A Deep Learning Approach of Blood Glucose Predictive Monitoring for Women with Gestational Diabetes

Huiqi Lu (yvonne.lu@eng.ox.ac.uk)
University of Oxford https://orcid.org/0000-0002-6140-3394
David Clifton
University of Oxford
Ping Lu
University of Oxford
Jane Hirst
Oxford University Hospitals NHS Foundation Trust
Lucy MacKillop
Oxford University Hospitals NHS Foundation Trust

Article

Keywords:

Posted Date: March 8th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2653408/v1

Abstract

Gestational diabetes is a subtype of diabetes that develops during pregnancy.Managing blood glucose (BG) within the healthy physiological range can reduce clinical complications for women with gestational diabetes. In this paper, we developed a Stacked Long and Short-Term Memory (Stacked-LSTM) model to predict before- and after-meal BG levels for women with gestational diabetes. A total of 190396 BG readings from 1110 patients were used for model development, validation and testing under three different prediction schemes: 7 days of BG readings to predict the next 7 or 14 days, and 14 days to predict 14 days. Our results show that the optimized BG schedule uses a 7-day observational window to predict the BG of the next 14 days, achieved accuracies of RMSE = 0.958±0.007, 0.876±0.003, 0.898±0.003, 0.622±0.003, 0.814±0.009, and 0.845±0.005 for the after-breakfast, after-lunch, after-dinner, before-breakfast, before-lunch and before-dinner predictions, respectively. The performance of our models is on par withthe prediction accuracies (RMSE) in benchmark BG prediction models using continuous glucose monitoring (CGM) readings. In conclusion, the Stacked-LSTM model is a promising approach for capturing the patterns in time series data, resulting in accurate predictions of BG levels. Using a deep learning model with routine fingerstick glucose collection is a promising, predictable and low-cost solution for BG monitoring for women with gestational diabetes.

I. Introduction

Gestational diabetes mellitus (GDM) is one of the most common medical complications of pregnancy and is defined as carbohydrate intolerance resulting in hyperglycemia of variable severity with onset or first recognition during pregnancy by the World Health Organization (WHO) [1]. GDM is associated with short- and long-term adverse health consequences for mothers and children. Due to increasing obesity and maternal age, the worldwide prevalence of GDM has doubled in the last five years. GDM currently affects approximately one in six pregnancies worldwide and one in five pregnancies in the UK [2-6]. Cost-effective, safe and clinically plausible management solutions are urgently needed to improve the quality of health and care in response to the increased demand for maternity and diabetes clinical services.

In many places around the world, including the UK, GDM management initially follows a ‘one size fits all approach’ with little consideration of individual differences. Management begins with education, dietary and lifestyle advice. Capillary blood glucose (BG) testing is recommended, fasting, pre- and post-meal between 4 to 6 times a day, until the birth of the baby. Depending on the BG response, some women with GDM need oral medications or insulin intervention for their BG management, whereas some can manage BG via diet and exercise. Conscious glucose monitoring (CGM) is a more expensive alternative for glucose monitoring and management. CGM repeats sampling of interstitial glucose by a fine capillary via an adhesive sensor on the skin. Whilst transformative for pregnant women with type 1 diabetes, CGM technologies are expensive with a lack of evidence to support their use in GDM. As such, fingerstick monitoring remains the recommended method of glucose monitoring in GDM.

The objectives of this paper are to (1) develop benchmark glucose prediction models with LSTM-based deep neural network (DNN) models using time-series data collected from the GDm-Health platform, (2) compare the prediction accuracy with published results of LSTM-based models using CGM data, and (3) suggest an optimized clinical review schedule with the potential to reduce the overall number of blood tests for mothers with stable and normal glucose measurements.

This study builds on the GDm-Health development at the Institute of Biomedical Engineering and the Nuffield Department of Women's Reproductive Health at the University of Oxford. GDm-Health is the UK's first clinical gestational diabetes management platform [7]. GDm-Health is a smartphone-based, Bluetooth-controlled BG monitoring platform with a bidirectional communication system between patients and clinicians to enable remote patient monitoring [8, 9]. The GDm-Health device is a CE-marked medical device and can be prescribed via health providers. The GDm-Health platform is currently being used by more than 50 NHS Foundation Trust Hospitals and covers more than 50% of the gestational diabetes population in England.

Ii. Results

Study Design and Participants

Model development was performed on data from a single site at the Women's Centre in the John Radcliffe Hospital, Oxford University (OUH) NHS Foundation Trust, Oxford, UK. Data were extracted from the electronic medical records of each woman from their first pregnancy booking appointment until the closure of that maternity encounter, following either delivery or discharge from postnatal care, loss to follow-up (e.g., moved away) or miscarriage/termination of pregnancy.

There were 1282 patients with GDM who were managed by the OUH and used the GDm-Health Apps between 14/08/2017 and 10/05/2021. Among them, 1110 patients (1148 pregnancies) were included in this study. The mean maternal age, BMI, parity and gravidity were 33.5 yrs, 31.0, 1.6 and 2.6, respectively.

Eligibility criteria were pregnant women with GDM at the OUH, with singleton pregnancies, and who recorded their BG data on the GDm-Health system and had an electronic health record for the pregnancy. For this study, GDM was defined as per the hospital guidelines at the time of pregnancy. Throughout the study period, all pregnant women underwent screening using clinical risk factors (NICE). Those with risk factors underwent a 2-hour 75-gram glucose tolerance test (GTT). Until April 2020, GDM was diagnosed using the thresholds from WHO/IADPSG (≥5.1 fasting, ≥10.0 at 1 hour, ≥8.5 mmol/L at 2 hours) at approximately 24-26 weeks of pregnancy. From April 2020-March 2021, GDm diagnosis was based on the RCOG COVID pathway (fasting glucose ≥5.3 mmol/L, HbA1c ≥39 mmol/mol at approximately 28 weeks). Following March 2021, NICE thresholds for the GTT were used to diagnose GDM (5.3 fasting. 7.8 mmol/L 2-hours). We excluded women who had not consented for their anonymised data to be shared through GDm-Health and women who opted out of the use of their data in health research.

