# Personalized Impact of Lifestyle on Type 1 Diabetes Patients: A Comprehensive Regression Analysis

Sinan Xie
*Department of Electrical & Computer Engineering*
*University of California, San Diego*
San Diego, USA
six004@ucsd.edu

Jared Leitner
*Department of Electrical & Computer Engineering*
*University of California, San Diego*
San Diego, USA
jjleitne@ucsd.edu

Sujit Dey
*Department of Electrical & Computer Engineering*
*University of California, San Diego*
San Diego, USA
sdey@ucsd.edu

***Abstract*—While lifestyle behaviors play important roles in Type 1 Diabetes (T1D), the individualized effects of lifestyle factors on T1D patients have not been studied. In this paper, we present a regression analysis-based approach to understand the personalized impact of lifestyle factors on blood glucose (BG) in T1D patients using the OhioT1DM dataset. Our method addresses the following challenges: (1) effectively structuring the multi-modal lifestyle, insulin, and BG data and (2) modeling the data to derive personalized insights. To solve the first challenge, a patient's data are segmented into time windows based on BG-affecting events. Consequently, we can better reveal each lifestyle factor's effect because windows capture BG trends at higher granularity and contain fewer concurrent BG-affecting factors. To model the data and derive personalized insights, we first utilized variance inflation factor (VIF) and correlation analysis to eliminate and combine lifestyle features to avoid multicollinearity issues. We then trained multiple machine learning models and found multivariate linear regression (MLR) yielded the best BG prediction. We derived personalized insights about the relationship between lifestyle factors and BG using the MLR model's statistical properties, including $\beta$ coefficients, P-values, and $R^2$ values. Our results show that T1D patients differ significantly in how lifestyle factors affect their BG, indicating that personalized lifestyle interventions are necessary for T1D management.**

***Index Terms*—type 1 diabetes, regression analysis, smart healthcare, digital health, lifestyle medicine**

## I. Introduction

Type 1 Diabetes (T1D) is a non-reversible, autoimmune disorder characterized by the destruction of $\beta$-cells in the pancreas, leading to a lack of insulin production. T1D management focuses on glycemic control, keeping blood glucose (BG) within a normal range to delay or prevent complications [1]. Effective glycemic control [2] involves exogenous insulin administration, continuous glucose monitoring (CGM), and following lifestyle guidelines regarding meal planning and eating patterns [3], [4], physical activity [5], [6], sleep [7], [8] and stress management [9].

Personalized lifestyle intervention is a growing trend for managing chronic diseases. By developing customized rather than one-size-fits-all lifestyle strategies, individuals result in better health outcomes [10]. For T1D, while previous literature has investigated how lifestyle factors affect the T1D population, the interpersonal differences in the impact of these factors among T1D patients have not been studied.

Due to the widespread use of mobile applications and wearables, we can track and generate ample lifestyle data. Combining lifestyle data with BG data collected from CGM presents an opportunity for a data-driven approach to analyze T1D patients' individual responses to lifestyle factors. In this study, we aim to analyze the personalized effect of lifestyle factors on T1D patients' BG using CGM, insulin, and lifestyle data to derive customized insights for better glycemic control. Considering the complexity of T1D management and BG dynamics, two challenges in developing such a system are identified:

1) ***Complexity of T1D Data***: Multi-modal lifestyle, insulin, and CGM data sampled at different frequencies complicate the data structuring, making it challenging to isolate and reveal each factor's impact on BG.

2) ***Personalized Lifestyle Insights***: Lifestyle factors are bi-directionally correlated with each other in T1D patients [11], posing difficulty in identifying the true effects of different factors on BG to generate personalized lifestyle insights.

To address these challenges, we propose a novel regression analysis-based system with specifically designed data segmentation and feature engineering. When structuring the multi-frequency T1D data, we segment and aggregate CGM and lifestyle data into temporal windows delimited by pre-defined events. These events indicate trends in BG change induced by new lifestyle factors, such as when individuals fall asleep, consume meals, inject bolus insulin, etc. We can thus capture BG dynamics at higher granularity and reduce the number of concurrent factors within a given window, making it easier to isolate each factor's impact on BG. As a result, each window corresponds to a dataset sample, consisting of aggregated lifestyle and insulin data as features and the change in BG within the window as the label.

