Virtual blood glucose monitoring and prediction using machine learning

Ashkan Dehghani Zahedani[1], Arvind Veluvali[1], Saransh Agarwal[1] Jiayu Zhou[1], Jingyi Ruan[1], Michael Snyder[1,2], Nima Aghaeepour[1,2]

[1]January AI

[2]Stanford University

Abstract

Various researchers have attempted to use artificial intelligence (AI) in the prediction of blood glucose (BGL), whether by predicting glucose into the future while the user is wearing a physical CGM device ("Continuous Glucose Prediction"), or by predicting glucose while the user is not wearing a physical CGM device ("Virtual CGM"). Here, we demonstrate the performance of January AI's method of BGL prediction, which incorporates data from a CGM device, a heart rate monitor, users' logs of nutrients and activity, and time of day. This method resulted in increased fidelity of BGL prediction across both continuous glucose prediction and virtual CGM, compared to standard machine learning algorithms.

Introduction

Continuous glucose monitors (CGMs) are medical devices that allow individuals to continuously track their blood glucose levels in real-time. While CGM technology is primarily used to manage diabetes, it can also be useful for individuals without diabetes who are looking to optimize their metabolic health and prevent the development of chronic conditions such as prediabetes and type 2 diabetes.

Maintaining healthy blood glucose levels is important for overall metabolic health, as consistently high or low blood glucose levels can have serious consequences. High blood glucose levels, also known as hyperglycemia, can lead to long-term complications such as nerve damage, kidney damage, and an increased risk of heart disease. On the other hand, low blood glucose levels, or hypoglycemia, can cause symptoms such as dizziness, confusion, and loss of consciousness, and can be potentially life-threatening if left untreated. By continuously monitoring their glucose levels, individuals can identify potential issues and make necessary changes to improve their metabolic health and reduce their risk of developing chronic conditions.

In addition to its benefits for metabolic health, CGM can also improve quality of life for individuals with diabetes. For example, CGM allows individuals to make informed decisions about their insulin doses and dietary choices based on the current state of their glucose levels, which can help to reduce the frequency of hyperglycemic and hypoglycemic events. This can lead to better glycemic control, which can in turn reduce the risk of long-term complications associated with diabetes. Additionally, CGM can provide individuals with diabetes with peace of mind by giving them real-time

insight into their glucose levels, which can help to reduce the stress and anxiety associated with managing this chronic condition.

Despite the many benefits of CGM, adoption of this technology has been slow, particularly among certain populations such as the elderly and those with limited access to healthcare. One reason for this slow adoption is the high cost of CGM systems, which can be a barrier for individuals who do not have insurance coverage or who have high deductible plans. In general, CGM systems can be expensive, with costs ranging from several hundred to several thousand dollars.

In addition to the cost, many individuals find CGM systems to be uncomfortable. The CGM sensor is typically worn under the skin. The sensor is inserted using a needle and is left in place for several days at a time before it needs to be replaced. Some individuals may find the insertion and removal of the sensor to be painful or uncomfortable, particularly if they have sensitive skin or are afraid of needles. Additionally, the sensor site may become irritated or inflamed, which can cause discomfort or pain.

Thus, given the benefits of CGM, as well as the hindrances to both accessing and using this technology, many researchers have sought to eliminate the need for a physical device by employing a predictive model. Georga et al established a predictive metabolic model for patients with type 1 diabetes, using free-living data on mobile devices. The Diabits app described by Kriventsov also attempts to predict blood glucose levels and display predicted values to users. Furthermore, Martinsson utilized recurrent neural networks to predict blood glucose values.

Literature has demonstrated that short-term glucose prediction can be improved by adding meal content information to CGM data. We hypothesized that the incorporation of values from a heart rate monitor (HRM), as well as information about users' nutrition and activity, would further increase the fidelity of short-term glucose prediction. Thus, we developed a unique machine learning model capable of predicting CGM values up to 2 hours in the future for users (continuous glucose prediction, or "CGP"); as well as CGM values for users who were not wearing a physical CGM device, but who had completed an AI training period during which they had worn a CGM device and HRM ("Virtual CGM"/"VCGM"). Embedding this model in our mobile application ("January V1"), we demonstrated that our model can produce both CGP and VCGM with a high degree of fidelity.

