Shortcomings in the Evaluation of Blood Glucose Forecasting

Jung Min Lee, Rodica Pop-Busui, Joyce Mee Kyoung Lee, Jesper Fleischer, Jenna Wiens

Submitted to: JMIR AI
on: January 19, 2023

Table of Contents

Shortcomings in the Evaluation of Blood Glucose Forecasting

Jung Min Lee[1] MEng; Rodica Pop-Busui[2] MD, PhD; Joyce Mee Kyoung Lee[3] MD, MPH; Jesper Fleischer[4, 5] MSc, PhD; Jenna Wiens[1] PhD

[1]Division of Computer Science and Engineering University of Michigan Ann Arbor US
[2]Division of Metabolism, Endocrinology and Diabetes Department of Internal Medicine University of Michigan Ann Arbor US
[3]Division of Pediatric Endocrinology Susan B. Meister Child Health Evaluation and Research Center University of Michigan Ann Arbor US
[4]Steno Diabetes Center Aarhus Aarhus DK
[5]Steno Diabetes Center Zealand Holbaek DK

Corresponding Author:
Jenna Wiens PhD
Division of Computer Science and Engineering
University of Michigan
2260 Hayward St
Ann Arbor
US

Abstract

Most artificial pancreas systems require a blood glucose (BG) forecasting model that captures the dynamics of the human metabolic system. Machine learning researchers train these models by optimizing for metrics such as root mean squared error (RMSE). However, we found that when combined with a standard controller, models that minimize RMSE do not necessarily yield a higher percent time-in-range (%TIR). We compared the predictive accuracy and control performance of two forecasters: a machine learning-based model that minimizes RMSE (LSTM) and a rule-based model (Loop). Despite achieving RMSE comparable to state-of-the-art (RMSE 15.24mg/dL at 30min), LSTM only achieved 24.35% (IQR 22.35-25.61) TIR. While Loop's prediction accuracy was worse (RMSE 19.50mg/dL at 30min, $p < 0.05$), it achieved higher TIR: 34.20% (IQR 31.25-41.02). Thus, the standard approach to evaluating BG forecasters could lead to poor model selection with respect to improving closed-loop control.

(JMIR Preprints 19/01/2023:45474)
DOI: https://doi.org/10.2196/preprints.45474

Original Manuscript

Short Paper

Shortcomings in the Evaluation of Blood Glucose Forecasting

Abstract

Most artificial pancreas systems require a blood glucose (BG) forecasting model that captures the dynamics of the human metabolic system. Machine learning researchers train these models by optimizing for metrics such as root mean squared error (RMSE). However, we found that when combined with a standard controller, models that minimize RMSE do not necessarily yield a higher percent time-in-range (%TIR). We compared the predictive accuracy and control performance of two forecasters: a machine learning-based model that minimizes RMSE (LSTM) and a rule-based model (Loop). Despite achieving RMSE comparable to state-of-the-art (RMSE 15.24mg/dL at 30min), LSTM only achieved 24.35% (IQR 22.35-25.61) TIR. While Loop's prediction accuracy was worse (RMSE 19.50mg/dL at 30min, $p < 0.05$), it achieved higher TIR: 34.20% (IQR 31.25-41.02). Thus, the standard approach to evaluating BG forecasters could lead to poor model selection with respect to improving closed-loop control.

Keywords: type 1 diabetes; predictive models; blood glucose forecasting; artificial intelligence; artificial pancreas; machine learning; closed-loop control

Introduction

Blood glucose (BG) forecasting models have emerged as a recent development in therapeutic solutions for type 1 diabetes (T1D) [1–4]. A BG forecast model uses a person's previous BG values and other variables such as insulin and meal history to predict future BG levels over a horizon (e.g., the next hour). Accurate BG forecasters can predict adverse BG events and improve BG management [5–7]. The eventual goal is to use these models to inform insulin dosing in the context of an artificial pancreas (AP), but creating accurate, individualized BG models remains a challenge [8].

To develop accurate BG forecasters, machine learning techniques have been proposed. During evaluation, accuracy is typically measured in terms of error-based metrics such as root mean squared error (RMSE). Recent work has thus been focused on developing new forecasters with the goal of minimizing RMSE [9–11]. However, while the general assumption has been that models with lower/ better RMSE will lead to better glycemic outcomes, we found that this was not necessarily true.

In this paper, we empirically investigate the relationship between BG forecasters' predictive accuracy in terms of RMSE, and the extent to which these models can lead to improved BG management. We also highlight key shortcomings in the current method used to evaluate BG forecasters.

Methods

Datasets

In our experiments, we considered both simulated and real data. We used an open-source version of the UVA/Padova T1D simulator with 10 virtual adults and paired the simulator with a realistic meal schedule (Appendix 1) for data generation [12,13].

Simulated – Patient Behavior Dataset

This dataset mimics a typical individual's behavior. Basal rate was set to default, and bolus size was determined using:

$$bolus=\frac{mealsize}{CR}+(b_g>150)\times\frac{b_g-b_t}{ISF}$$

where CR is the carbohydrate ratio, ISF the insulin sensitivity factor, b_g the current BG, and b_t the target BG (140 mg/dL). To mimic errors in carbohydrate estimation, $mealsize$ was randomly set as 80-120% of the true meal size. Boluses occurred anywhere between 15 minutes prior and after a meal.

Real – Ohio T1D Dataset

The OhioT1DM dataset is a clinical dataset released for the Blood Glucose Level Prediction Challenge.[14] It contains 8 weeks' worth of continuous glucose monitoring, insulin dose, and self-reported life-event data for 12 people (5 females) with T1D [15].

Forecasting Models

We investigated the accuracy of two forecasting models.