Once we processed the data, we used feature engineering techniques and regression analysis to model the relationships between lifestyle features and BG to gain personalized insights into T1D patients' data. To minimize the correlations between lifestyle and insulin features, we employed correlation analysis and variance inflation factor (VIF) to guide feature

engineering. This results in a feature set that provides better interpretability by avoiding multicollinearity issues (i.e., several independent variables are highly correlated). We compared multiple machine learning and statistical regression models using the OhioT1DM dataset and found that multi-variate linear regression (MLR) achieved the best prediction performance. Furthermore, MLR provides a higher degree of interpretability by demonstrating both the magnitude and statistical significance of the associations between variables. Our regression analysis results demonstrate that interpersonal variance exists in how lifestyle factors affect BG in T1D patients, highlighting the need for personalized T1D lifestyle intervention. Moreover, we discuss personalized lifestyle insights drawn from the regression analysis and how to support the development of a customized recommendation system. To the best of our knowledge, this is the first work that investigates the personalized effect of lifestyle factors on BG for T1D patients.

## II. Methodology

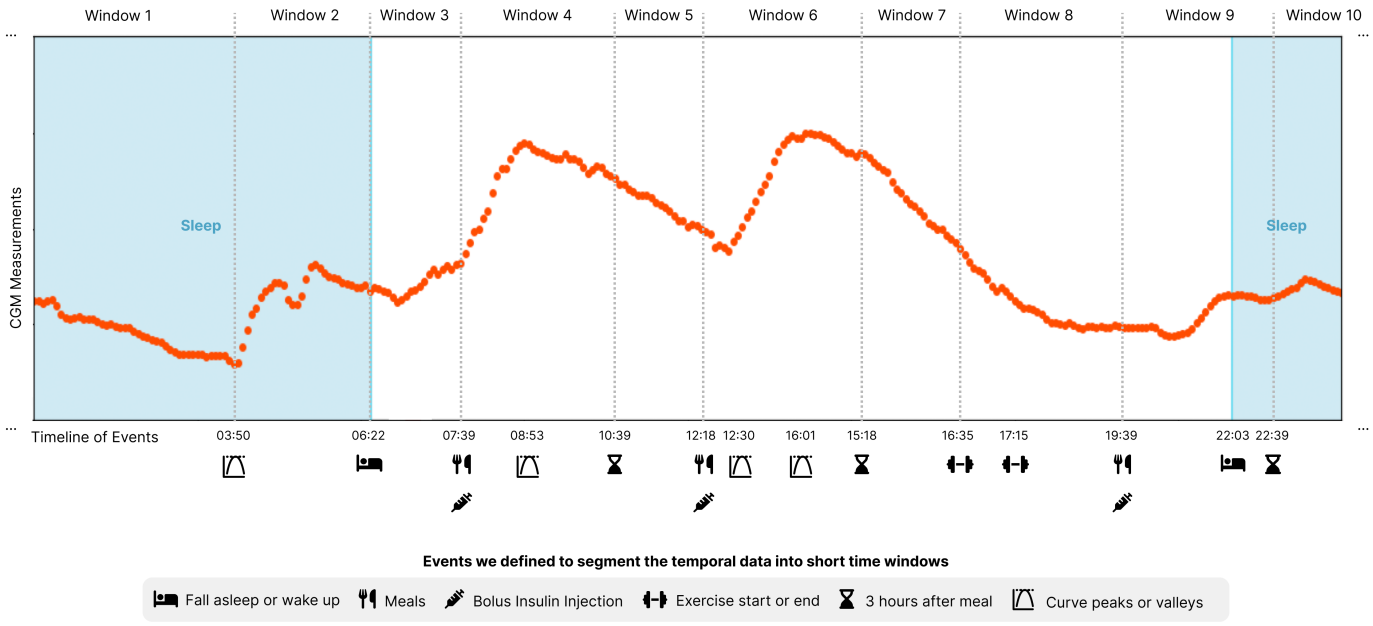
### A. Dataset Overview

We evaluated our approach using the publicly available OhioT1DM dataset [12]. The dataset includes eight weeks of CGM, insulin, physiological, and movement data, as well as self-reported lifestyle events collected from 12 T1D patients aged 20 to 60. All patients used basal-bolus therapy, i.e., taking multiple doses of long-acting insulin (basal) combined with rapid-acting insulin (bolus) to stabilize daily BG levels. Lifestyle events were reported through a custom smartphone app, while physiological and movement data were tracked using fitness bands. After eliminating patients with insufficient lifestyle data (e.g., frequently missing sleep or meal data), 7 patients were analyzed for this study.