Literature Review

Researchers have attempted to predict CGM curves for at least 10 years. Georga et al. used Support Vector regression (SVR) and Random Forest regression (RFR) to predict CGM curves for 27 patients with Type 1 Diabetes. Both models took as input a user's CGM profile over time, and simultaneously integrated insulin injection data, food intake, and exercise expenditure. SVR was able to achieve a test error of 5.21 mg/dL and 7.14 mg/dL at 30 and 60 minutes, respectively. RF was able to achieve 8.15 mg/dL and 9.25 mg/dL at 30 and 60 minutes, respectively. Despite these promising results, other researchers have raised concerns about the model; the correlation of 0.99 could be a sign of overfitting, and, furthermore, George et al. have as of this writing not provided information concerning the data and what standards upon which they based their test data.

In 2016, Chiara Zecchin interrogated the utility of incorporating insulin data in CGM prediction. Zecchin's research evaluated the benefit of including exogenous insulin and carbohydrate data into a

neural network-based model incorporating four inputs (CGM, DCGM, Insulin, and Carbohydrates) in 15 T1D subjects over 3 consecutive days. Zecchin's research determined that insulin data can improve CGM prediction only to a certain extent. Indeed, because insulin's effect is delayed, this hormone is not very suitable for immediate prediction (within 60 minutes), but can significantly assist long-term prediction.

In 2018, the release of the Ohio T1DM (Type 1 Diabetes Mellitus) data set gave the field of CGM prediction a unified test standard. John Martinsson used LSTM neural network and negative log-likelihood loss as well as Physiological Loss Function on the 2018 Ohio T1DM data set, and got the errors of 18.86 mg/dL and 31.4 mg/dL in the time window of 30 and 60 minutes, respectively.

Stan Kriventsov was able to achieve better results in a retrospective study conducted in 2020. Kriventsov took a hybrid approach, incorporating machine learning alongside a biophysical model. Kriventsov used Gradient Boosted Decision Trees and SVM to achieve 86.7% and 70.6% accuracy in the 30 and 60-minute prediction intervals, respectively. This model was also tested on the Ohio dataset, and achieved 18.68 mg/dL error in the 30-minute prediction window.

Kriventsov had taken a different approach than Ignacio Rodriguez. Rodriguez aimed to design automatic BGL prediction models that incorporated *only* CGM data, which came from the real-time data of the FGM sensor. The author tried three models (namely, Autoregressive Integrated Moving Average, Random Forest and Support Vector Machine); Random Forest achieved the best results. Additionally, the author found that shorter sampling intervals of CGM data resulted in higher prediction accuracy. In a test with a sampling interval of 15 minutes, 15.43mg/dL and 25.9mg/dL were obtained on the predictions of 30 and 60 minutes, respectively; while the two figures obtained

at the sampling interval of 5 minutes were 14.63 mg/dL and 22.12 mg/dL at 30 and 60 minutes, respectively.

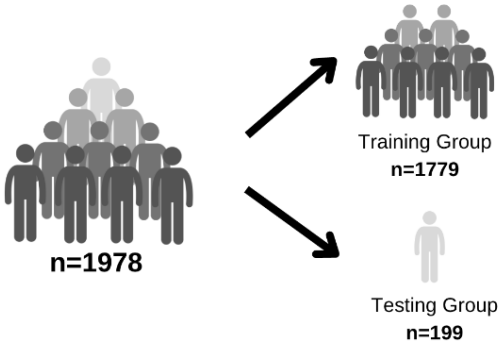
Our approach improves upon the previous body of work, with CGM prediction of up to 2 hours into the future (compared to 30 and 60 minutes), and increased accuracy as demonstrated by a superior RMSE score. Furthermore, our work allows us to not only predict BGL up to 2 hours in the future, but also to learn the user's biology sufficiently well as to not need a physical CGM device, a standard that has not been demonstrated elsewhere.

Results

For this study, we recruited 1978 users of the January V1 mobile application. These users were a subset of our overall user base, and who met a simple set of criteria: 3 "good days" of logging, where a "good day" is defined as at least 1 hour of CGM coverage per day (here, "coverage" refers to CGM's ability to read and display a user's BGL value); and not logging more than 10k calories per day.

This group of 1978 users was split into two groups: a training set (1779 people), and a testing set (199 people), as reflected in Fig. 1, below.

