11.8%	1.4%	0.2%	0.0%	80.7%	16.7%	2.5%	0.2%	0.0%	**73.7%**	21.3%	4.8%	0.2%	0.0%
Multimodal Sensor 2															
0	**96.7%**	3.3%	0.0%	0.0%	0.0%	**95.2%**	4.8%	0.1%	0.0%	0.0%	**93.0%**	7.0%	0.0%	0.0%	0.0%
1	**97.5%**	2.5%	0.0%	0.0%	0.0%	**96.0%**	4.0%	0.1%	0.0%	0.0%	**93.5%**	6.5%	0.1%	0.0%	0.0%
2	**97.6%**	2.4%	0.0%	0.0%	0.0%	**96.4%**	3.6%	0.1%	0.0%	0.0%	**94.0%**	6.0%	0.1%	0.0%	0.0%
3	92.3%	6.2%	1.5%	0.0%	0.0%	**91.2%**	7.9%	0.9%	0.0%	0.0%	**85.4%**	13.4%	1.2%	0.0%	0.0%
4	93.2%	5.6%	1.2%	0.0%	0.0%	90.1%	8.9%	1.0%	0.0%	0.0%	83.6%	14.2%	2.2%	0.0%	0.0%
Unimodal Sensor 2															
0	96.6%	3.3%	0.0%	0.0%	0.0%	94.7%	5.2%	0.0%	0.0%	0.0%	92.7%	7.2%	0.0%	0.0%	0.0%
1	97.5%	2.4%	0.0%	0.0%	0.0%	96.0%	4.0%	0.1%	0.0%	0.0%	93.5%	6.4%	0.1%	0.0%	0.0%
2	97.4%	2.5%	0.0%	0.0%	0.0%	96.0%	4.0%	0.0%	0.0%	0.0%	93.9%	6.1%	0.0%	0.0%	0.0%
3	94.4%	5.0%	0.6%	0.0%	0.0%	89.9%	8.9%	1.2%	0.0%	0.0%	83.6%	14.9%	1.5%	0.0%	0.0%
4	94.3%	5.0%	0.7%	0.0%	0.0%	**90.4%**	8.6%	1.0%	0.0%	0.0%	**84.4%**	14.3%	1.3%	0.0%	0.0%

Table 7. Percentage distribution comparison of Parkes' grid error zones for multimodal and unimodal architectures for both Menarini (Sensor 1) and Abbot (Sensor 2).

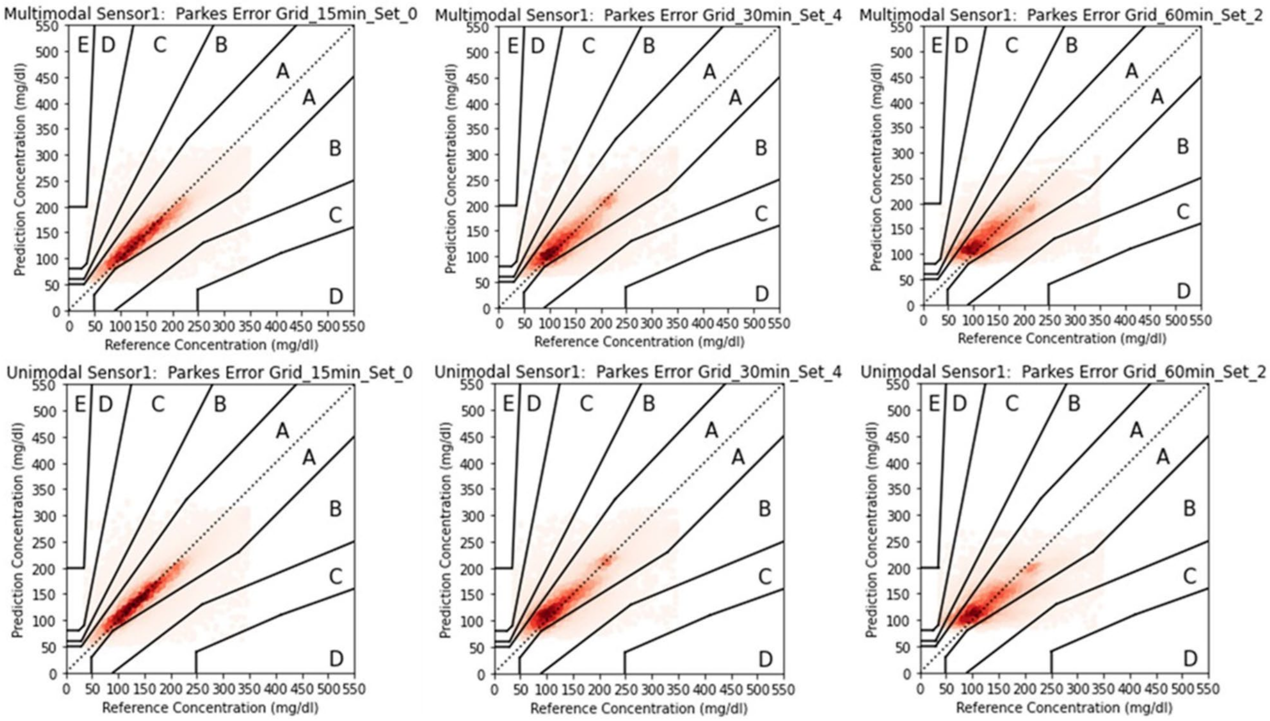
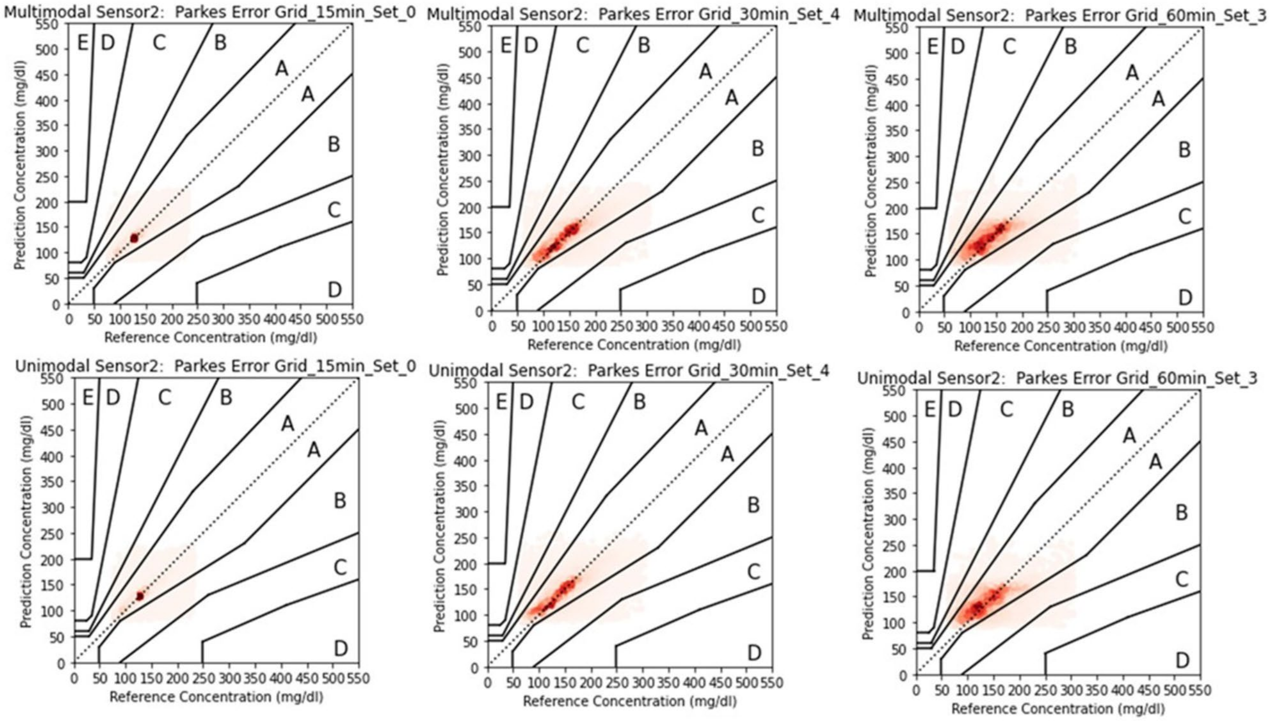