The OhioT1DM dataset's wealth of lifestyle data makes it ideal for analyzing the personalized effects of various lifestyle factors on the BG of T1D patients. Listed below are data types from the OhioT1DM dataset we include in our study:

**1. BG Levels**: CGM measurements recorded every 5 minutes.
**2. Meals**: meal times and estimated carbohydrate intake in grams.
**3. Sleep**: sleep start-end times and self-rated qualities (1 to 3).
**4. Exercise**: exercise start-end times and self-rated intensities (1 to 10).
**5. Work**: work start-end times and self-rated intensities (1 to 10).
**6. Bolus**: bolus insulin injection times and dosages.
**7. Basal**: basal insulin start-end times and the rates of continuous injection (basal rate).
**8. Steps**: step counts recorded by a fitness tracker, aggregated every 5 minutes.

These data constitute the main BG-affecting factors: diet, sleep, physical activity, and exogenous insulin. Note that work, either sedentary or physical, could affect both physical activeness and psychological stress and lead to changes in BG.




**Note:** Not all events lead to the start/end of windows because of time constraints related to meals and minimum window length. For instance, at 08:53, 12:30, 16:01 and 22:03, the window does not end due to a former undigested meal, while at 17:15, the window does not end as the window length would be less than 1 hour.

Fig. 1. Illustration of our data segmentation on a sample day. On the timeline, we display the types and times of pre-defined events that indicate window segmentation. The start and end times of each window are illustrated by dashed lines.

Due to this, we include it in the scope of our study.

### B. Data Segmentation

Temporal segmentation of CGM, insulin, and lifestyle data into time windows should be carefully designed in order to accurately reflect the impact of each lifestyle factor on an individual's BG levels. Short windows are vulnerable to picking up transient fluctuations in BG and CGM sensor noise. When using long windows, such as 12 or 24 hours, the change in BG of each window would be attributed to the combined effect of all BG-affecting factors over this period. These factors may cancel out each other's effect, making it challenging to isolate the impact of a single factor on BG.

Another crucial factor to consider when segmenting data is the carbohydrate absorption rate. To accurately represent a meal's effect on BG over a period of time, it's best to consider the amount of absorbed carbohydrates, rather than the total amount consumed. However, calculating absorbed carbohydrates is difficult as the time for carbohydrates to fully break down into glucose and enter the bloodstream can vary from 0.5 to over 3 hours based on the type of carbohydrates, glycemic index, and its combination with protein and fat [13]. Keeping in mind these concerns, we propose a set of segmentation events where each window should begin or end at:

- *BG-affecting Events*: meals, bolus insulin injections, sleep times, wakeup times, and exercise start times.
- *BG Peaks and Valleys*: local maxima or minimal of BG measurements.
- *Carbohydrate Absorption Completed*: a fixed time $t_{absorb}$ after each meal.

Segmenting data at BG-affecting events reduces the number of BG-affecting events that take place simultaneously in each window while segmenting at BG peaks and valleys allows BG dynamics to be captured in greater granularity, both of which serve to isolate and reveal the impact of lifestyle factors on BG. To deal with the varied carbohydrate absorption rate, we constrain each window to start $t_{absorb}$ after any previous meals and end $t_{absorb}$ after any meals within the window. By doing so, we can assume that the absorbed carbohydrates are equivalent to the consumed carbohydrates. Each window has a minimum length of $t_{min}$ to avoid picking up noise. In this study, we set $t_{absorb}$ to 3 hours and $t_{min}$ to 1 hour. Windows that are missing CGM readings within 15 minutes of the start and end time are excluded. Figure 1 shows a sample of data segmentation on 24 hours of data using our method. A total of 1789 windows were generated for 7 patients, with each patient having more than 200 windows.

### C. Feature Engineering

Lifestyle factors have both long-term (up to 24 to 48 hours) and short-term effects on T1D BG levels [11], [14]. For example, being physically active and getting adequate and high-quality sleep on the former night may reduce stress levels and increase insulin sensitivity on the next day, leading to better glycemic control. This has motivated us to consider different time periods when generating lifestyle features. As a result, we named two kinds of features for each window - *window* and *24hrs* features. *window* features are aggregated within each window to capture the short-term effects of lifestyle on BG. *24hrs* features are aggregated 24 hours before the window start time to capture the feature's long-lasting effect on BG change. A complete list of the generated features (original features) is shown in the left column of Table I. It's worth noting that when calculating the bolus insulin features, we use ingested rather than consumed bolus insulin amounts similar to carbohydrates. We utilized an exponential curve implemented by the OpenAPS API to calculate the insulin's cumulative activity over time [15]. For the target variable, we use BG difference which is calculated by subtracting the starting BG from the ending BG of each window since we are interested in the BG change instead of the absolute value of BG.