Fig. 1

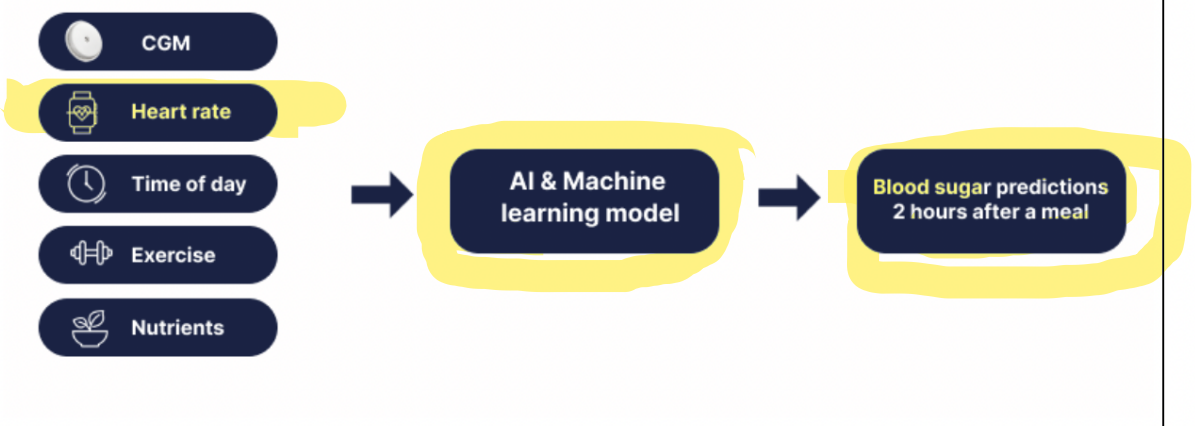


Study Design. 1978 users were split into a training group (n=1779) and a testing group (n=199).

Inputs

We collected a number of inputs for all users. First, we collected user's CGM data, which came from CGM devices worn by users during the 14-day AI training period; this data was reflective of our target, as we attempted to predict CGM whether into the future (CGP), or when the user was not not wearing a physical CGM device (VCGM) (Fig. 2 and Fig. 3)

Fig. 2



The inputs to the CGP model (where BGL is predicted up to 2 hours after a meal) are CGM, HR, time of day, exercise, and nutrients.

Fig. 3

The inputs to the VCGM model (where BGL is predicted without the use of a physical CGM device) are HR, time of day, exercise, and nutrients.

We also collected HR data, generated by HRM devices worn by users. HR information has two implications for BGL and the resultant predictions. First, increased HR due to activity leads to a depression in BGL. Conversely, increased HR due to stress leads to an *increase* in BGL. As a corollary to collecting HR data, we also collected exercise information logged by the user, including the type and duration of said exercise.

An additional input was all foods logged by users, along with the corresponding nutritional information (which was populated by our application). Food logging is critical to CGM predictions, as foods (and especially carbohydrates) are directly correlated with BGL.

We collected time-of-day information, which was passed to the model as both sine and cosine waves (with the period of each wave being 24 hours). While a user's food consumption influences their BGL level, so too does their circadian rhythm. For example, most users see a modest spike in BGL immediately before waking, even after fasting while asleep. Thus, collection of time-of-day information allowed us to learn across users around what time BGL rises, or is steady.

Finally, we collected "one hot" vectors for CGM, activity, meals, and HR. For our study, each "Time Step" (TS) spanned fifteen minutes. For all other values other than activity, each "one hot" vector was binary; for activity, we measured intensity and thus measured four values ("activity mild"; "activity moderate"; "activity intense"; and "no activity"). One-hot vectors were utilized in order to let us tell the model whether the features described above were available, or not. Further discussion of "one hot" vectors can be found below.

Data Summary

We collected 46,655 days' worth of data from all users. This equates to 4,478,880 TS (there are 96 15-minute TS in a 24-hour interval). Of this total, for all users, we collected 4,039,528 TS of CGM coverage; 3,231,375 TS of HRM data; and 43,632 activities were logged.

We collected 41,907 days' worth of data from our training users (4,023,072 TS). Our training users logged 96,589 meals and logged 39,083 activities. Furthermore, our training users had CGM coverage of 3,627,766 TS; and HRM coverage of 2,898,004.

We collected 4,748 days' worth of data from our testing users (455,808 TS). Our testing users logged 10,872 meals and 4,549 activities. Furthermore, our testing users had CGM coverage of 411,762 TS; and HRM coverage of 333,371 TS.

Metrics

We calculated a number of metrics, which reflect the accuracy of our model for CGP and/or VCGM.

We first collected RMSE Peak data, which represents the difference between the *peaks* of the predicted and actual values for BGL. Because of this, RMSE peak only captures one data point in each 2hr 15m prediction window. RMSE Peak is important for counterfactuals, which tell the user about actual, projected, or avoided spikes (by means of food substitution or exercise).

We also collected RMSE point-by-point data. This represents the difference between the predicted and actual values for BGL, at 10 points separated by 15-minute intervals across the 2hr 15m prediction window. While RMSE Peak represents the *height* of a user's BGL curve, RMSE point-by-point represents the *shape* of the user's curve, and is similarly important for counterfactual exercises.