Women using the GDm-health system to monitor their BG levels were suggested to test a minimum of three days of the week, measuring their BG four times a day (before-breakfast: BB, one hour after-breakfast: AB, one hour after-lunch: AL, and one hour before-dinner: BD). Before-lunch (BL) and before-dinner (BD) tests were also recommended when taking glucose-lowering medication (metformin or insulin). We refer to these six measurements as “Tags AB, AL, AD, BB, BL, BD” in this paper. The flow chart of patients, data preprocessing and the number of time series used in the three models are reported in Fig. 1.

The distributions of BG readings after window preparations are shown in Fig. 2. In general, all six tags demonstrated a good fit of normal distribution with one peak and moderate symmetricity. The distribution of all after-meal measurements demonstrated good symmetrical distributions, and all before-meal measurements had a small tail (stew right) that represented the Hyper group. This observation matches the time-series size reported in Table 1.

Beyond the time-series BG readings of three prediction windows, readings were subgrouped into Hyper, Normal, and Hypo, which refers to the value of glucose measurements above, within or under the NICE guidance [10]. The Hyper (High) group had one of more BG readings >7.8 mmol/L for 1-hour postprandial measurements or >5.3 mmol/L for fasting measurements within the observation window. The Normal group had all readings within the target range, and the Hypo group (Low) had at least one reading <3.0 in the observation window.

Model performance

We first developed and tested the Stacked-LSTM models for 7 predict 7 days, 7 predict 14 days, and 14 predict 14 days for all windows. By comparing the three models’ mean absolute error (MAE) and root mean square error (RMSE) side by side, as shown in Fig. 3, we can confirm that (1) before-breakfast is the most “predicable” BG among all six before- and postmeal measurements and (2) the 7-day-predict-14 day is the optimized BG monitoring timeframe.

We then chose the 7-day predicted 14-day window frame as an example to evaluate the performance of Stacked-LSTM under two subgroups, namely, the Normal and Hyper subgroups, as defined in the Methods section. For women who have hypoglycaemia, our system will provide an alert; no prediction model was trained on the Hypo subgroup data. The characteristics and model performance of the subgroups are listed in Table 1. The results shown in the Table 1 confirmed that BG prediction in the Normal subgroup had smaller RMSE and MAE values than that in the Hyper subgroup. In conclusion, all pre-meal predictions are more accurately predictable than the post-meal predictions. For the overall group (shown in Fig. 3) and the normal and Hyper subgroups, the RMSE values are higher than the MAE values, suggesting a moderate number of large errors during model testing.

Table 1. Stacked-LSTM model performances in subgroups

	Groups	Time-series*	RMSE (SD)	MAE (SD)
AB: After-breakfast	Hyper (High)	20579	1.169 (0.008)	0.869 (0.005)
	Normal	21230	0.772 (0.003)	0.601 (0.004)
	Overall*	41842	**0.958 (0.007)**	0.728 (0.007)
AL: After-lunch	Hyper (High)	18447	0.876 (0.002)	0.657 (0.003)
	Normal	21273	0.776 (0.001)	0.597 (0.001)
	Overall*	39736	**0.876 (0.003)**	0.656 (0.003)
AD: After-dinner	Hyper (High)	14931	1.010 (0.002)	0.809 (0.002)
	Normal	22905	0.834 (0.004)	0.627 (0.005)
	Overall*	37876	**0.898 (0.003)**	0.692 (0.002)
BB: Before-breakfast	Hyper (High)	32070	0.832 (0.004)	0.472 (0.003)
	Normal	16044	0.316 (0.003)	0.242 (0.003)
	Overall*	48188	**0.622 (0.003)**	0.358 (0.004)
BL: Before-lunch	Hyper (High)	11145	1.002 (0.057)	0.720 (0.051)
	Normal	2720	0.502 (0.010)	0.419 (0.015)
	Overall*	13887	**0.814 (0.009)**	0.597 (0.010)
BD: Before-dinner	Hyper (High)	9750	0.989 (0.010)	0.655 (0.004)
	Normal	3819	0.506 (0.020)	0.400 (0.018)
	Overall*	13645	**0.845 (0.005)**	0.570 (0.004)

* Total number of BG time-series windows used in model development, evaluation and testing, split to 80:10:10 in proportions

** The Overall group include all three subgroups: High, Normal and Hypo.

Table 2. Performance comparisons on GDm-Health and CGM-based LSTM models

Study	**Rabby** *et al.* [11]	**Wang** *et al.* [12]	**Doorn** *et al.* [13]	**Pustozerov** *et al.* [14]	**This paper**
Monitoring Method	CGM	CGM	CGM	CGM	Fingerstick
Patient size for modelling / Diabetes Type	6 / T1DM	56 / T1DM	540 / mixed of normal glucose metabolism, prediabetes, and T2DM	62 / mixed of 48 GDM and 14 normal glucose tolerance	943 / GDM
Patient size for testing / Diabetes Type	No independent testing	No independent testing	6 / T1DM	No independent testing	105 / GDM
Prediction Model	Three-layer Stacked-LSTM	(1) LSTM (2) VMD-LSTM (3) PSO-LSTM (4) VME-PSO-LSTM	LSTM	Linear regression with Lasso regularization	Three-layer Stacked-LSTM
Observation Window	Up to 8 weeks	125 hours	≥ 48 hours	7 days	7 days
Prediction Window	Dynamic One-hour	Dynamic One-hour	Dynamic One-hour	One-hour after meal: next day	One-hour before or after meals **
Accuracy* (RMSE in mmol/L)	0.958	(1) 1.182 (2) 0.507 (3) 0.385 (4) 0.246	1.730	0.870	(1) After meal: **0.911** (2) Before meal: **0.760**

***** Accuracy reported on the test set unless there are no test set results

****** After meal is the mean of AB, AL, and AD, and before meal is the mean of BB, BL, and BD, result reported for the overall cohort model performance for 7 predicted 14-day windows.