LSTM

A long short-term memory (LSTM) model is a recurrent neural network used in many BG forecasting models [9,16–18]. In our experiments, we used an LSTM model proposed by Mirshekarian et al [18], which takes the previous 2 hours of BG, meal, bolus, and basal insulin data sampled every 5 minutes as input. Model parameters were tuned to minimize the mean squared error between the predicted and true BG values at the next time step, and the final model was selected based on validation RMSE. During inference time, the model was autoregressively applied until the desired prediction horizon was reached (Appendix 2).

Loop

We used the forecaster provided in Loop, an open source DIY AP system [19]. The two parameters for insulin effect were set to the default setting recommended for adults, and the expected absorption time for each meal was set to 1.5 hours. The parameters were verified by matching the predictions made by Loop and the simulator. Note that by design, Loop captures the separate effects of carbohydrates and insulin.

As an upper bound on forecast performance, we also considered a perfect model i.e., an '*Oracle*', which corresponds to the simulator in our experiments.

Control Algorithm

To allow a continuous range of insulin doses, we used random shooting combined with model predictive control (MPC) [20] as the control algorithm and used it to generate simulated data. This algorithm generates k random insulin sequences at each time step. For each insulin sequence, the corresponding BG trajectory is generated using the forecaster. A risk score, defined as the cumulative discounted Magni risk [21], is assigned to each BG trajectory. The insulin sequence that yields the trajectory with the lowest risk score is selected, and only the first insulin dose of the sequence is administered. This process is repeated for every time step.

In our setting, doses were limited to boluses and only occurred at the very beginning of the dosing sequence. Bolus sizes were randomly sampled from a patient specific distribution. A 4 hour prediction horizon was used as this encouraged better control performance. 50 candidate sequences were tested at each time step (details in Appendix 3).

Evaluation

The forecasters (and Oracle) were evaluated in terms of predictive accuracy and glycemic outcome when paired with the controller. Prediction accuracy was measured in terms of RMSE on both simulated and real datasets. This was measured on 150 days of test data for each patient for prediction horizons of 30 minutes and 4 hours using the following equation:

$$RMSE(f,h,D)=\sqrt{\frac{1}{|D|}\sum_{t\in D}\left(f(t+h)-y(t+h)\right)^2}$$

where f is the forecaster, D the entire dataset, h the prediction horizon, $f(t+h)$ the hth point in the prediction made by the forecaster at time t, and $y(t+h)$ the hth point in the prediction made by the Oracle at time t. 95% confidence intervals were calculated using 1000 bootstrapped samples.

Glycemic outcome was measured on 100 days of test data for each individual, broken into 20 independent, 5 day-long episodes. Three metrics were measured: % time-in-range (>70 and <180 mg/dL), % time-below-range (≤ 70 mg/dL), and % time-above-range (≥ 180 mg/dL), with the interquartile range across all individuals.

Sensitivity Analyses

Each forecaster's ability to distinguish the individual effects of carbohydrates and insulin was also evaluated by measuring the RMSE over a subset of data where only carbohydrates or insulin was present in the input. For qualitative evaluation, we aligned and compared predictions that were made 25 minutes after an insulin or carbohydrate event.

Results

Applied to the simulated test data, LSTM achieved an RMSE comparable to that of state-of-the-art machine learning forecasters [18] and better than Loop (RMSE 15.24 mg/dL (15.21-15.27) vs 19.50 mg/dL (19.45-19.54) at 30 minutes), but this did not translate into better glycemic outcomes (Table 1). LSTM had lower time-in-range (TIR) compared to Loop (24.35% (22.35-25.61) vs. 34.20% (31.25-41.02)). These findings contradict the widely held assumption that a better forecaster (in terms of RMSE) will necessarily lead to better control.

When only insulin (Figure 1) or carbohydrate (Figure 2) is present in the input, Loop predicts a BG trajectory that aligns with clinical understanding (i.e., insulin leads to a decrease and carbohydrates lead to an increase). On the other hand, LSTM predicts trajectories that mimic those when insulin and carbohydrates are administered in tandem. However, in terms of RMSE over this subset of data (Figures 3, 4), LSTM outperforms Loop in most cases with the exception of longer prediction horizons in the presence of carbohydrates.

Table 1. Prediction accuracy [□] and glycemic outcome [b] for each forecaster.

Forecaster [c]	Prediction Horizon	RMSE Patient Behavior	RMSE OhioT1D (real)	Control Performance on Simulated Data [d]		
				% TIR	%TBR	%TAR

		(simulated)				
Oracle (upper bound)	30 min	0.00 (0.00, 0.00)	–	92.83 (87.80, 96.44)	0.69 (0.00, 4.73)	4.29 (1.04, 10.26)
	4 hours	0.00 (0.00, 0.00)	–			
LSTM	30 min	15.24 (15.21, 15.27)	22.03 (21.32, 22.75)	24.35 (22.35, 25.61)	73.91 (72.98, 74.39)	2.17 (0.78, 4.57)
	4 hours	29.42 (29.32, 29.52)	75.31 (73.52, 77.21)			
Loop	30 min	19.50 (19.45, 19.54)	60.32 (48.59, 72.10)	34.20 (31.25, 41.02)	0.00 (0.00, 0.15)	64.19 (58.31, 67.79)
	4 hours	44.07 (43.95, 44.18)	80.82 (69.34, 92.40)			

a. Prediction accuracy is measured in terms of RMSE at prediction horizons of 30 minutes and 4 hours. Values in parentheses indicate 95% confidence intervals.

b. Control performance (glycemic outcome) is measured in terms of percent of TIR, TBR, and TAR when the forecaster is paired with the control algorithm. Values in parentheses indicate interquartile range.

c. The forecasters are listed in order of decreasing prediction accuracy.

d. Control performance is evaluated on a simulated dataset generated by the selected control algorithm.