RMSE point-by-point shifted data was also collected. Based on where the actual maximum peak happened in a user's CGM, we shift the CGP predictions retroactively so that the peak lies directly beneath the CGM peak (i.e., at the same time). Here, we calculate the RMSE of *overlapping* points (because, in shifting, we might lose the furthermost points).

We also collected data for the Pearson Correlation Coefficient, percent error, and root means square error (RMSE). Each of these metrics reflects the accuracy of our predictions, and equations for each are given below.

Pearson Correlation Coefficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

$$\delta = \left|\frac{v_A - v_E}{v_E}\right| . 100\%$$

δ = percent error
v_A = actual value observed
V_E = expected value

$$RMSE = \sqrt{\frac{\sum_{i=i}^{N} (predicted_i - Actual_i)^2}{N}}$$

Delay

When a user scans their CGM device, they collect not only an instantaneous BGL measurement; but also all of the measurements from the 8 hours that preceded their scan (for example, if the user scans their device at 8PM, they collect all of the measurements from 12PM-8PM).

Our model requires an input at the time a prediction is requested; however, per the above logic, it is that users will scan their CGM devices in such a way that they have gaps in their CGM coverage. Since our model was trained on data that was passed months in the past, we realized that gaps in CGM coverage which might have been reflected at the original time, might not be reflected at the time when the model was being trained. For example, if Person A scans every hour, and Person B scans every 6 hours, our data set will reflect full coverage for both people (because those scans happened months ago); in reality, however, Person B did *not* have full coverage at the time he requested a prediction.

Thus, in training, our model would have accustomed itself to account for coverage that was artificially without gaps; that is, our model would *expect* CGM values 15 minutes before a prediction is requested. As explained above, however, this level of coverage might not reflect reality; due to infrequent scanning, we cannot guarantee that there will be coverage at the time a user wants to see a prediction. Therefore, in order to train the model and account for a lack of coverage, we introduced a delay. This delay was a Gaussian distribution, with a mean of 2 hours' delay, and a standard deviation of 8 hours (with a maximum of 0 hours' delay). As expected, percent error was lower without delay (Table 1).

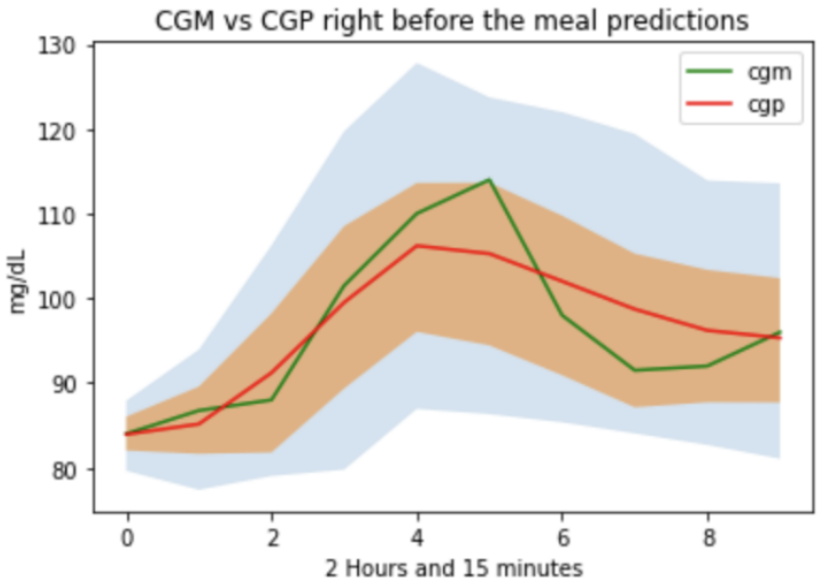
Table 1.

	CGP2.0 without delay	CGP2.0 with delay
RMSE PEAK	20.5 mg/dL	22.6 mg/dL
RMSE point by point	13.2 mg/dL	14.3 mg/dL
RMSE point by point shifted	12.8 mg/dL	13.6 mg/dL
Correlation	0.71	0.71
Percent Error	10.3%	11.1%

Eliminating delay from the CGP2.0 model leads to lower RMSE Peak, RMSE point-by-point, RMSE point-by-point shifted, and lower percent error.

CGP

Fig. 4



Our CGP algorithm operates on a stochastic basis, generating at each TS 100 different potential BGL values, along with the corresponding likelihood of each value occurring. The CGP value displayed is the weighted average of those 100 different values. In the figure above, the red line reflects CGP; the green line represents CGM; the orange zone is the range between the 25th and 75th percentiles of the potential values that are generated by the CGP algorithm; and the blue zone is the range between the 25th and 75th percentiles of the potential values that are generated by the CGP algorithm. This figure shows that the actual curve falls within the confidence interval of the predicted curve.