In Table 2, we summarize the study design and the performances of the current state-of-the-art BG prediction models using CGM data and then compare them with our model. The comparison results confirmed that the Stacked-LSTM deep learning model can achieve excellent glucose prediction performance similar to that of the CGM system. Fingerstick BG tests can be seen as a snapshot of CGM.

The “events” labels AB, AL, AD, BB, BL, and BD naturally captured the oscillation of the BG pattern. The compression results demonstrated that by using routine glucose collection schedules (4-6 times a day) under clinical platforms such as GDm-Health or similar fingerstick glucose management systems, we can use the Stacked-LSTM model as a promising approach to capture the patterns and seasonality in BG data.

Iii. Discussion

Innovations derived from digital health and machine learning have the potential to improve clinical care delivery and personalize glucose monitoring schedules for people with GDM [15]. The James Lind Alliance Priority Setting Partnerships suggested that a top research priority for women’s health during pregnancy was to use digital health technology to improve pregnancy, birth and mother and child health outcomes [16].

In this paper, we reported, for the first time, a BG predictive monitoring system based on time-series fingerstick BG readings. The results show that the Stacked-LSTM model can predict before- and after-meal BG in the next 7-14 days with an estimate of 10-15% error range (MAE in proportion of mean measurements). Second, the accuracy of our model is on par with the benchmark performance of 1-hour prediction models using CGM data for type 1 and 2 diabetes [11] [12] [13]. As over half of the women with GDM manage their glucose with diet and exercise changes alone, the high additional cost of CGM will be difficult to justify. Our results suggested that adapting the predictive glucose monitoring algorithm into routine fingerstick blood glucose monitoring has the potential to perform comparable patient monitoring as CGM and can potentially improve clinical outcomes and improve the equality of clinical resources.

Clinical machine learning models can identify patterns (digital markers) and disease phenotypes. Several machine learning approaches, including linear regression, decision trees, random forests, and artificial neural networks [11-13, 17], have been used for predicting BG continuously (every 15, 30, 45 and 60 minutes) with CGM glucose machines for type 1 and type 2 diabetes. In these studies, LSTM-based models outperformed other machine learning models. In our study, we used Stacked-LSTM as a benchmark deep learning model with fingerstick BG readings to predictively monitoring the BG trend for women with GDM. LSTM networks have several advantages for BG prediction, including the ability to capture patterns in time-series data, robustness to missing data, high performance and ability to model multivariate data. These advantages make LSTM a promising approach for BG prediction and for improving the management of diabetes and the prevention of complications.

Our models predict the 1 hour before- or after-meal BG for the next 7 and 14 days after each blood test, which is better scheduling of predictive monitoring for clinical intervention. In contrast, studies with CGM use a dynamic 1-hour prediction window running as a background, which is beneficial for type 1 diabetes in pregnancy [18]. There is no evidence that CGM improves clinical outcomes in women with GDM. Our previous model [19] predicted which patients were likely to have glucose readings above target from

those who would not in the next 72 hours. However, given that the effects of medication or dietary changes may take some days, a 7- to 14-day period is more appropriate. Therefore, in women with GDM, we argue that a 7- to 14-day prediction window may be more suitable for clinical management purposes.

One limitation of our work is that the predictive models were built using data from a single site. In addition, the fingerstick test is uncomfortable and inconvenient, especially for pregnant women. Although the test cost of fingersticks is much lower than that of CGM tests, the cost of the test strips is not insignificant in low- and medium-income counties or health systems where women directly pay the test cost.

Whilst our work is a proof-of-principle, we see several potential benefits for clinical decision support to predict 7-to-14-day glucose in patients with GDM. For example, patients achieving their glucose targets could reduce the monitoring frequency. Conversely, predicted high BG readings could provoke early review by the diabetes team for medication review and dose titration. In resource-limited settings, glucose testing strips could be better rationalized. Women who dislike monitoring could be given more personalized testing schedules based on their prior readings encouraging compliance. Caution will also be needed in those receiving insulin therapy where there is a risk of hypoglycaemia. Before clinical implementation, further work is needed to understand how clinicians would use the information from the algorithm and how they would explain this to patients and involve them in decision making about monitoring frequencies. We plan to expand the current study by using federated learning or transfer learning-based LSTM architecture on multiple hospitals in the UK as the next step linking the algorithm with data in other hospitals.

Iv. Methods

Ethics

This study is a part of the ongoing clinical study “Predictive monitoring and management of pregnant women with gestational diabetes mellitus” (REC reference: 21/HRA/3733), which starts in September 2021. This study used routinely collected anonymised, retrospective data. The study was approved by the ethical committee of the UK Health Research Authority, and conducted in accordance with the Declaration of Helsinki. Participants gave their consent in writing.

Data Preprocessing

We used four steps in preprocessing: data cleaning, missing data imputation, observation-prediction sequence preparation, and finally splitting data into training, validation and testing sets (80:10:10 in proportions).