RMSE, root mean square error; TIR, time in range; TBR, time below range; TAR, time above range

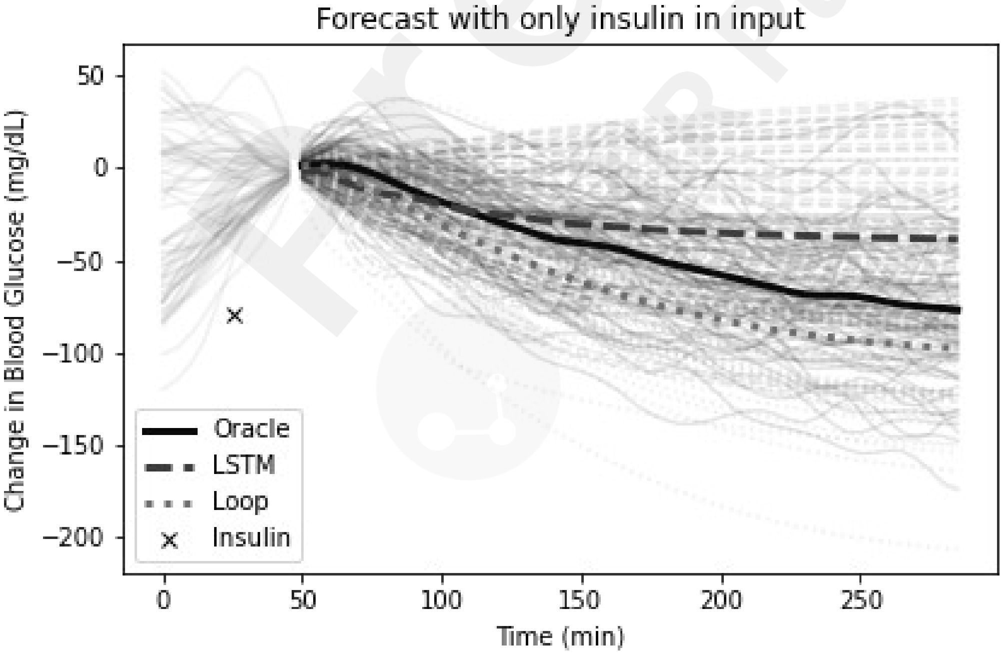


Figure 1. 150 predicted trajectories made by each forecaster when there is a single insulin in the input. Mean trajectories are indicated in bold and predictions were made 25 minutes after the insulin event.

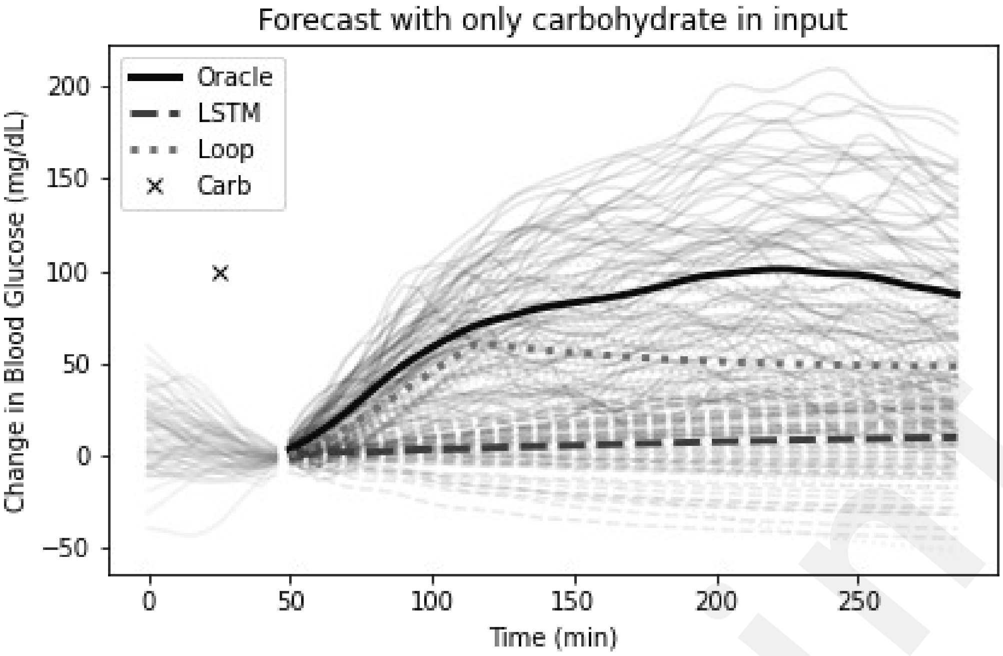


Figure 2. 150 predicted trajectories made by each forecaster when there is a single carbohydrate in the input. Mean trajectories are indicated in bold and predictions were made 25 minutes after the carbohydrate event.