Virtual CGM

Fig. 5

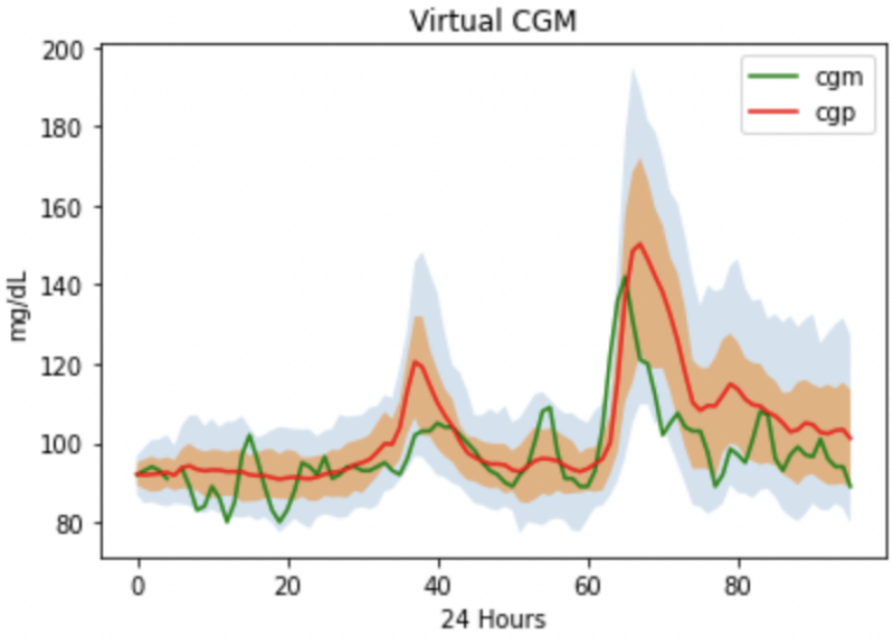


Figure 5 reflects a full day's BGL without any actual CGM values available. This figure demonstrates that users' BGL falls within the confidence interval of our BGL predictions (VCGM).

While CGP always predicts BGL values 2 hours and 15 minutes into the future, VCGM's predictions are live. Furthermore, unlike CGP, VCGM values are displayed while the user is *not* wearing a CGM device. Stated differently, VCGM doesn't use any CGM inputs, while CGP uses CGM inputs.

Thus, out of the metrics investigated for CGP, there are fewer that apply to VCGM. Those metrics are reflected in Table 2.

Table 2

	VCGM
RMSE point by point	17.8 mg/dL
Correlation	0.83
Percent Error	13.0%

The most applicable metrics for VCGM are RMSE point by point, correlation, and percent error.

Benchmarks

Comparison with baselines/benchmarks allows us to determine the efficacy of our prediction algorithm. Overall, our metrics are more reflective of actual values than other comparable algorithms. January's CGP saw a percent error of only 10.3% (the next closest benchmark was 15.1%). January's VCGM saw a percent error of 13.0% (the next closest benchmark was 14.2%). (Table 3).