Fig. 5. Parkes' Grid error comparison between multimodal and unimodal architectures at selected variable set for prediction horizon of 15 min, 30 min and 60 min for Menarini sensor.


Multimodal Sensor2: Parkes Error Grid_15min_Set_0

Multimodal Sensor2: Parkes Error Grid_30min_Set_4

Multimodal Sensor2: Parkes Error Grid_60min_Set_3

Unimodal Sensor2: Parkes Error Grid_15min_Set_0

Unimodal Sensor2: Parkes Error Grid_30min_Set_4

Unimodal Sensor2: Parkes Error Grid_60min_Set_3

Fig. 6. Parkes Grid error comparison between multimodal and unimodal architectures at selected variable set for prediction horizon of 15 min, 30 min and 60 min for Abbot sensor.

The CGM training pipeline was then concatenated with the personalized baseline information pipeline to inform variations in the CGM training through additive concatenation, leading towards multimodal information. The multimodal architecture performance was compared with the unimodal architecture which involved a CGM pipeline only for predicting with prediction horizon of 15, 30 and 60 min. The leave-day cross validation protocol was introduced in which a day was kept out for testing purposes whereas the rest of days were used as training set with day-window sliding after every cycle. The prediction performance was also compared in terms of clinical significance using *Parkes' Grid error*, a graphical tool used to evaluate both accuracy and clinical relevance of glucose predictions in Type 2 diabetes subjects. The results show that informing the CGM variations based on personalized baseline information improves the prediction performance at the cohort level.

The multimodal model architecture significantly outperformed the unimodal architecture in terms of MAPE for first four baseline variable sets. For a prediction horizon of (i) 15 min, (ii) 30 min and (iii) 60 min, the MAPE was between (i) 14–16 mg/dL, (ii) 19–21 mg/dL and (iii) 25–26 mg/dL respectively compared to unimodal architectures with MAPE between (i) 14–16 mg/dL, (ii) 21–23 mg/dL and (iii) 26–27 mg/dL respectively. Besides, there has been higher concentration in zone A of Parkes' Grid error for multimodal architecture of these variable sets which shows high clinical significance. Of course the increase of prediction horizon reduced the prediction performance in terms of both MAPE and Parkes' Grid error. Nevertheless, the performance drop for multimodal architecture while moving from a prediction horizon of 15 min to 60 min was lower compared to unimodal architectures as shown in APE distribution in Figs. 3 and 4.

In this study, we also observed the MAPE in terms of chronic events such as Hyperglycaemia and Hypoglycaemia. We observed that the performance of predicting high interstitial glucose values (Hyperglycaemia) was even better within 15 min and 30 min prediction horizon. This results in significance of our multimodal architecture in the case of Type 2 diabetes as there had been high number of hyperglycaemic events in Type 2 diabetes. On the other hand, our multimodal architecture performance was relatively poor in predicting blood glucose in hypoglycaemic range. This is because there had been low number of hypoglycaemic events occurred in Type 2 diabetes; resulting in data imbalance problem.