Regression analysis is a statistical technique widely used to investigate the association between two or more variables. In this study, we selected multi-variate linear regression (MLR) as the model, as described in subsection III-A. One issue with using MLR is multicollinearity, which occurs when a regression model's predictors are highly correlated. Multicollinearity leads to low accuracy of the estimated coefficients and a loss in the model's statistical power [16].

TABLE I
The originally generated features and the final features derived to avoid the multicollinearity issue in regression analysis. Features ending with _window are aggregated within each time window, while those ending with _24hrs are aggregated 24 hours before the window's start time.

| **Original Features** | **Final Features** |
|---|---|
| carb_window, sleep_duration_window, sleep_duration_24hrs, sleep_quality_window, sleep_quality_24hrs, exercise_duration_window, exercise_duration_24hrs, exercise_intensity_window, exercise_intensity_24hrs, work_duration_window, work_duration_24hrs, work_intensity_window, work_intensity_24hrs, bolus_window, basal_window, steps_window, steps_24hrs | carb_window, sleep_score_window, sleep_score_24hrs, exercise_load_window, exercise_load_24hrs, work_load_window, work_load_24hrs, bolus_window, basal_window, steps_window, steps_24hrs |

A common method to detect multicollinearity is variance inflation factor (VIF), which is calculated as follows:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where $R_i^2$ is the $R^2$ value of a linear regression model that regresses the $i^{th}$ predictor using all the other predictors. In practice, a VIF value less than 4 indicates a weak correlation between one predictor and the other predictors [17]. The original features in Table I yield high VIF values, suggesting high correlations between features. To identify which features are most correlated and guide the elimination of certain features, we generate a feature correlation heatmap for each patient. Figure 2 shows a heatmap of the correlation matrix for one patient, with darker shades of red or blue representing stronger positive or negative correlations, respectively, between the two features. As highlighted by yellow circles on the heatmap, there exist strong positive correlations between sleep duration and quality, exercise duration and intensity, and work duration and intensity, either short-term or long-term. To eliminate high correlations within the data while preserving a broad range of lifestyle features, we replace the original sleep, exercise, and work features with "sleep scores", "exercise loads" and "workloads" by multiplying the duration features with the intensity or quality features. For example, "sleep score" is calculated by multiplying sleep duration by the reported sleep quality. By doing so, the VIF value of all predictors falls below 4 for all patients. The final features for regression analysis are summarized in the right column of Table I.

### *D. Regression Analysis for Personalized Lifestyle Insights*

Personalized MLR models are developed for each participant using the final features generated in the previous section as predictors and the BG change within each window as the target variable. Then, we utilize the $R^2$ value for each individual MLR model, the $\beta$ coefficient, and its statistical significance to derive personalized lifestyle insights. Note that standardizing the features on a per-patient basis allows for the comparison of their association with BG change using coefficient values.

We start by gaining an overall understanding of how a T1D patient's BG changes are affected by lifestyle factors using the $R^2$ value and the number of significant predictors for each patient model. In a regression model, the $R^2$ value captures the extent of variance in the target variable that is explained by the predictors. Therefore, a high $R^2$ indicates that the T1D patient's BG changes could be mostly explained by lifestyle features, suggesting lifestyle intervention would be pivotal in managing the patient's BG levels. On the other hand, a low $R^2$ value implies that the lifestyle features are insufficient to explain the BG changes of a patient. This could mean that other factors besides lifestyle should be considered in the patient's BG management. Moreover, the number of