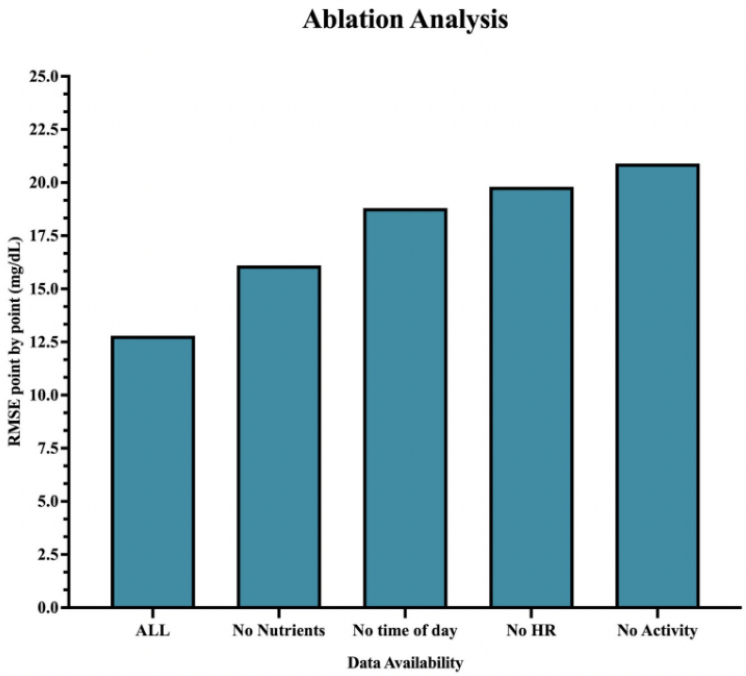
Table 3: Comparison with Benchmarks

	Predict average for all users		Predict average for each user		XGBoost		Random Forest		January AI	
	CGP	**VCGM**	**CGP**	**VCGM**	**CGP**	**VCGM**	**CGP**	**VCGM**	**CGP**	**VCGM**
RMSE Peak	44.5 mg/dL	-	35.2 mg/dL	-	24.0 mg/dL	-	27.1 mg/dL	-	**20.5 mg/dL**	-
RMSE Point by Point	20.3 mg/dL	32.1 mg/dL	19.3 mg/dL	22.5 mg/dL	19.5 mg/dL	18.7 mg/dL	21.5 mg/dL	19.0 mg/dL	**13.2 mg/dL**	**17.8 mg/dL**
RMSE Point by Point Shifted	21.8 mg/dL	-	19.7 mg/dL	-	18.2 mg/dL	-	20.2 mg/dL	-	**12.8 mg/dL**	-
Correlation	-	-	-	-	0.58	0.30	0.54	0.29	**0.71**	**0.83**
Percent Error	18.2%	18.5%	15.1%	15.0%	15.7%	14.2%	15.9%	14.9%	**10.3%**	**13.0%**

Ablation Analysis

We investigate our most important feature in terms of CGP, using ablation analysis to determine which input to the model has the most impact on our ability to predict BGL (Fig. 6).

Fig. 6

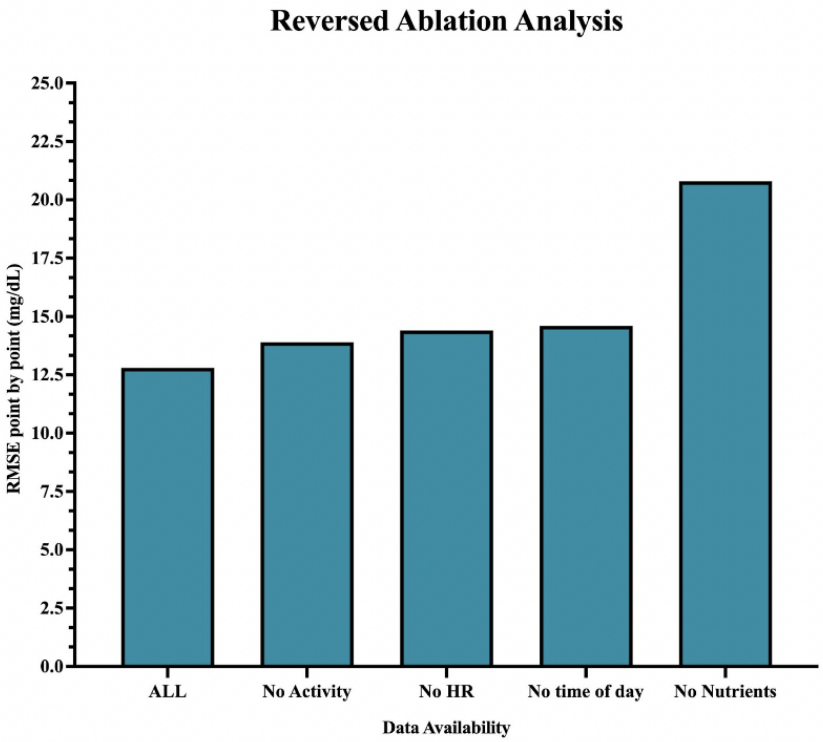


Ablation analysis was conducted to determine the most important input to the fidelity of the model. In order of importance, the most important inputs are nutrients; time of day; HR; and activity.

Reverse Ablation analysis

While ablation analysis *removes* inputs to determine which input is most important, reverse ablation analysis *adds* inputs. Fig. 7, below, demonstrates the results of our reverse ablation analysis in terms of various inputs' effects on our model's fidelity.