Limitations

This study had some limitations. Firstly the bottleneck associated with multimodal architecture performance for large baseline variable set was their availability. The variable sets were defined based on their availability with respect to the individual patients. As shown in Table 5, increase in number of baseline variables reduced the number of patients as there were only few patients with every baseline information. The low number of patients actually impacted the multimodal architecture performance as it worked better for a higher number of patients. Secondly, the CGM values collected for individual subjects were not consistent as data acquisition varied from total number from 4 to 10 days. Our upcoming studies may incorporate augmenting the dataset based on probabilistic distribution at the individualized physiology. Thirdly, the dataset size was relatively small (n = 40) and unevenly split across the two CGM devices (Sensor 1 and Sensor 2). Although population models were trained separately for each sensor to account for device-specific variation, domain adaptation techniques were not applied in this study. Future work will explore such approaches to enhance model generalizability across CGM systems and broader T2D populations.

In this study, we opt for additive concatenation approach due to the limitation of size of the dataset. This is because adding model complexity (such as transformer mechanism) resulted in prediction performance. As a part of our future studies, we aim to develop the advanced Graphical Neural Networks which can train the CGM variations based on counterfactual analysis of underline comorbidities. Besides, due to limitations in terms of dataset size, the performance depreciation was observed while increasing the underline comorbidities due to their availability with limited patients.

Methods

Study protocol

This study is part of the GATEKEEPER strategy for the Multinational Large-Scale Piloting of an eHealth Platform[23]. The data were collected in the frame of the Central Greece High Complexity Phase I pilot study: a non-interventional, prospective observational study. The applied eligibility criteria encompass elderly patients with T2D and comorbidities, aged 60 years and older. These patients belonged to the intermediate and poor health groups according to clinical guidelines mentioned in[24]. These groups had 3 or more non-diabetic chronic illnesses with mild to severe cognitive impairment. Specifically, people with T2D participated in Phase I, using either the GlucoMen Day Menarini® Continuous Glucose Monitoring (CGM) system (15-min sampling interval) or the Libre Abbot® system (5-min sampling interval) for a monitoring period of up to 4 weeks.

Calibration is the key difference among both sensors. Abbott's FreeStyle Libre is factory calibrated and remains stable throughout its lifespan, while Menarini's GlucoMen Day requires periodic user calibration, which can introduce variability. Sensor placement also differs: Abbott sensors are worn on the back of the arm, and Menarini sensors may be placed elsewhere, affecting accuracy due to variations in skin thickness, blood flow, and fat. Additionally, Abbott's sensor measures every 15 min, while Menarini's system measures every minute, which can impact data quality, especially during rapid glucose changes. Both sensors can be influenced by environmental factors like temperature, sweating, and physical activity, with Menarini's sensor being more sensitive to skin perspiration. These factors, along with individual physiological differences, contribute to discrepancies between the two devices.

Patients with severe hearing or vision problems or any other acute or chronic condition that would limit the ability of the user to participate in the study were excluded[25]. All the data collection methods were performed in accordance with the relevant guidelines of the Institutional Review Board of Larisa University Hospital, which are aligned with the Declaration of Helsinki[26]. The names of all participants and other HIPAA identifiers[27] have been removed prior to data sharing. Furthermore, informed consent has been obtained from all participants and/or their legal guardians. The timeline for the study protocol (incl. ethical approval, study design, data acquisition and integration into GATEKEEPER high performance big data platform) has been presented in Supplementary Fig. 1. For more information about study participants, please check clinical trials.gov ID NCT05461716.

Data curation and preprocessing

We applied exploratory techniques to visualize each patient's CGM data, including histograms, autocorrelation plots, partial autocorrelation plots, and the Augmented Dickey-Fuller (ADF) test. We also checked for duplicates and outliers in each time series. To handle missing values in the glucose sensor data, linear interpolation was applied to ensure continuity in the time series. Specifically, missing values in the glucose sensor readings were imputed using linear interpolation. Additionally, for patients using the Abbott sensor, the data was kept at its original sampling frequency of 15 min, as provided by the sensor. However, for patients using the Menarini sensor, the data was resampled to a 5-min interval. Patients were excluded from the analysis if more than 50% of their CGM data was missing, either due to sensor dropouts or user non-compliance. This threshold was set to ensure that the remaining data was sufficiently complete to maintain the integrity and reliability of the analysis. Each type of data was processed separately before merging. Min–Max scaling was applied to normalize the CGM data to a range of [0, 1], while no normalization techniques were applied to the baseline variables.

Outcome and predictors definition

The output of the predictive model describes the concentration of glucose concentration in the interstitial fluid at time $t + PH$ for a prediction horizon (PH) equal to 15, 30 or 60 min. The univariate models' input comprises the history of interstitial glucose concentration values, as recorded by the CGM system. In the case of multimodal models, the input includes additionally specific EHR variables. T2D participated in this study used either the GlucoMen Day Menarini® Continuous Glucose Monitoring (CGM) system with sampling interval of 15 min; or the Libre Abbot® system with sampling interval of 5 min; for a monitoring period of up to 4 weeks. Besides, for informing interstitial glucose variation with underline comorbidities, we have also included baseline variables representing these comorbidities as predictors.