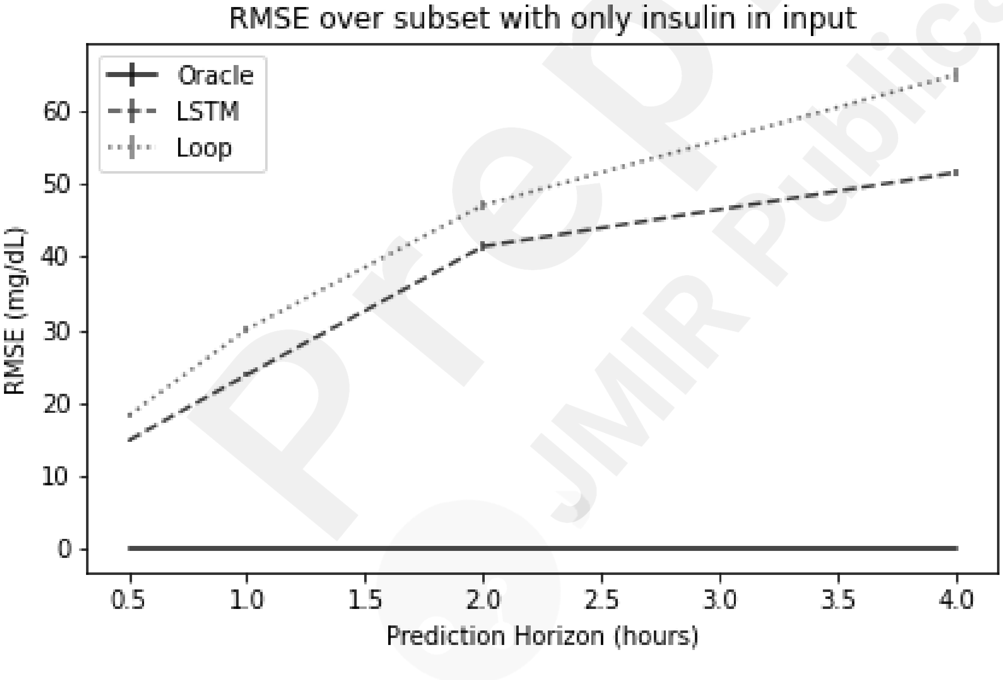


Figure 3. RMSE of each forecaster over a subset of the data where only insulin is in the input.

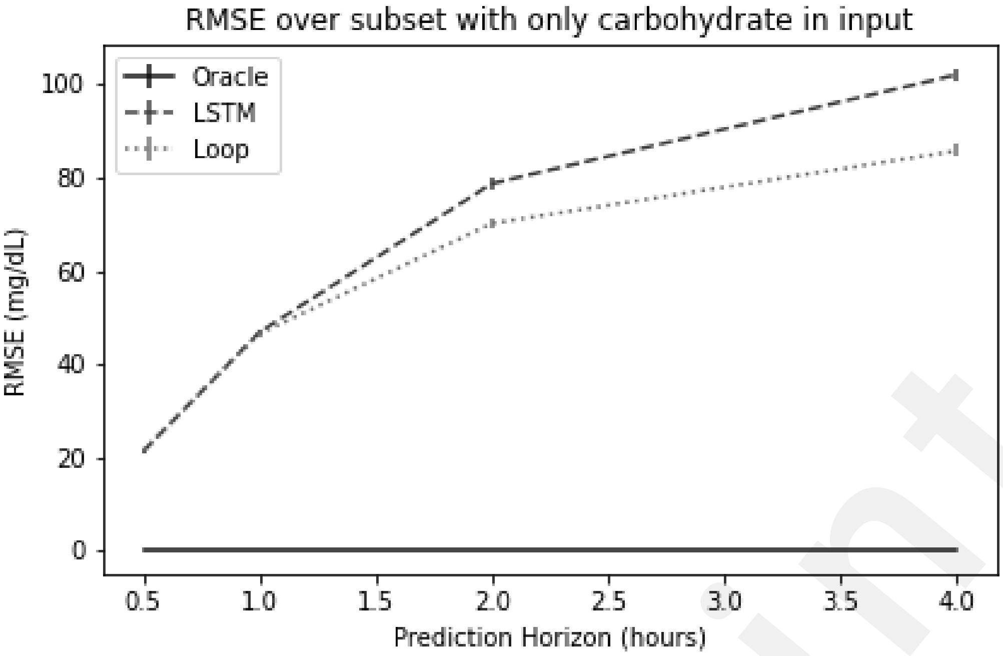


Figure 4. RMSE of each forecaster over a subset of the data where only carbohydrate is in the input.

Discussion

We demonstrate that the current approach of evaluating BG forecasters is not sufficient when selecting forecasters for closed-loop control. We believe our finding is pertinent as recent work has focused on building ‘better’ models that have lower RMSE over short prediction horizons. We thus encourage researchers to explore different evaluation metrics and training strategies when developing predictive models for closed-loop control.

Analyses showed that LSTM, while accurate in terms of RMSE, failed to accurately capture the effects of carbohydrates and insulin. Our qualitative analysis of LSTM’s behavior when carbohydrate or insulin is administered in isolation indicates that LSTM may be conflating the effects of these two variables. We hypothesize that this behavior arises due to the training data, where in almost all cases insulin and carbohydrates are administered together. This restricts the model from learning the individual effects of each variable. In contrast, Loop correctly models the individual effects, in part because they are explicitly modeled and summed together based on prior knowledge.

A key strength of our analysis is the focus on evaluating a forecaster’s ability to predict the individual effects of carbohydrates and insulin. To the best of our knowledge, no other BG forecasting evaluation metric – such as mean absolute error, mean absolute percentage error, or Clarke error grid [22] – have considered this setting explicitly. Our evaluation of these metrics (Appendix 4) show that they too are unable to capture a forecaster’s ability to distinguish the effects of carbohydrates and insulin.

Our study is not without limitations. First, our study cohort was limited to adults. Further study will be required to verify our findings remain consistent across other age groups. Second, while we used a single control algorithm, successful commercialization of hybrid closed-loop systems such as Tandem IQ or Medtronic Automode indicates that other control algorithms may be able to obtain better control. Further study will be necessary to validate our findings with other control algorithms.

Conclusions

Our findings show that forecasters with lower overall RMSE do not necessarily result in better control. We also illustrate how current evaluation techniques fail to identify cases where the forecaster conflates the effects of carbohydrates and insulin, which could contribute to poorer control performance. Going forward, we encourage researchers to evaluate BG forecasters for their ability to accurately predict the independent effects of carbohydrates and insulin, especially when developing models for closed-loop control.

Acknowledgements

J.M.L. and J.W. conceived and designed the study. J.M.L. analyzed the data and drafted the manuscript. J.W. and Joyce M.L. supervised the study. J.W. obtained funding. All authors critically reviewed the article for intellectual content. This work was supported by JDRF Center of Excellence at U of M grant. R.P.B. was also supported by R01DK107956, U01DK119083, 1U01 DK0945157, and R01DK116723.