To clean data, we first remove invalid tags, including “PRANDIAL-TAG-OTHER” and “PRANDIAL-TAG-NONE”. Then, we removed BG readings < 1.5 or > 30. The last step is to remove gestational GDays > 294 days (42 weeks 7 days) or < 78 (12 weeks 0 days). After data cleaning, time-series BG data were restructured into multiple time-series BG windows.

A 5-day moving average method was used for missing data imputation by replacing missing values with the average of the values surrounding the missing value based on a specified window size. We hypothesize that blood glucose can be accurately estimated using the previous BG measurements, and that the seasonality of BG remains stable during the observational window and subsequent prediction window. For model testing on unseen data, we used the moving average on the observational window only. Sequence padding is used for the prediction windows by using an index mask to indicate the missing/invalid value (mask = 0) or true value (mask = 1). Compared to linear interpolation, the moving average method is more similar to a human’s physiological response system, which changes gradually over time based on previous BG values.

The next step is preparing the time-series sequence (observation and prediction windows) for model training, validation and testing. In this study, we designed two observation windows (7 and 14 days) and two prediction windows (7 and 14 days). The rationale for the design of seven- and fourteen-day prediction windows is based on the one- to two-week patient review period in clinical practice. This design could enable patients and clinicians to have an accurate estimate of BG status in the upcoming one to two weeks following each blood test. Each extracted sequence will consist of n time steps (days) of BG readings as the observational window, and the model is trained to predict the BG at the next m steps (days). In this paper, we tested on the pairs of n = 7,14 and m = 7 and 14 days, namely, the prediction periods of 7 days prediction 7 or 14 days, and 14 days predict 14 days. Therefore, for each pregnancy, we used overlapping sliding windows with fixed lengths of 14, 21 and 28 days for sequence extraction.

Stacked-lstm Prediction Model

In this paper, we used a deep learning model with a Stacked-LSTM architecture for BG predictive model development. We developed and tested on three models on three monitoring windows, namely, 7 predict 7 days, 7 predict 14 days, and 14 predict 14 days. After choosing the optimized monitoring window, we then developed two independent Stacked-LSTM models to evaluate the model performances of the Hypo and Normal subgroups.

LSTM is a recurrent neural network (RNN) architecture used in deep learning to overcome the vanishing and exploding gradient problems that are often encountered when training time-series data [19]. The memory cell in an LSTM allows information to persist for a self-defined period of time compared to traditional RNNs. The input, output, and forget gates control the flow of information into and out of the memory cell, allowing the network to selectively retain or discard information as needed. This makes LSTMs particularly effective for modelling long-term dependencies in sequential data.

As shown in Fig. 4, the LSTM architecture includes memory cells, gates (input, output, and forget), and activation functions (Adam in our paper). The gate designs allow the LSTM to learn when to “remember” and “forget” information from previous time steps and predict sequential outputs, making it well suited for a two-week BG prediction task. Details of a single LSTM are demonstrated in the dashed line box of Fig. 4, and the equations are shown below.

Let the current input be X_T, and the previous hidden state is H_{T-1}. Then,

$$I_T = \sigma(W_{XI} * X_T + W_{HI} * H_{T-1} + W_{CI} \circ C_{T-1} + B_I) \text{ (Eq. 1)}$$

$$F_T = \sigma(W_{XF} * X_T + W_{HF} * H_{T-1} + W_{CF} \circ C_{T-1} + B_F) \text{ (Eq. 2)}$$

$$O_T = \sigma(W_{XO} * X_T + W_{HO} * H_{T-1} + W_{CO} \circ C_T + B_O) \text{ (Eq. 3)}$$

$$C_T = F_T \circ C_{T-1} + I_T \circ tanh(W_{XC} * X_T + W_{HC} * H_{T-1} + B_C) \text{ (Eq. 4)}$$

$$H_T = O_T \circ tanh(C_T) \text{ (Eq. 5)}$$

Here, $*$ is the convolution operator, and $\circ$ is the Hadamard product (also called the elementwise product). W_{XI}, W_{XF}, W_{XO}, W_{XC}, W_{HI}, W_{HF}, W_{HO} and W_{HC} represent the convolutional filters. B_I, B_F, B_O and B_C are the biases for each layer. The input gate I_T controls which part of the new input information will be kept in the long-term state. The forget gate F_T decides which part of the long-term state is removed. The output gate O_T decides which part of the long-term state is read. C_T is the long-term state. Short-term state H_T is the representation of BG in the most recent days; its variation is more likely caused by diet, exercise and medication.

These equations define the forwards pass of a single LSTM cell at each time step. In a three-layer LSTM network, the hidden state from one LSTM layer is passed as input to the next two layers at each time step. The final hidden state can then be used for prediction, such as regression in our model.

Our models have three stacked LSTMs to increase the likelihood of detecting long-term dependencies of the BG in the next few days. The advantage of using a stacked LSTM network over a single-layer LSTM network is that a stacked LSTM network has a greater capacity to model complex relationships in the input data. By stacking multiple LSTM layers, the network can learn increasingly abstract representations of the input data, which can improve its ability to make predictions and generalize to new data [20].

We ran the LSTM architecture with a number of hyperparameter settings. The number of input channels matches the observation window (7 or 14, respectively), and there is one hidden channel in each LSTM layer. We set the dropout to 0.2, which means the dropout probability on the outputs of each LSTM layer except the last layer. The final output of the three LSTM layers will provide two outputs, BG prediction and hyperglycemia and hypoglycemia alerts. The output of BG prediction is a BG sequence responding to a

fixed prediction window (7 or 14 days). In our models, we have two prediction windows, 7 or 14 days after the last BG test.

Model Training, Validation And Testing

The LSTM model is trained and validated under the pipeline shown in Fig. 2. The inputs of the model are six time-series BG data in response to the readings before and after each meal. The readings among different tags (AB, AL, AD, BB, BL, and BD) are treated as independent of each other.