Multimodal architecture for CGM prediction

We developed multimodal architectures built upon deep neural networks which have the capability to model real-time CGM variations while informing these variations via appropriate information fusion methods. The CGM variations have been informed by patient electronic health records (e.g. demographics, or anthropometrics, as shown in Table 1). CGM data had been acquired from T2DM patients under real-time conditions. We compiled and compared the results by performing cross validation using the first 30 min of CGM data for training and predicting the CGM values after prediction horizon of (i) 15 min, (ii) 30 min and (iii) 60 min. The training

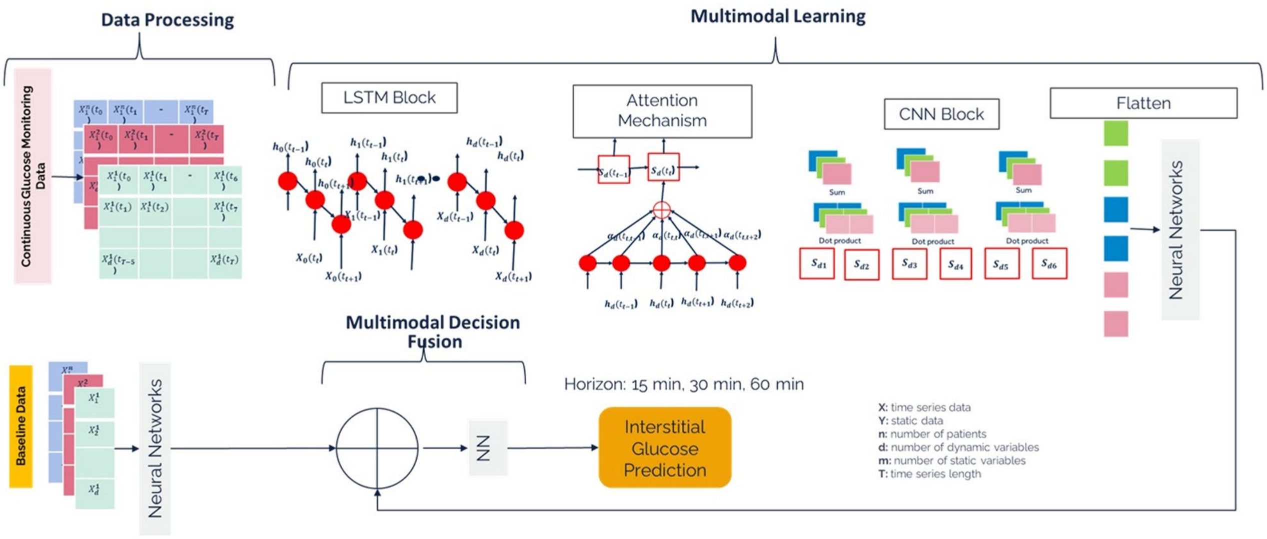


Fig. 7. Multimodal Architecture for predicting blood glucose.

and test sets were derived based on setting the interval length of test set of one day which were sliding from beginning till end of the dataset.

At the first instance, we derived the CGM-only trained population model to find optimal deep neural network architecture; we call it unimodal architecture.

At the second instance, we derived the deep neural networks pipeline which was trained on the T2D patient baseline information. First, we involved two baseline variables which were present in most of the subjects using Abbot and Menarini CGM devices. The inclusion of more baseline variables led to a reduction in the number of available subjects. At the end, there were only 6 subjects who had 14 baseline variables. The list has been presented in Table 5. This enabled us to develop 7 multimodal architectures due to 7 baseline variables subsets. The output of the baseline deep neural network was then fused into the CGM-only training pipeline (unimodal architecture) via additive fusion methods followed by the deep neural network training on CGM variation features informed by the baseline network.

We trained and compared unimodal and multimodal architectures across 7 variable sets for predicting CGM values for the aforementioned prediction horizons of 15, 30 and 60 min. The comparison for both sensors was performed separately using violin plot, showing the distribution of absolute point errors between predicted CGM values and real CGM values, along with quartile markings at 25%, 50% and 75%. Due to instability in the model performance for baseline variable set 5 and onwards, we included the performance comparison from variable set 0 to variable set 4 only.

Clinical explainability of prediction performance

We further assessed the multimodal architecture prediction performance under clinical settings based on *Parkes grid error*[22]. The Parkes 'grid error classifies the scatter plot of predicted interstitial glucose and reference interstitial glucose for type 2 diabetic subjects in five different zones: A, B, C, D and E. The estimation in zone A would be considered as ideal, whereas estimation in zone B would be considered as clinically acceptable.

Model development

The block diagram of architecture has been presented in Fig. 7. The architecture has been designed to predict the interstitial glucose at defined time horizon based on (i) time series historical values from continuous glucose monitoring (CGM) and (ii) static baseline health record information. Let CGM is represented as X has the dimension of n x T; where n is number of users and T is the length of temporal dimension of the CGM input. The CGM values had been acquired from multiple users across different number of days. Each user had been provided one out of two types of CGM devices. One type of device had sampling frequency of 5 min whereas other type of device had sampling frequency of 15 min. For users with a 15-min sampling interval (i.e., Abbot Sensor 2), CGM time series were upsampled to 5 min intervals using linear interpolation. Similarly, Menarini sensor with sampling interval of 1-min was downsampled to 5 min to ensure consistent temporal resolution with Abbot Sensor 2. Considering data acquisition spanning around a couple of days, we performed the data window scheme acquiring $t_0 - t_5, t_1 - t_6, t_{T^n-5} - t_{T^n}$; where T^n is total number of samples for user n. We used 30-min sample to predict interstitial glucose with time horizon of (i) 15 min, (ii) 30 min and (iii) 60 min.

As mentioned in Figs. 1 and 7, the local and temporal features of CGM have been acquired by 1D BiLSTM network with attention layer followed by staked 1D CNN layer. The model initially prepares the CGM values based on the aforementioned window scheme. Concurrently, the baseline data is treated as separate input which is preprocessed and learned separately using a set of dense layers to extract representative deep features. After acquisition of local and temporal features from CGM and representative deep features from baseline data, we added a fusion layer concatenating both types of features followed by a dense layer with sigmoid activation for regressing the CGM values.