Conflicts of Interest

Joyce M.L. is on the medical advisory board at Goodrx and is a consultant for Tandem diabetes care. The rest of the authors declare that they have no conflicts of interest concerning this article.

Abbreviations

AP: artificial pancreas
BG: blood glucose
LSTM: long short-term memory
RMSE: root mean squared error
T1D: type 1 diabetes
TAR: time above range
TBR: time below range
TIR: time in range

Multimedia Appendix 1

Algorithm for meal schedule generation.

Multimedia Appendix 2

LSTM model architecture and training details.

Multimedia Appendix 3

Control algorithm implementation details.

Multimedia Appendix 4

Evaluation results for additional metrics.

References

1. Pappada SM, Cameron BD, Rosman PM, Bourey RE, Papadimos TJ, Olorunto W, Borst MJ. Neural Network-Based Real-Time Prediction of Glucose in Patients with Insulin-Dependent

Diabetes. Diabetes Technol Ther Mary Ann Liebert, Inc., publishers; 2011 Feb;13(2):135–141. doi: 10.1089/dia.2010.0104

2. Amar Y, Shilo S, Oron T, Amar E, Phillip M, Segal E. Clinically Accurate Prediction of Glucose Levels in Patients with Type 1 Diabetes. Diabetes Technol Ther Mary Ann Liebert, Inc., publishers; 2020 Aug;22(8):562–569. doi: 10.1089/dia.2019.0435

3. Kushner T, Breton MD, Sankaranarayanan S. Multi-Hour Blood Glucose Prediction in Type 1 Diabetes: A Patient-Specific Approach Using Shallow Neural Network Models. Diabetes Technol Ther Mary Ann Liebert, Inc., publishers; 2020 Dec;22(12):883–891. doi: 10.1089/dia.2020.0061

4. Oviedo S, Vehí J, Calm R, Armengol J. A review of personalized blood glucose prediction strategies for T1DM patients. Int J Numer Methods Biomed Eng 2017;33(6):e2833. doi: 10.1002/cnm.2833

5. Daskalaki E, Nørgaard K, Züger T, Prountzou A, Diem P, Mougiakakou S. An Early Warning System for Hypoglycemic/Hyperglycemic Events Based on Fusion of Adaptive Prediction Models. J Diabetes Sci Technol SAGE Publications Inc; 2013 May 1;7(3):689–698. doi: 10.1177/193229681300700314

6. Imrisek SD, Lee M, Goldner D, Nagra H, Lavaysse LM, Hoy-Rosas J, Dachis J, Sears LE. Effects of a Novel Blood Glucose Forecasting Feature on Glycemic Management and Logging in Adults With Type 2 Diabetes Using One Drop: Retrospective Cohort Study. JMIR Diabetes 2022 May 3;7(2):e34624. doi: 10.2196/34624

7. Garg SK, Weinzimer SA, Tamborlane WV, Buckingham BA, Bode BW, Bailey TS, Brazg RL, Ilany J, Slover RH, Anderson SM, Bergenstal RM, Grosman B, Roy A, Cordero TL, Shin J, Lee SW, Kaufman FR. Glucose Outcomes with the In-Home Use of a Hybrid Closed-Loop Insulin Delivery System in Adolescents and Adults with Type 1 Diabetes. Diabetes Technol Ther 2017 Mar;19(3):155–163. doi: 10.1089/dia.2016.0421

8. Bequette BW. Challenges and recent progress in the development of a closed-loop artificial pancreas. Annu Rev Control 2012 Dec 1;36(2):255–266. doi: 10.1016/j.arcontrol.2012.09.007

9. Rubin-Falcone H, Fox I, Wiens J. Deep Residual Time-Series Forecasting: Application to Blood Glucose Prediction. :5.

10. Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: A Deep Learning Framework for Accurate Glucose Forecasting. IEEE J Biomed Health Inform 2020 Feb;24(2):414–423. doi: 10.1109/JBHI.2019.2931842

11. Zhu T, Li K, Chen J, Herrero P, Georgiou P. Dilated Recurrent Neural Networks for Glucose Forecasting in Type 1 Diabetes. J Healthc Inform Res 2020 Sep 1;4(3):308–324. doi: 10.1007/s41666-020-00068-2

12. Kovatchev BP, Breton M, Dalla Man C, Cobelli C. In Silico Preclinical Trials: A Proof of Concept in Closed-Loop Control of Type 1 Diabetes. J Diabetes Sci Technol SAGE Publications Inc; 2009 Jan 1;3(1):44–55. doi: 10.1177/193229680900300106

13. Xie J. simglucose. 2022. Available from: https://github.com/jxx123/simglucose [accessed Oct

29, 2022]

14. The Blood Glucose Level Prediction Challenge Rules. Available from: http://smarthealth.cs.ohio.edu/bglp/bglp-rules.html [accessed Oct 29, 2022]

15. Marling C, Bunescu R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. CEUR Workshop Proc 2020 Sep 1;2675:71–74. PMID:33584164

16. Jaloli M, Cescon M. Long-term Prediction of Blood Glucose Levels in Type 1 Diabetes Using a CNN-LSTM-Based Deep Neural Network. J Diabetes Sci Technol SAGE Publications Inc; 2022 Apr 25;19322968221092784. doi: 10.1177/19322968221092785

17. Idriss TE, Idri A, Abnane I, Bakkoury Z. Predicting Blood Glucose using an LSTM Neural Network. 2019 Fed Conf Comput Sci Inf Syst FedCSIS 2019. p. 35–41. doi: 10.15439/2019F159

18. Mirshekarian S, Bunescu R, Marling C, Schwartz F. Using LSTMs to learn physiological models of blood glucose behavior. 2017 39th Annu Int Conf IEEE Eng Med Biol Soc EMBC 2017. p. 2887–2891. doi: 10.1109/EMBC.2017.8037460