Fig. 7



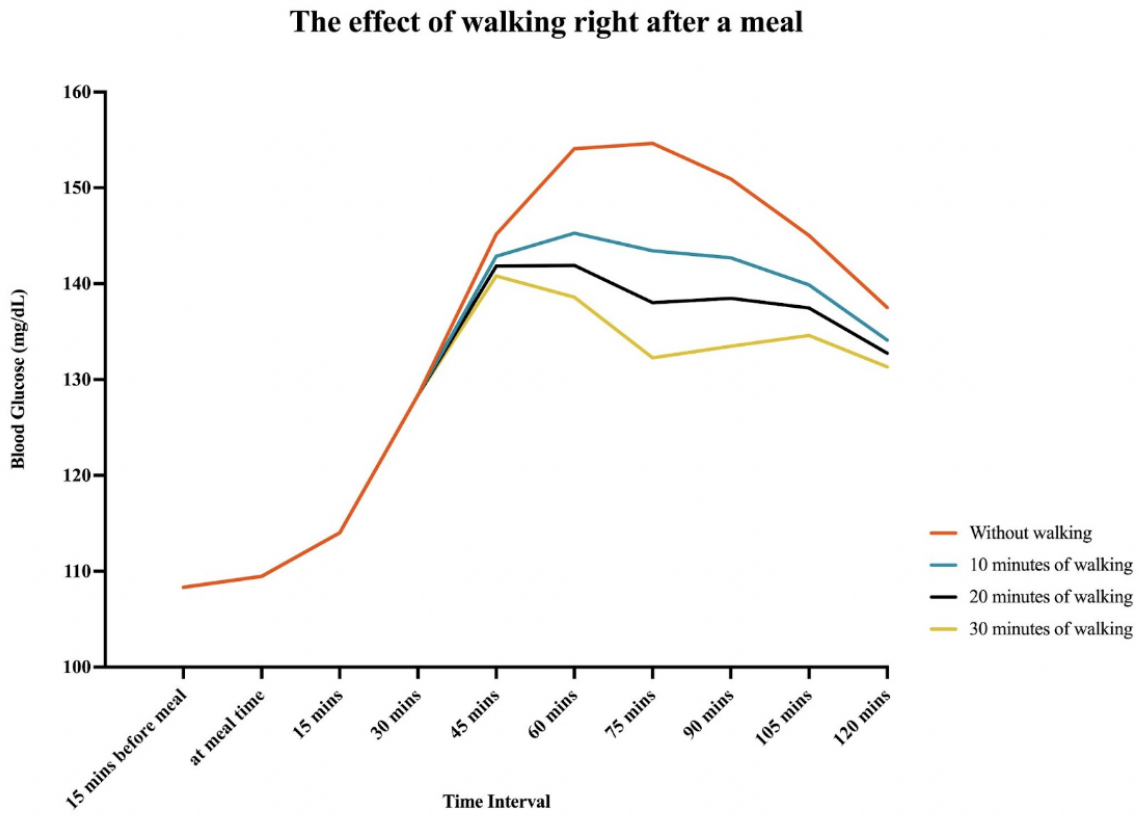
Reverse ablation analysis was conducted to determine the most important input to the fidelity of the model. In order of importance, the most important inputs are nutrients; time of day; HR; and activity.

Exercise Forcing

Because exercise was logged so infrequently by our users within 2 hours of mealtime, our model discounted the impact of exercise on BGL when compared to other levers like nutrients.

In the example below (Fig. 8), we hypothetically added activity to a user's CGP by extrapolating the effect that activity had on the BGL levels of patients who *had* logged activity. In doing so, we were able to demonstrate the effect of activity on BGL levels and thus CGP.

Fig. 8



Counterfactuals are used to demonstrate the effect of activity (in this case, a post-meal walk) on a user's BGL level. As shown, even a walk of modest length can lower BGL spikes, and, the longer the walk, the greater the attenuation of the spike.

Discussion

Overall, we were able to demonstrate that our machine learning-based algorithm for both continuous glucose prediction (CGP) and Virtual CGM (VCGM) demonstrated a high degree of fidelity, with performance superior to comparable benchmarks across various metrics. By consolidating food logging and HRM data, along with CGM data during the training period, we produced better predictions than baseline models; CGP was ~31% better, and VCGM was ~8.5% better than baseline.

It should be noted that, though our V1 algorithm demonstrated superior performance, even this algorithm is less effective than our current, V2 algorithm. This is because the methods reflected in this paper imposed far less stringent requirements upon users for what constituted a “good day”, *vis a vis* the number of daily logged meals, and the hours of CGM/HRM usage during the training period.

For our V2 product, however, we required users to collect at least 5 days of: 1) 12 hours of HRM coverage; 2) 12 hours of CGM coverage; 3) at least 3 meals logged.

Thus, given the increased amount of data as a result of this product decision, we expected the model to be more accurate than is reflected in this paper; indeed, that was the case. Results for users who had 4 or more “good days”, as well as users who had 7 or more “good days”, are reflected in Table 6.