To model the temporal context of the CGM data, we first deployed BiLSTM layers; allowing temporal patterns from CGM to be extracted. The core structure of the LSTM cell is the use of three gates i.e. the input gate (i_{t_T}), the forget gate (f_{t_T}), and the output gate (o_{t_T}). These gates control the update, maintenance, and deletion of information contained in a cell state $\widetilde{C}_{t_T}$; $C_{t_{T-1}}$, and C_{t_T} respectively whereas h_{t_T} is the value of the hidden layer at time t_T. θ s represent set of weight matrices and b s represent set of biases vectors which are updated following backpropagation algorithm with each temporal iteration. Besides, θ s and the b s are the set of weight matrices and biases vectors, respectively, updated following the backpropagation through time algorithm. In addition, $\otimes$ represents the Hadamard product; σ is the standard logistic sigmoid function; $\oplus$ is the concatenation operator; and φ the output activation function. Equations (1)–(7) give the transmission of information in the memory cell at each step.

$$f_{t_T} = \sigma(\theta_f \cdot \left[h_{t_{T-1}}, X_{t_T}\right] + b_f) \tag{1}$$

$$i_{t_T} = \sigma(\theta_i \cdot \left[h_{t_{T-1}}, X_{t_T}\right] + b_i) \tag{2}$$

$$\widetilde{C}_{t_T} = tanh(\theta_c \cdot \left[h_{t_{T-1}}, X_{t_T}\right] + b_c) \tag{3}$$

$$C_{t_T} = f_{t_T} \otimes C_{t_{T-1}} \oplus i_{t_T} \otimes \widetilde{C}_{t_T} \tag{4}$$

$$o_{t_T} = \sigma(\theta_o \cdot \left[h_{t_{T-1}}, X_{t_T}\right] + b_o) \tag{5}$$

$$h_{t_T} = o_{t_T} \otimes tanh(C_{t_T}) \tag{6}$$

$$y_T = \varphi(\theta_y h_{t_T} + b_y) \tag{7}$$

In order to take the advantage of temporal context in both directions, we deployed *BiLSTM* which combines input from two separate hidden LSTM layers in opposite direction to the same output. Let's consider $X^1(t_{0:5})$=($X^1(t_0)$, $X^1(t_1)$, $X^1(t_2)$, $X^1(t_3)$, $X^1(t_4)$, $X^1(t_5)$); for which LSTM hidden layer becomes $\overrightarrow{h}^n_t = \left(\overrightarrow{h}^n_{t_0}, \overrightarrow{h}^n_{t_1}, \overrightarrow{h}^n_{t_2}, \overrightarrow{h}^n_{t_3}, \overrightarrow{h}^n_{t_4}, \overrightarrow{h}^n_{t_5}\right)$ towards forward hidden sequence and $\overleftarrow{h}^1_t = \left(\overleftarrow{h}^1_{t_0}, \overleftarrow{h}^1_{t_1}, \overleftarrow{h}^1_{t_2}, \overleftarrow{h}^1_{t_3}, \overleftarrow{h}^1_{t_4}, \overleftarrow{h}^1_{t_5}\right)$ towards backward hidden sequence. Thus, Eq. (7) is now driven as:

$$\overrightarrow{h}^n_{t_T} = \sigma\left(\theta_{\overrightarrow{h}^n_T} . \left[\overrightarrow{h}^n_{t_{T-1}}, X^n\left(t_T\right)\right] + b_{\overrightarrow{h}^n_T}\right) \tag{8}$$

$$\overleftarrow{h}^n_{t_T} = \sigma\left(\theta_{\overleftarrow{h}^n_T} . \left[\overleftarrow{h}^n_{t_{T+1}}, X^n\left(t_T\right)\right] + b_{\overleftarrow{h}^n_T}\right) \tag{9}$$

$$(\overrightarrow{h}^n_{t_0}, \overleftarrow{h}^n_{t_0}) \ldots (\overrightarrow{h}^n_{t_T}, \overleftarrow{h}^n_{t_T}) = BiLSTM(X^n\left(t_0\right), X^n\left(t_1\right), ..., X^n(t_5)) \tag{10}$$

$$y^n_t = \varphi(\theta_{y^n_t \overrightarrow{h}^n_T} \overrightarrow{h}^n_{t_T} + \theta_{y_t \overleftarrow{h}^n_T} \overleftarrow{h}^n_{t_T} + b_{y^n_t}) \tag{11}$$

The output y^n_t is used as an input to the *self-attention layer* which had been deployed to highlight the local CGM features under consideration based on the temporal context. This can be represented by $\sigma^a(q^a, v^a)_{t,t\prime}$ which is the softmax function between query (context) of the attention layer q^a and value of attention layer v^a at time t and $t\prime$.

$$\sigma^a(q^a, v^a)_{t,t\prime} = \frac{e^{dot(q^a_t, v^a_{t\prime})}}{\sum_{t=0}^{l_f} e^{dot(q^a_t, v^a_{t\prime})}} \tag{12}$$

where l_f=6 is the number of output units of the BiLSTM later. Since it is the self-attention mechanism, the input to both is y^n_t.