As shown in Fig. 2, the model training includes hyperparameter tuning via batch learning and local optimization. Batch learning is used in our model to efficiently train deep learning models. To implement batch learning in the Stacked-LSTM model, we divided the training data into batches of size 32. Batch normalization was carried out to regularize the activations in the network, reducing overfitting and improving generalization.

The LSTM network weights were first initialized. Within each batch, the training model performed a forwards pass through the three-layer LSTM network to calculate the predictions. The mean squared error, which is the square of RMSE, is used as the loss function. The loss between the predictions of the BG sequence and the imputed BG values (preprocessed data with imputation) was calculated before performing a backwards pass to calculate the gradients of the loss for weight update. After the third layer of the LSTM, Adam, an optimization algorithm, is used to perform 100-epoch hyperparameter tuning. In our experiment, the loss converges at approximately 20–30 epochs in all Stacked-LSTM model training. The weight of the Stacked-LSTM is then updated. Using batch learning can help to regularize the activations in the network, reducing overfitting and improving generalization. After the Stacked-LSTM models have been trained, the next step is to evaluate the models using performance metrics with the evaluation data.

At the model testing step, we used trained models to make predictions on a separate independent holdout dataset, which is 10% of the BG time-series. Model testing will evaluate the model generalization on unseen data and highlights issues such as model overfitting. We run each model 5 times, calculate the mean and the standard deviation of the performance metrics on the test dataset and report the mean and standard deviation of RMSE and MAE. One difference between the model test and evaluation is that we calculated the errors between the predicted values and the true values (raw data without imputation).

The performance metrics for the training, validation and test sets are the same, including RMSE and MAE. RMSE is the square root of the average squared errors before errors are averaged. MAE is the mean absolute error. RMSE will highlight the occurrence of large errors in comparison to MAE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(pred_i - target_i)^2} \quad \text{(Eq. 6)}$$

$$MAE = \frac{1}{N}\sum_{i}^{N}|(pred_i - target_i)| \quad \text{(Eq. 7)}$$

Declarations

DATA AVAILABILITY

The dataset used in this study is protected under the ethical approval protocol.

CODE AVAILABILITY

Code is available from the corresponding author on reasonable request. The data preprocessing is carried out using MATLAB R2022a. Deep learning model training and evaluation were carried out using PYTHON 3.6.

AUTHOR CONTRIBUTION

H.L. conceptualized the study, provided research governance and protocol development, analyzed data, and drafted the manuscript. P.L. analyzed data and drafted the manuscript. J.H. provided protocol development and collected data. L.M. conceptualized the study and provided research governance. D.C. provided protocol development and funding support. All authors approved the completed manuscript.

ACKNOWLEDGEMENTS

H.L. is supported by the Royal Academy of Engineering Daphne Jackson Trust Fellowship grant and an EPSRC Healthcare Technologies Challenge Award (EP/N020774/1). P.L. is supported by the Wellcome Trust under grant 217650/Z/19/Z. J.H. is supported by the UKRI Future Leaders Fellowship grant. D.C. was supported by the NIHR Oxford Biomedical Research Centre, an NIHR Research Professorship grant, an RAEng Research Chair grant, the InnoHK Hong Kong Centre for Cerebrocardiovascular Health Engineering, and the Pandemic Sciences Institute. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

COMPETING INTERESTS:

The commercial licence of GDM-Health was transferred to Sensyne Health and then acquired by HUMA Health. The authors have no commercial affiliation or collaboration with HUMA Health plc. LM is a part-time employee of EMIS plc.

References

1. K. G. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation," *Diabet Med,* vol. 15, no. 7, pp. 539-53, Jul 1998.
2. A. Garrison, "Screening, diagnosis, and management of gestational diabetes mellitus," (in English), *Am Fam Physician,* vol. 91, no. 7, pp. 460-467, Apr 1 2015.
3. W. W. Zhu, H. X. Yang, C. Wang, R. N. Su, H. Feng, and A. Kapur, "High prevalence of gestational diabetes mellitus in Beijing: effect of maternal birth weight and other risk factors," *Chin Med J,* vol. 130, no. 9, pp. 1019-1025, May 5 2017.
4. C. C. Martis R, Shepherd E, Alsweiler J, Downie MR, Brown J., "Treatments for women with gestational diabetes mellitus: an overview of cochrane systematic reviews " *Cochrane Db Syst Rev,* REVIEW no. 8, 2018.
5. E. Bain, P. Middleton, and C. A. Crowther, "Progressing towards standard outcomes in gestational diabetes Cochrane reviews and randomised trials," *Aust N Z J Obstet Gynaecol,* vol. 56, no. 1, pp. 113-6, Feb 2016.
6. P. Saravanan, G. Diabetes in Pregnancy Working, G. Maternal Medicine Clinical Study, O. Royal College of, and U. K. Gynaecologists, "Gestational diabetes: opportunities for improving maternal and child health," *Lancet Diabetes Endocrinol,* vol. 8, no. 9, pp. 793-800, Sep 2020.
7. The UK National Institute of Health and Care Excellence. "Health app: GDm-Health for people with gestational diabetes." https://www.nice.org.uk/advice/mib131 (accessed on 15 Jan 2023)
8. J. E. Hirst *et al.*, "GDm-health: development of a real-time smartphone solution for the management of women with gestational diabetes mellitus (GDM)," (in English), *Bjog-Int J Obstet Gy,* vol. 122, pp. 403-403, Apr 2015.
9. J. E. Hirst *et al.*, "GDm-Health: A Pilot Study Examining Acceptability of Mobile Phone Assisted Remote Blood Glucose Monitoring for Women with Gestational Diabetes Mellitus," (in English), *Reprod Sci,* vol. 21, no. 3, pp. 252a-253a, Mar 2014.
10. The UK National Insittute of Health and Care Excellence. "Diabetes in pregnancy: management from preconception to the postnatal period." https://www.nice.org.uk/guidance/ng3/chapter/1-recommendations#antenatal-care-for-women-with-diabetes-2 (accessed on 15 Jan 2023)
11. M. F. Rabby, Y. Tu, M. I. Hossen, I. Lee, A. S. Maida, and X. Hei, "Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction," *BMC Medical Informatics and Decision Making,* vol. 21, no. 1, p. 101, 2021/03/16 2021.
12. W. Wang, M. Tong, and M. Yu, "Blood Glucose Prediction With VMD and LSTM Optimized by Improved Particle Swarm Optimization," *IEEE Access,* vol. 8, pp. 217908-217916, 2020.
13. W. van Doorn *et al.*, "Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study," *PLoS One,* vol. 16, no. 6, p. e0253125, 2021.