Table 6:

	4+ “Good Days”	**7+ “Good Days”**	**V1 Testing Group Users**
RMSE point by point	12.6 mg/dL	11.9 mg/dL	17.8 mg/dL
Percent Error	9.8%	9.3%	13%

The marked improvement in the accuracy of our CGP and VCGM can be accounted for by better user logging; indeed, the accuracy of these models was judged solely on the basis of users’ good days, where those good days are much more stringently defined.

Further discussion of these results is beyond the scope of this paper.

---

1. Freckmann, Guido. "Basics and use of continuous glucose monitoring (CGM) in diabetes therapy" *Journal of Laboratory Medicine*, vol. 44, no. 2, 2020, pp. 71-79. https://doi.org/10.1515/labmed-2019-0189.

2. Klonoff DC, Nguyen KT, Xu NY, Gutierrez A, Espinoza JC, Vidmar AP. Use of Continuous Glucose Monitors by People Without Diabetes: An Idea Whose Time Has Come? Journal of Diabetes Science and Technology. 2022;0(0). doi:10.1177/19322968221110830.

3. Müller M, Canfora EE, Blaak EE. Gastrointestinal Transit Time, Glucose Homeostasis and Metabolic Health: Modulation by Dietary Fibers. *Nutrients*. 2018; 10(3):275. https://doi.org/10.3390/nu10030275.

4. John I. Malone; Diabetic Central Neuropathy: CNS Damage Related to Hyperglycemia. *Diabetes* 1 February 2016; 65 (2): 355–357. https://doi.org/10.2337/dbi15-0034.

5. Mendez, C.E., Der Mesropian, P.J., Mathew, R.O. *et al.* Hyperglycemia and Acute Kidney Injury During the Perioperative Period. *Curr Diab Rep* 16, 10 (2016). https://doi.org/10.1007/s11892-015-0701-7.

6. Steven M. Haffner, The Importance of Hyperglycemia in the Nonfasting State to the Development of Cardiovascular Disease, *Endocrine Reviews*, Volume 19, Issue 5, 1 October 1998, Pages 583–592, https://doi.org/10.1210/edrv.19.5.0343.

7. Kalra S, Mukherjee JJ, Venkataraman S, Bantwal G, Shaikh S, Saboo B, Das AK, Ramachandran A. Hypoglycemia: The neglected complication. Indian J Endocrinol Metab. 2013 Sep;17(5):819-34. doi: 10.4103/2230-8210.117219. PMID: 24083163; PMCID: PMC3784865.

8. Engler, S., Fields, S., Leach, W. *et al.* Real-Time Continuous Glucose Monitoring as a Behavioral Intervention Tool for T2D: A Systematic Review. *J. technol. behav. sci.* 7, 252–263 (2022). https://doi.org/10.1007/s41347-022-00247-5.

9. Predictive metabolic modeling for type 1 diabetes using free-living data on mobile devices. E.I. Georga, V.C. Protopappas and D.I. Fotiadis, J.C. Lin, K.S. Nikita (Eds.), Wireless Mobile Communication and Healthcare: Second International ICST Conference, MobiHealth 2010, Ayia Napa, Cyprus, October 18-20, 2010. Revised Selected Papers (2011), pp. 187-193.

10. The Diabits App for Smartphone-Assisted Predictive Monitoring of Glycemia in Patients With Diabetes: Retrospective Observational Study", Stan Kriventsov, JMIR 2020.

11. "Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks", John Martinsson, JHIR 2019.

12. How Much Is Short-Term Glucose Prediction in Type 1 Diabetes Improved by Adding Insulin Delivery and Meal Content Information to CGM Data? A Proof-of-Concept Study", Chiara Zecchin, JDST 2016.

13. "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes", Georga, E.I., 2012.

14. "A predictive model of subcutaneous glucose concentration in type 1 diabetes based on random forests", Goerga, E.I., 2012.

15. "How Much Is Short-Term Glucose Prediction in Type 1 Diabetes Improved by Adding Insulin Delivery and Meal Content Information to CGM Data? A Proof-of-Concept Study", Chiara Zecchin, JDST 2016.

16. Favero SD, Facchinetti A, Cobelli C. A glucose-specific metric to assess predictors and identify models. 2012;59(5):1281–1290.

17. "Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks", John Martinsson, JHIR 2019.

18. "The Diabits App for Smartphone-Assisted Predictive Monitoring of Glycemia in Patients With Diabetes: Retrospective Observational Study", Stan Kriventsov, JMIR 2020.

19. "Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through Machine Learning Techniques", Ignacio Rodriguez, 2019