The 1D *Convolutional Neural Network (CNN)* blocks had been deployed to model the local features provided by self-attention layer based on temporal context of the CGM. 1D CNN can learn attention driven temporal context time series univariate data where convolution is done separately along the time dimension for every input vector. Formally if input $\sigma^a(q^a, v^a)_{t,t\prime} \in \mathbb{R}^{l_f \times 1}$ and kernel K is $m \times 1$ then convolutional output in new feature space would be $\sigma'^a(q^a, v^a)_{t,t\prime} \in \mathbb{R}^{[\frac{l_f - m}{d+1}, 1]}$, where d is the step size. Based on number of filters, the CNN expands the attention output to more abstract and informative features, called feature maps. Each value p_i of the feature map p is then fed into activation function, $\varnothing$, to calculate $p_i = \varnothing\left(K^T \times \sigma^{a(i:i+j-1)} + b\right)$, where activation function $\varnothing$ is non-linear activation function $RELU\left(x\right) = \max(o, x)$, b is the bias and $\sigma^{a(i:i+j-1)}$ is the j observation from σ^a. The CNN networks have been followed by 20% dropout to avoid overfitting. The kernel in the convolutional layer had been initialized by Glorot Uniform which initializes the convolutional weights based on uniform distribution within range [-limit, limit] where limit = $\sqrt{\frac{6}{f_{in}+f_{out}}}$ where f_{in} is number of input units and f_{out} is number of output units.

As empirical experimentation, we put size of kernel K as 3 for CNN with number of filters as 100. The CNN network was then followed by 10% dropout.

The *multimodal fusion network* allows to fuse the representations learned from the CGM values and the baseline data. Considering that the learned representation from CGM values (i.e. CNN output) is Z^1 and the learned representations from fully connected neural networks trained on baseline data is Z^2. The fusion of both representations are learned by multi-layer fully connected neural network. This can be represented as:

$$Z^3 = G(Z^1 \oplus Z^2, W_3) \tag{13}$$

where $\oplus$ is the fusion operator, W_3 is the matrix of trainable weights and G is the multilayer fully connected neural network. Following the multimodal fusion, we deployed dense layer regressor to predict the interstitial glucose with specified prediction horizon. The regressor is a fully connected neural network followed by sigmoid function. The final results of the regressor and classifier are represented as $\widehat{Y}^{n}_{T^n} \in T^n \times 1$ where T^n is total number of time samples for the subject n.

The *objective loss function* of estimating interstitial glucose is log likelihood function represented as:

$$\mathcal{L} = \sum_{k=1}^{n} \sum_{i=1}^{T_k} \left(\widehat{Y}_i^k - log \left(\sum_{j \epsilon t_i} exp(\widehat{Y}_j^k) \right) \right) \tag{14}$$

where t_i is the prediction horizon for estimating interstitial glucose of the subject k. Noting that loss function is the summation of predicting interstitial glucose for every subject k with their respective samples T_k.

Model evaluation

Mean Absolute Point Error (MAPE) has been selected to evaluate the model which measures average magnitude of error produced by a model with the advantage of scale-independency and interpretability[20]. It can be calculated as:

$$MAPE = 100\frac{1}{n}\sum_{t=1}^{n}\left|\frac{CGM_t^A - CGM_t^P}{CGM_t^A}\right| \tag{15}$$

where CGM_t^A is actual CGM value and CGM_t^P is the predicted CGM value.

To evaluate model performance, we employed a leave-one-day-out cross-validation approach. In this method, each day's data was sequentially designated as the test set while the remaining data was used for training. This sliding window technique ensured that each data point was tested at least once while maximizing the amount of training data available for each iteration. For each iteration, data preceding the test day and data following the test day were combined to form the training set, while the designated day was held out as the test set. This data partitioning strategy is commonly used in time-series forecasting studies where temporal dependencies are critical.

The multimodal architecture training and validation had been implemented on GATEKEEPER Big Data platform where all the data from the pilot has been hosted and deep learning packages have been trained and tested in the platform. The total training time was 1 min to run 50 iterations in each cross-validation cycle.

Related work

Deep learning has emerged as a leading approach in interstitial glucose predictions, with a primary focus on applications in Type 1 Diabetes Mellitus (T1DM)[28]. Initial work using LSTM-based models on the OhioT1DM dataset[29,30] showed limited gains over feature-engineered traditional Machine Learning (ML) methods. More sophisticated architectures, including attention-based Gated Recurrent Units (GRU)s[31] and CNNs[32], have since demonstrated improved performance across T1D, T2D, and gestational diabetes datasets.

A growing number of studies aim to improve individual-level prediction accuracy while ensuring generalizability across diverse populations and data sources. In the context of T1D, Zhu et al.[33] utilised meta-learning and evidential deep learning (i.e., including an attention-based bidirectional Recurrent Neural Networks (RNN) and evidential regression) to quantify uncertainty and personalize glucose forecasting. Daniels et al.[34] introduced a multitask learning architecture that jointly models shared and individual-specific representations of glucose dynamics in T1D patients. Regarding T2D, Deng et al.[35] employed deep transfer learning with data augmentation to improve robustness under limited data conditions. Sun et al.[36] developed a Bayesian structural time series model that incorporates clinical data priors (i.e., anthropometric and biochemical characteristics) to address inter-individual variability in T2D. Similarly, Yang et al.[37] proposed a clustering-based domain adaptation approach, enabling more personalized modelling by aligning latent representations across patient subgroups.

Complementary to models based solely on CGM, Montaser et al.[38] proposed a seasonal stochastic local modelling framework that explicitly incorporates variable-length, time-stamped events such as meals and physical activity. This work underscores the relevance of irregular but clinically significant behavioural factors in interstitial glucose prediction. Other contributions have emphasized model interpretability in multivariate glucose predictive modelling; a graph-attentive RNN (GARNN) framework[39] captures detailed interactions among CGM and self-reported event data, enhancing both prediction accuracy and transparency.