14. E. Pustozerov, P. Popova, A. Tkachuk, Y. Bolotko, Z. Yuldashev, and E. Grineva, "Development and Evaluation of a Mobile Personalized Blood Glucose Prediction System for Patients With Gestational Diabetes Mellitus," *JMIR mHealth and uHealth,* vol. 6, no. 1, pp. e6-e6, 2018.
15. R. W. Beck, R. M. Bergenstal, L. M. Laffel, and J. C. Pickup, "Advances in technology for management of type 1 diabetes," *Lancet,* vol. 394, no. 10205, pp. 1265-1273, Oct 5 2019.
16. G. Ayman *et al.*, "The top 10 research priorities in diabetes and pregnancy according to women, support networks and healthcare professionals," *Diabet Med,* vol. 38, no. 8, p. e14588, Aug 2021.
17. M. Armandpour, B. Kidd, Y. Du, and J. Z. Huang, "Deep Personalized Glucose Level Forecasting Using Attention-based Recurrent Neural Networks," *International Joint Conference on Neural Networks,* pp. 1-8, 2021.
18. D. S. Feig *et al.*, "Continuous glucose monitoring in pregnant women with type 1 diabetes (CONCEPTT): a multicentre international randomised controlled trial," (in eng), *Lancet,* vol. 390, no. 10110, pp. 2347-2359, Nov 25 2017.
19. J. Yang *et al.*, "Machine Learning-Based Risk Stratification for Gestational Diabetes Management," (in eng), *Sensors (Basel),* vol. 22, no. 13, Jun 25 2022.
20. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional LSTM Network: a machine learning approach for precipitation nowcasting," Proceedings of International Conference on Neural Information Processing Systems, v1, Montreal, Canada, 2015.
21. P. Lu, H. Qiu, C. Qin, W. Bai, D. Rueckert, and J. A. Noble, "Going Deeper into Cardiac Motion Analysis to Model Fine Spatio-Temporal Features," Springer International Publishing, in Medical Image Understanding and Analysis, pp. 294-306, 2020.