19. LoopDocs. Available from: https://loopkit.github.io/loopdocs/ [accessed Oct 29, 2022]

20. Nagabandi A, Kahn G, Fearing RS, Levine S. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning. 2018 IEEE Int Conf Robot Autom ICRA 2018. p. 7559–7566. doi: 10.1109/ICRA.2018.8463189

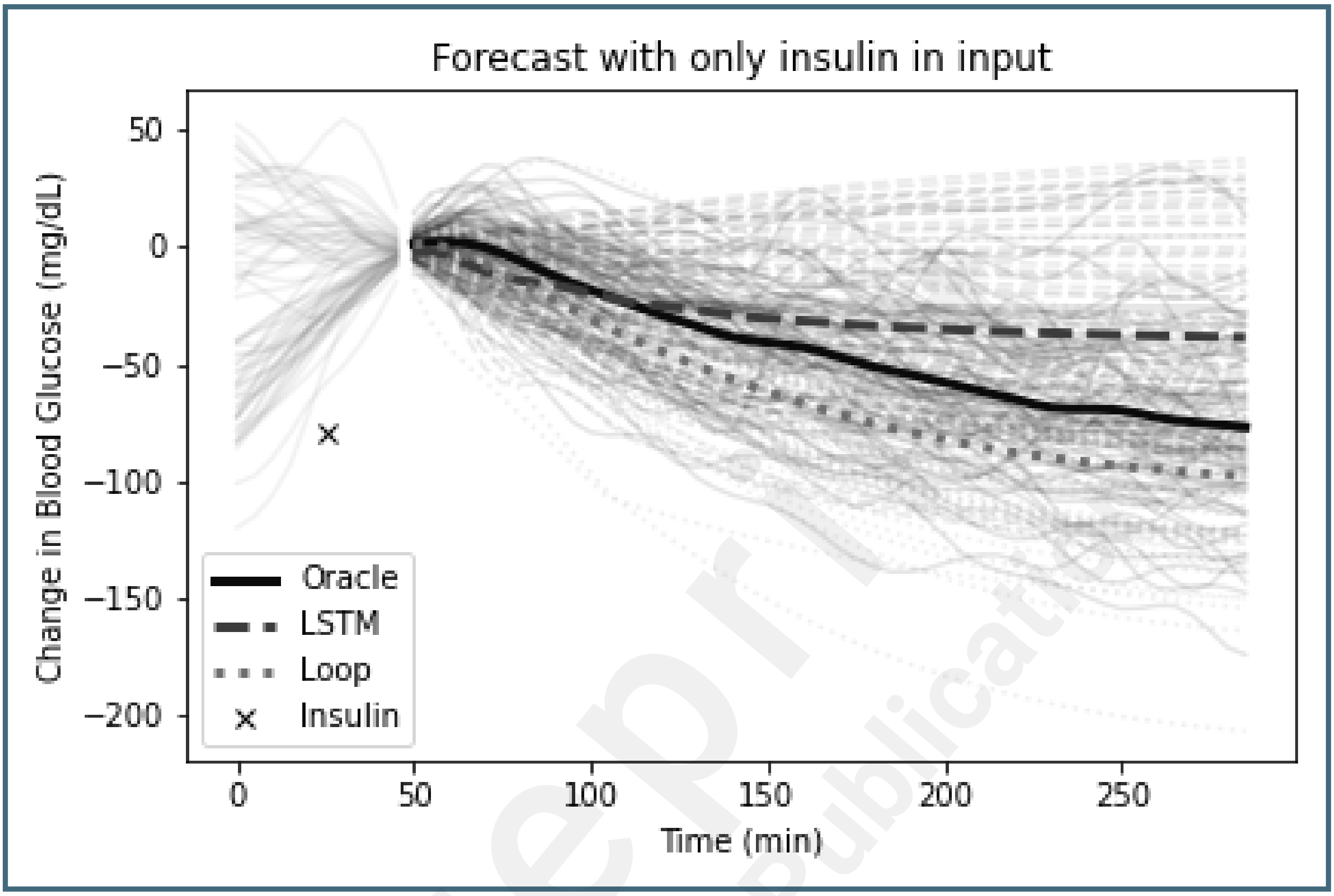
21. Magni L, Raimondo DM, Bossi L, Dalla Man C, De Nicolao G, Kovatchev B, Cobelli C. Model Predictive Control of Type 1 Diabetes: An in Silico Trial. J Diabetes Sci Technol SAGE Publications Inc; 2007 Nov 1;1(6):804–812. doi: 10.1177/193229680700100603

22. Clarke WL. The Original Clarke Error Grid Analysis (EGA). Diabetes Technol Ther 2005 Oct;7(5):776–779. doi: 10.1089/dia.2005.7.776