Conclusion

In this paper, we designed and developed a novel generalized multimodal architecture based on 30-min CGM values informed by baseline physiological information of Type 2 diabetic patients for predicting CGM values with prediction horizon of 15, 30 and 60 min. To the best of our knowledge, this is the first study of predicting interstitial glucose values where CGM variation were informed by individual physiology. Compared to unimodal architecture, we achieved the mean absolute point error of (i) 14–16 mg/dL, (ii) 19–21 mg/dL and (iii) 25–26 mg/dL for predicting CGM values with prediction horizon of 15, 30 and 60 min respectively while addressing the clinical trustworthiness of our model. Besides, the multimodal architectures had lower MAPE for predicting interstitial glucose compared to unimodal architectures in hypoglycaemic as well as in hyperglycaemic range. The model had limitations due to the non-availability of baseline physiological information for every patient along with the lower number of participants in the study. Therefore, as our planned future work, we aim to develop the methodologies to augment missing information based on probabilistic distribution of the dataset. Nevertheless, this model managed to predict the interstitial glucose for prediction horizon of up to 60 min with adequate prediction accuracy which can serve as a first step for generalized interstitial glucose prediction model. Besides, we also aim to conduct the studies based on impact of meal and exercises on interstitial glucose variation.

Data availability

The data that support the findings of this study are available upon reasonable request to the pilot managers from University Hospital of Larisa such as Alexandra Bargiota (abargio@med.uth.gr) and University of Thessaly such as George E Dafoulas (gdafoulas@uth.gr).

Code availability

The code used in this study is available upon request to the corresponding author Dr. Muhammad Salman Haleem salman.haleem@warwick.ac.uk via University of Warwick license.

Received: 22 January 2025; Accepted: 13 June 2025
Published online: 29 July 2025

References

1. Zimmet, P., Alberti, K. G., Magliano, D. J. & Bennett, P. H. Diabetes mellitus statistics on prevalence and mortality: Facts and fallacies. *Nat. Rev. Endocrinol.* **12**(10), 616–622 (2016).
2. Galicia-Garcia, U. et al. Pathophysiology of type 2 diabetes mellitus. *Int. J. Mol. Sci.* **21**(17), 6275 (2020).
3. Sun, H. et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **183**, 109119 (2022).
4. Parker, E. D. et al. Economic costs of diabetes in the US in 2022. *Diabetes Care* **47**(1), 26–43 (2024).
5. Bellary, S., Kyrou, I., Brown, J. E. & Bailey, C. J. Type 2 diabetes mellitus in older adults: clinical considerations and management. *Nat. Rev. Endocrinol.* **17**(9), 534–548 (2021).
6. Kaufman, J. M., Thommandram, A. & Fossat, Y. Acoustic analysis and prediction of type 2 diabetes mellitus using smartphone-recorded voice segments. *Mayo Clinic Proceedings: Digital Health* **1**(4), 534–544 (2023).
7. N. S. Padhye, J. Wang, Pattern of active and inactive sequences of diabetes self-monitoring in mobile phone and paper diary users, in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015: IEEE, pp. 7630–7633.
8. Ernawati, U., Wihastuti, T. A. & Utami, Y. W. Effectiveness of diabetes self-management education (DSME) in type 2 diabetes mellitus (T2DM) patients: Systematic literature review. *J. Public Health Res.* **10**(2), jphr-2021 (2021).
9. Wagner, J., Malchoff, C. & Abbott, G. Invasiveness as a barrier to self-monitoring of blood glucose in diabetes. *Diabetes Technol. Ther.* **7**(4), 612–619 (2005).
10. Bailey, T., Bode, B. W., Christiansen, M. P., Klaff, L. J. & Alva, S. The performance and usability of a factory-calibrated flash glucose monitoring system. *Diabetes Technol. Ther.* **17**(11), 787–794 (2015).
11. A. D. Association. The American Diabetes Association Releases Standards of Care in Diabetes—2025. https://www.diabesrelief.online/newsroom/press-releases/american-diabetes-association-releases-standards-care-diabetes-2025 (accessed 02–05, 2025).
12. Kebede, M. M., Zeeb, H., Peters, M., Heise, T. L. & Pischke, C. R. Effectiveness of digital interventions for improving glycemic control in persons with poorly controlled type 2 diabetes: A systematic review, meta-analysis, and meta-regression analysis. *Diabetes Technol. Ther.* **20**(11), 767–782 (2018).
13. Sobel, S. I., Chomentowski, P. J., Vyas, N., Andre, D. & Toledo, F. G. Accuracy of a novel noninvasive multisensor technology to estimate glucose in diabetic subjects during dynamic conditions. *J. Diabetes Sci. Technol.* **8**(1), 54–63 (2014).
14. Andellini, M. et al. Artificial intelligence for non-invasive glycaemic-events detection via ECG in a paediatric population: Study protocol. *Heal. Technol.* **13**(1), 145–154 (2023).
15. Cisuelo, O. et al. Development of an artificial intelligence system to identify hypoglycaemia via ECG in adults with type 1 diabetes: Protocol for data collection under controlled and free-living conditions. *BMJ Open* **13**(4), e067899 (2023).
16. Fokkert, M. et al. Performance of the FreeStyle Libre Flash glucose monitoring system in patients with type 1 and 2 diabetes mellitus. *BMJ Open Diabetes Res. Care* **5**(1), e000320 (2017).
17. van Doorn, W. P. et al. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. *PLoS ONE* **16**(6), e0253125 (2021).
18. Liu, K. et al. Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis. *JMIR Med. Inform.* **11**(1), e47833 (2023).
19. Service, F. J. Glucose variability. *Diabetes* **62**(5), 1398–1404 (2013).
20. Kim, S. & Kim, H. A new metric of absolute percentage error for intermittent demand forecasts. *Int. J. Forecast.* **32**(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003 (2016).
21. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68**(6), 540–546 (2015).