Figures

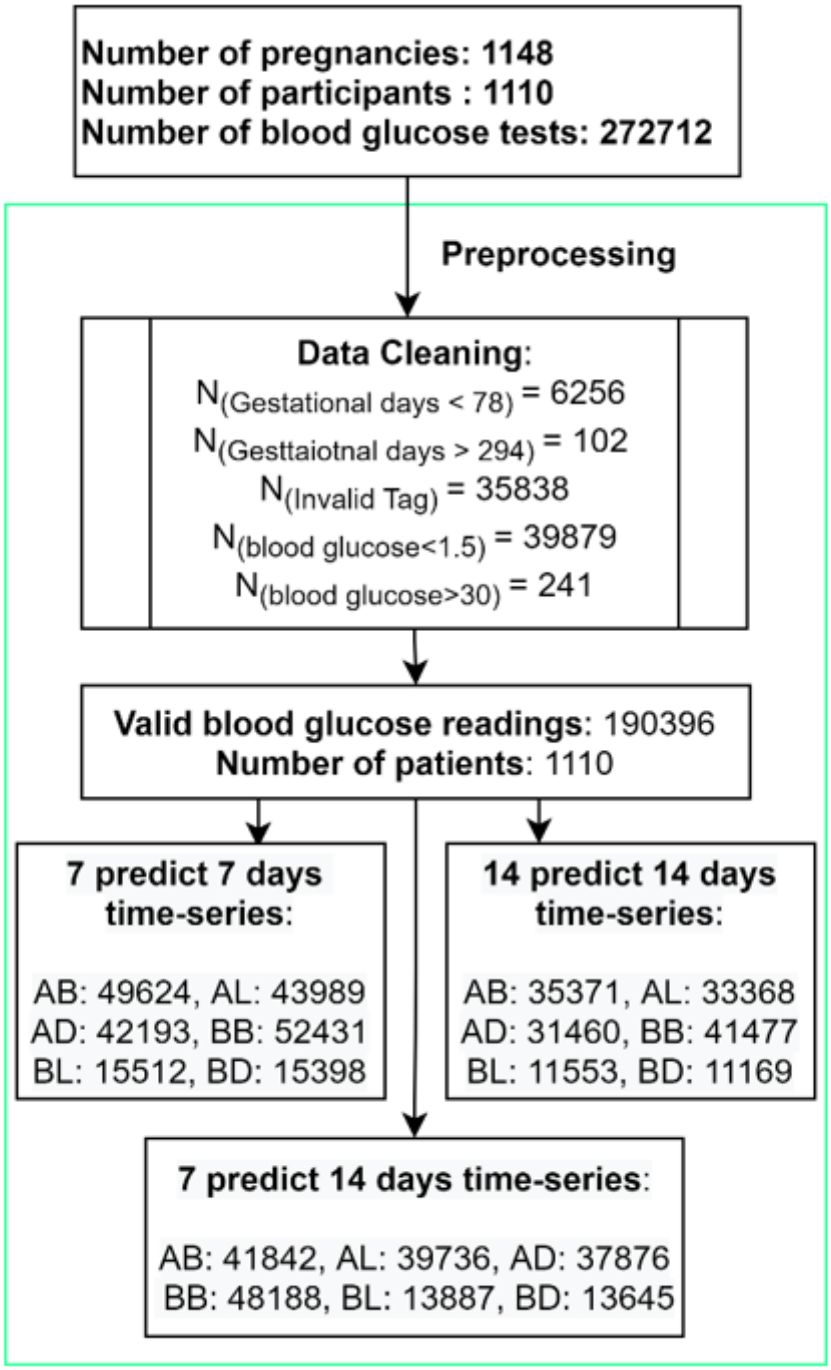

Figure 1

Flow chart of participants, data cleaning and model preparation

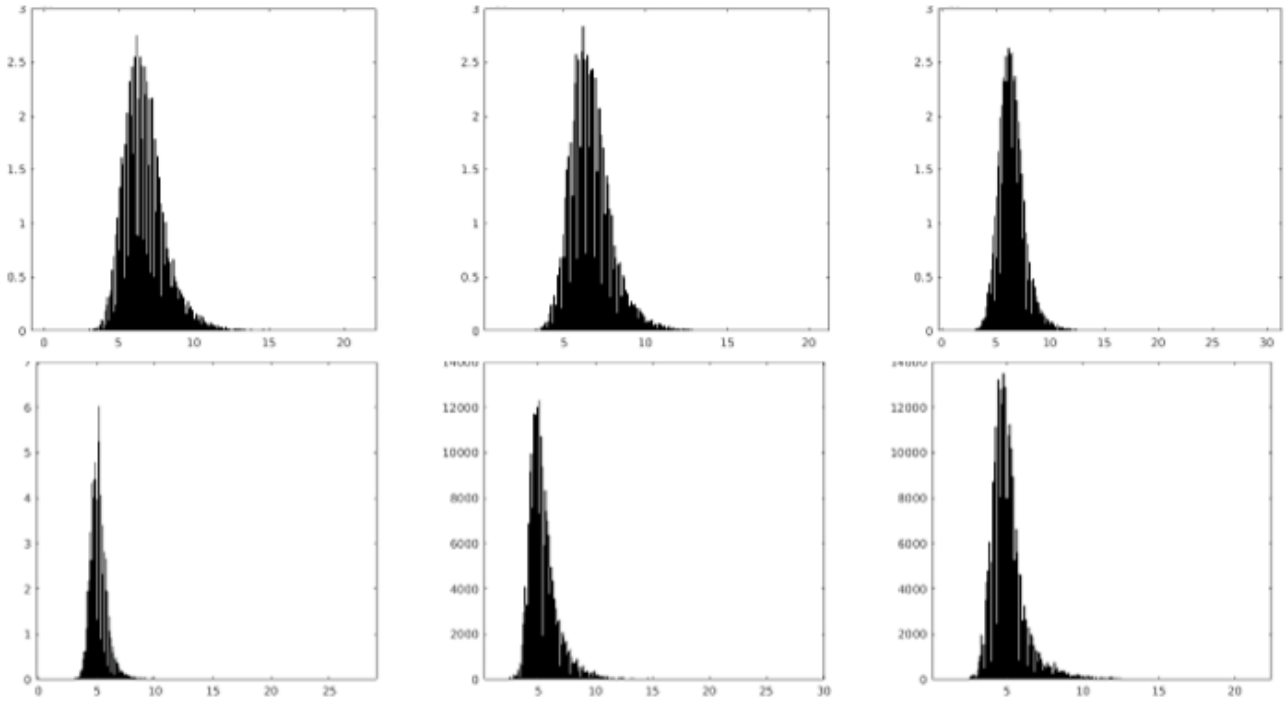

Figure 2

Distribution of BG readings for 7 days predict 14 days (in the order of AB, AL, AD, BB, BL, BD left to right and then the first to the second row), as an exemplar.