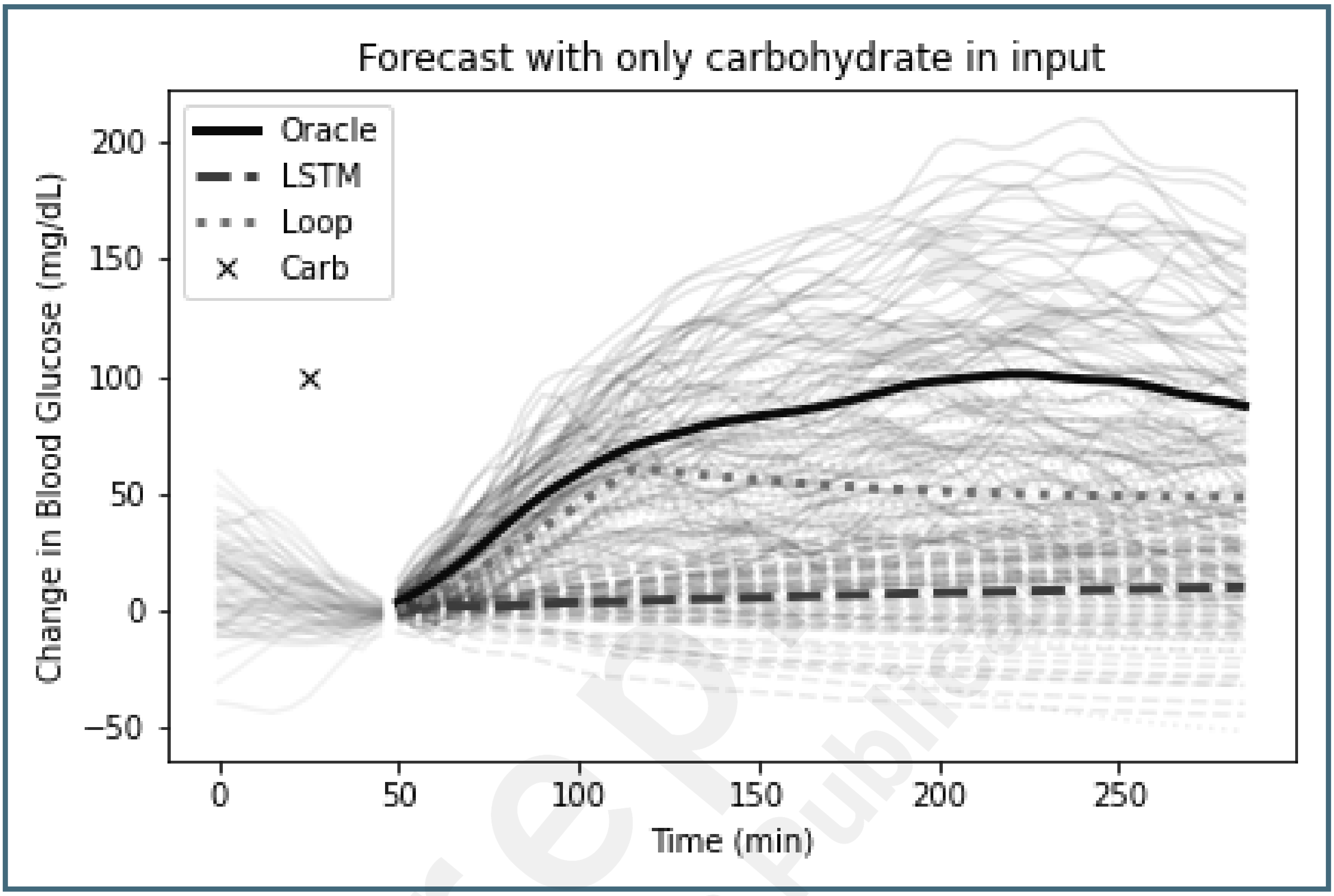
Supplementary Files

Figures

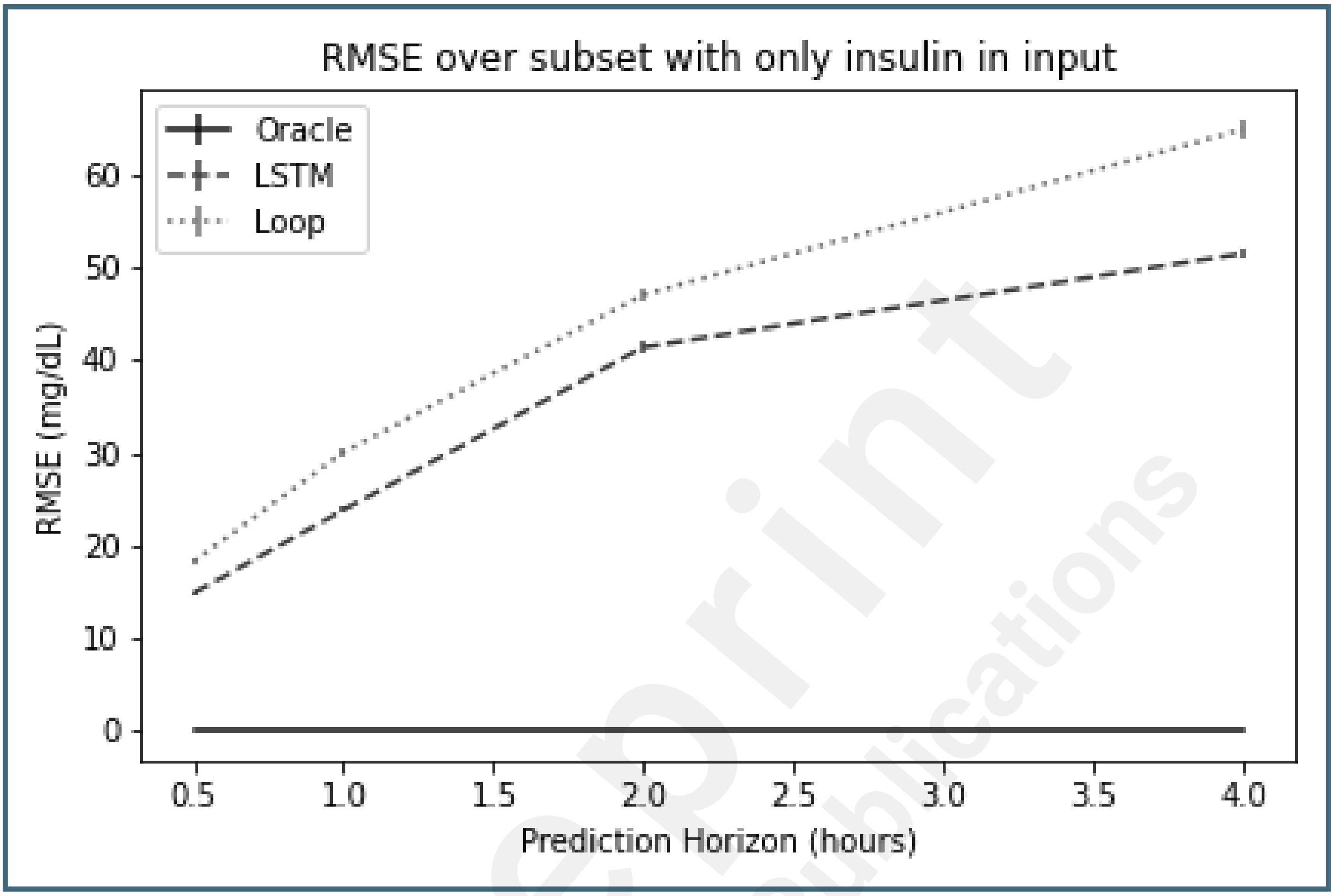
150 predicted trajectories made by each forecaster when there is a single insulin in the input. Mean trajectories are indicated in bold and predictions were made 25 minutes after the insulin event.