22. Pfützner, A., Klonoff, D. C., Pardo, S. & Parkes, J. L. Technical aspects of the Parkes error grid. *J. Diabetes Sci. Technol.* **7**(5), 1275–1281 (2013).
23. de Batlle, J. et al. GATEKEEPER's strategy for the multinational large-scale piloting of an eHealth platform: tutorial on how to identify relevant settings and use cases. *J. Med. Internet Res.* **25**, e42187. https://doi.org/10.2196/42187 (2023).
24. LeRoith, D. et al. Treatment of diabetes in older adults: An endocrine society* clinical practice guideline. *J. Clin. Endocrinol. Metab.* **104**(5), 1520–1574. https://doi.org/10.1210/jc.2019-00198 (2019).
25. G. E. Dafoulas. Incidence of Hypoglycaemia Events in Patients With Stable Insulin-treated Type 2 Diabetes Mellitus Based on Continuous Glucose Monitoring. https://clinicaltrials.gov/study/NCT05461716?term=George%20Dafoulasrank=1#participation-criteria (accessed 02-05, 2025).
26. W. M. Association, WMA Declaration of Helsinki–ethical principles for medical research involving human participants, 2013, ed.
27. Portability, I. & Act, A. *Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule* (Human Health Services, 2012).
28. Felizardo, V., Garcia, N. M., Pombo, N. & Megdiche, I. Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review. *Artif. Intell. Med.* **118**, 102120 (2021).
29. Martinsson, J., Schliep, A., Eliasson, B. & Mogren, O. Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res.* **4**, 1–18 (2020).
30. C. Marling, R. Bunescu, The OhioT1DM dataset for blood glucose level prediction: Update 2020, in CEUR workshop proceedings, 2020, vol. 2675: NIH Public Access, p. 71.
31. Koca, Ö. A., Kabak, H. Ö. & Kılıç, V. Attention-based multilayer GRU decoder for on-site glucose prediction on smartphone. *J. Supercomput.* **80**(17), 25616–25639 (2024).
32. Seo, W., Park, S.-W., Kim, N., Jin, S.-M. & Park, S.-M. A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study. *Comput. Methods Programs Biomed.* **211**, 106424 (2021).
33. Zhu, T., Li, K., Herrero, P. & Georgiou, P. Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning. *IEEE Trans. Biomed. Eng.* **70**(1), 193–204 (2022).
34. Daniels, J., Herrero, P. & Georgiou, P. A multitask learning approach to personalized blood glucose prediction. *IEEE J. Biomed. Health Inform.* **26**(1), 436–445 (2021).
35. Deng, Y. et al. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. *NPJ Digital Med.* **4**(1), 109 (2021).
36. Y. Sun, P. Kosmas, Integrating bayesian approaches and expert knowledge for forecasting continuous glucose monitoring values in type 2 diabetes mellitus, IEEE Journal of Biomedical and Health Informatics 2024.
37. Yang, T., Yu, X., Tao, R., Li, H. & Zhou, J. Blood glucose prediction for type 2 diabetes using clustering-based domain adaptation. *Biomed. Signal Process. Control* **105**, 107629 (2025).
38. Montaser, E., Díez, J.-L. & Bondia, J. Glucose prediction under variable-length time-stamped daily events: A seasonal stochastic local modeling framework. *Sensors* **21**(9), 3188 (2021).
39. Piao, C. et al. GARNN: an interpretable graph attentive recurrent neural network for predicting blood glucose levels via multivariate time series. *Neural Netw.* **185**, 107229 (2025).

Acknowledgements

The authors would like to express their sincerest gratitude to all the partners and participants of the GATEKEEPER project. This project has received funding from the European Union's Horizon 2020 research and innovation program (under grant 857223), cofounded by the European Regional Development Fund (ERDF), "A way to make Europe." The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

MSH: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data Processing, Writing—Original Draft, Review and Editing, Visualisation; DK: Conceptualisation, Methodology, Formal analysis, Investigation, Data Processing, Writing—Original Draft, Review and Editing; EG: Conceptualisation, Investigation, Data Processing, Writing—Original Draft, Review and Editing; GD: Domain Expert, Data Collection, Clinical Study Lead and Clinical Interpretation of Results; AB: Domain Expert, Clinical Study Definition, Data Collection; LLP: Review, Editing and Writing; MR: Review, Editing and Writing; GF: Funding Acquisition, Review, Editing and Writing; LP: Funding Acquisition, Review, Editing and Writing; DF: Funding Acquisition, Review, Editing and Writing. The authors read and approved the final manuscript and are accountable for ensuring accuracy and integrity of any part of the work.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The ethical approval of this study has been obtained through Institutional Review Board of Larisa University Hospital prior to commencement of data collection. The study protocol has been registered in clinicaltrials.gov (NCT05461716). All the methods were performed in accordance with the relevant guidelines of the Institutional Review Board of Larisa University Hospital which are aligned with the Declaration of Helsinki. The names of all participants and other HIPAA identifiers have also been removed prior to data sharing. Besides, informed consent has been obtained from all participants and/or their legal guardians.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-07272-3.

Correspondence and requests for materials should be addressed to M.S.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Gatekeeper Consortium

Claudio Caimi[8], Christian Tamporale[8], Mirko Manea[8], Chiara Bonferini[8], Eugenio Gaeta[6], Gloria Cea Sánchez[6], Ioanna Drympeta[9], Konstantinos Votis[9], Frans Folkvord[10,11] & Jordi Battle[12,13]

[8]Hewlett-Packard Italiana, Milan, Italy. [9]Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece. [10]PredictBy Research and Consulting, Barcelona, Spain. [11]Tilburg School of Humanities and Digital Sciences, Tilburg, The Netherlands. [12]Hospital Universitari Arnau de Vilanova and Santa Maria, Lleida, Spain. [13]Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid, Spain.