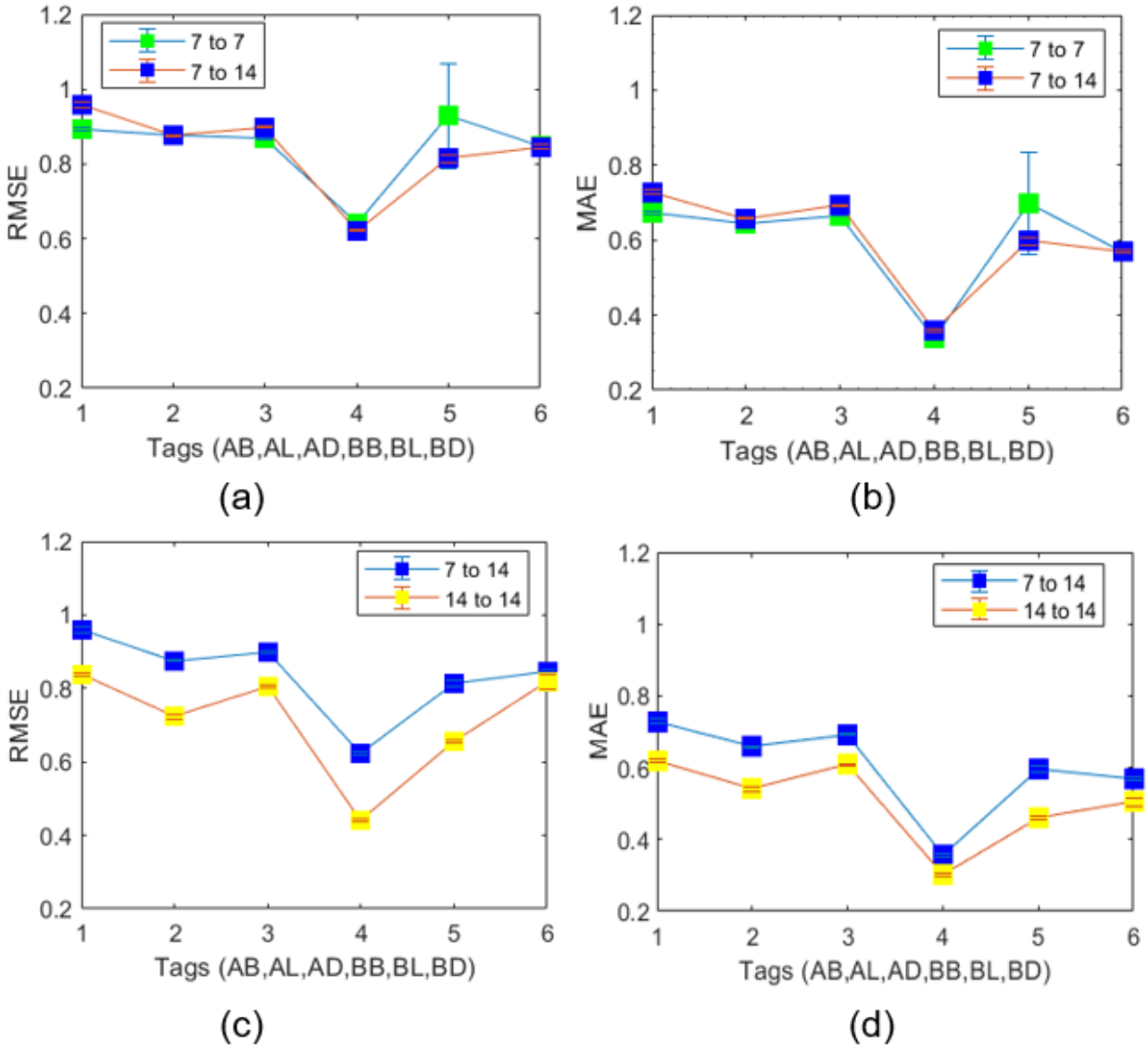

Figure 3

LSTM Prediction Performance Comparisons: (a) RMSE results of 7 predict 7 days *vs* 7 predict 14 days, (b) MAE results of 7 predict 7 days *vs* 7 predict 14 days, (c) RMSE results of 7 predict 14 days *vs* 14 predict 14 days, and (d) MAE results of 7 predict 14 days *vs* 14 predict 14 days.

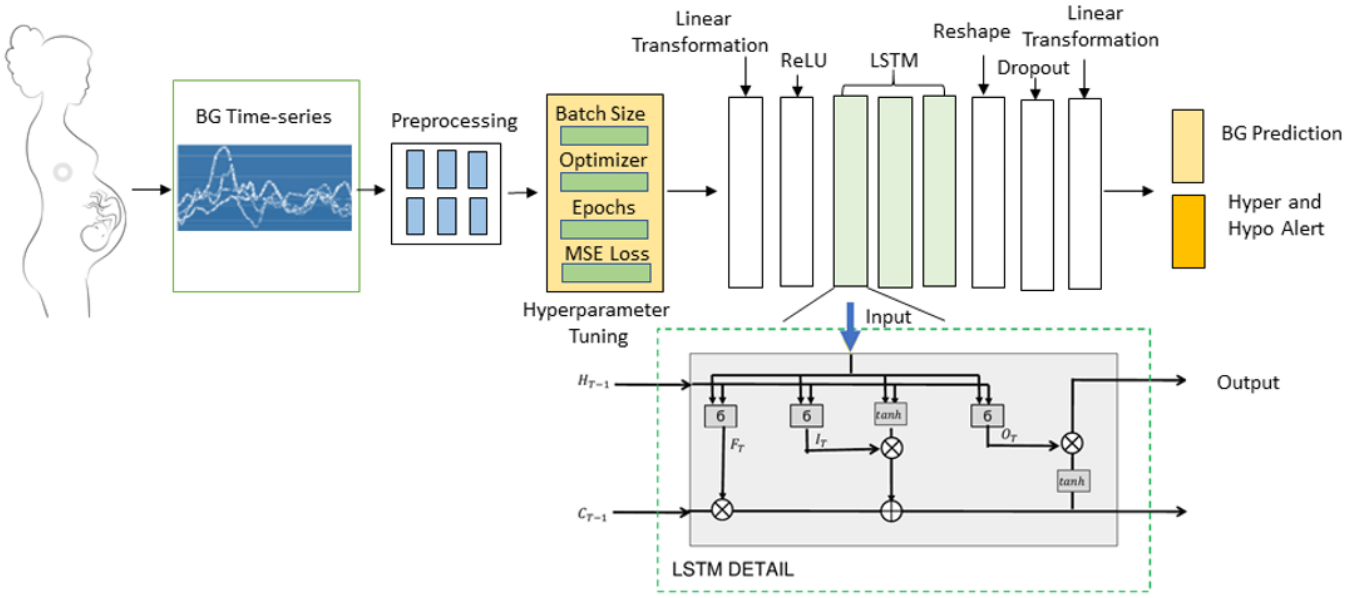

Figure 4

Architecture of three-layer stacked-LSTM and its model development pipeline