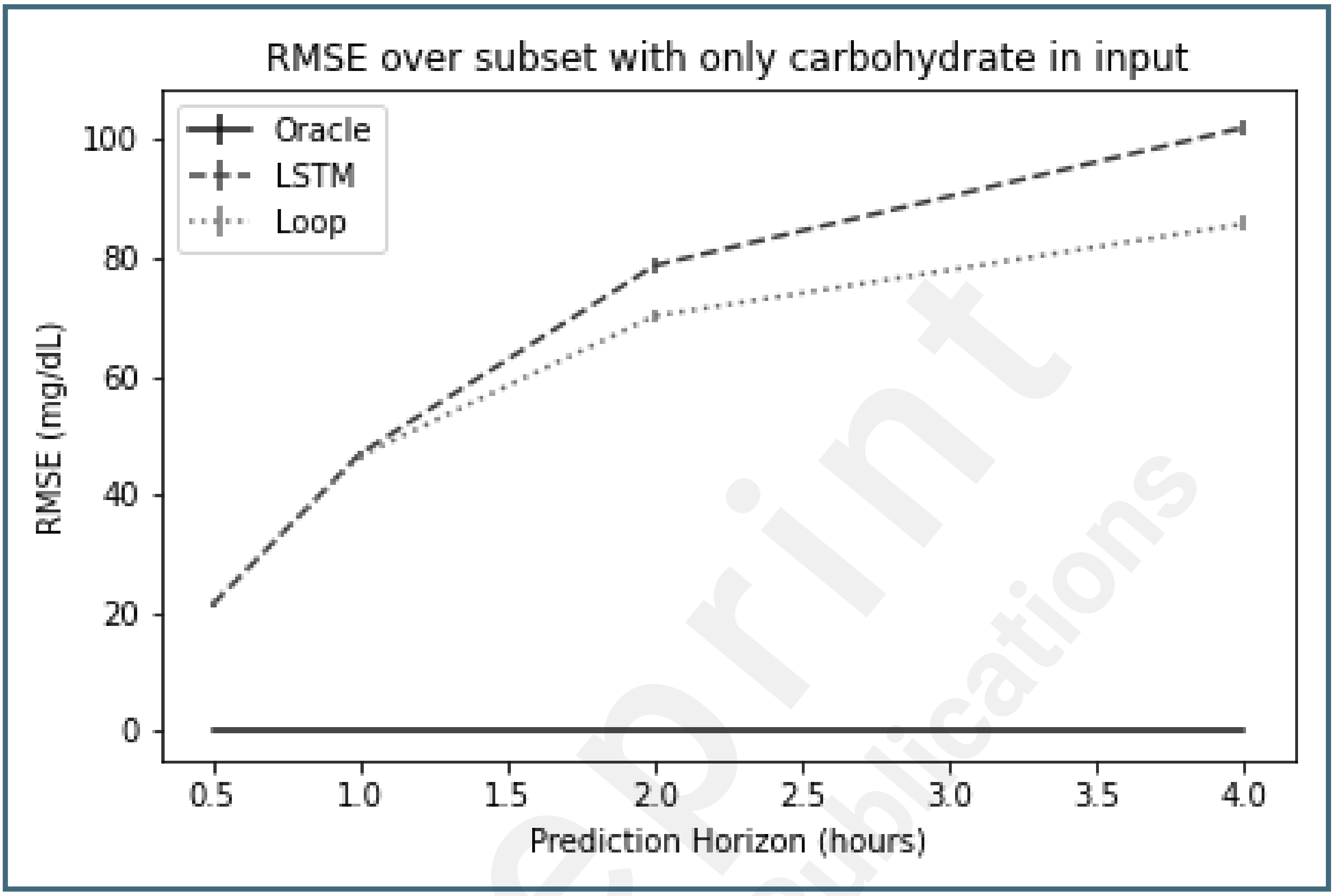
150 predicted trajectories made by each forecaster when there is a single carbohydrate in the input. Mean trajectories are indicated in bold and predictions were made 25 minutes after the carbohydrate event.



RMSE of each forecaster over a subset of the data where only insulin is in the input.



RMSE of each forecaster over a subset of the data where only carbohydrate is in the input.



Multimedia Appendixes

Algorithm for meal schedule generation.
URL: http://asset.jmir.pub/assets/5b295ed1b2a8613d36d69989301c2819.docx

LSTM model architecture and training details.
URL: http://asset.jmir.pub/assets/182dbe1f07fd114f562d31e941d661fe.docx

Control algorithm implementation details.
URL: http://asset.jmir.pub/assets/daa175f911641d816fb3a398ed9ff603.docx

Evaluation results for additional metrics.
URL: http://asset.jmir.pub/assets/a800ea91311bb08036beaaff2568d7f5.docx