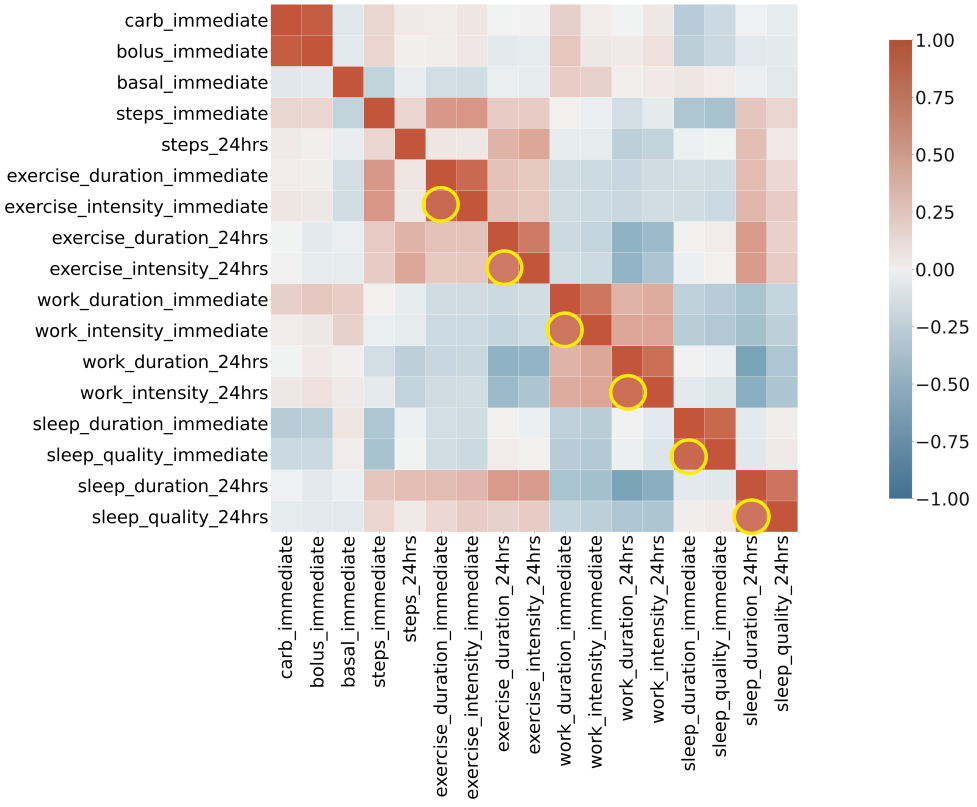



Fig. 2. The correlation heatmap of one patient. Notable correlations are circled in yellow.

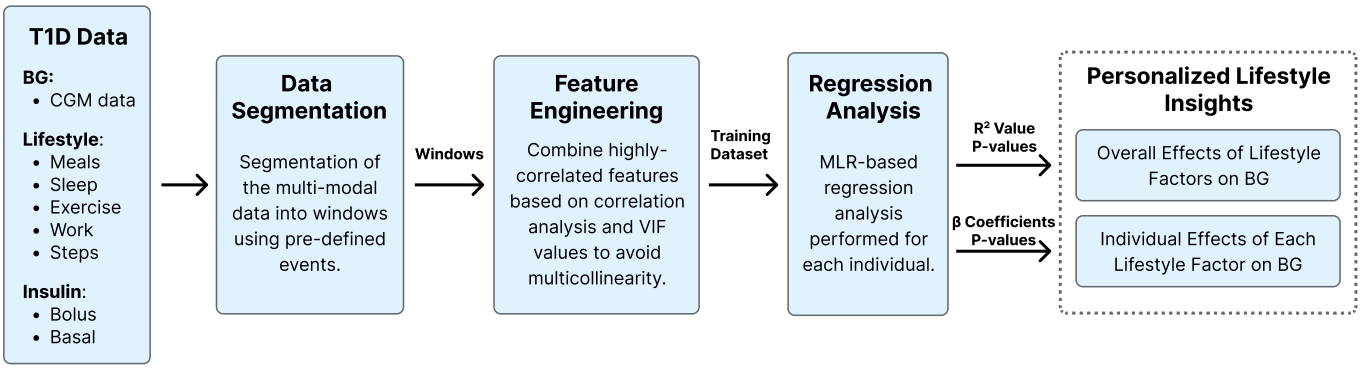



Fig. 3. Block Diagram of our proposed method of deriving personalized lifestyle insights.

significant predictors shows whether multiple or only a few lifestyle features are potentially influencing the patient's BG levels, which is also valuable in forming a customized lifestyle intervention strategy.

After analyzing the overall effects of lifestyle factors on BG, we take a deeper look at each predictor's $\beta$ coefficients and the significance levels across patients to investigate whether their effects on BG vary interpersonally. For each predictor, we pick all patients having significant coefficients and compare their directionality. If we observe variations in the directionality, i.e., both positive and negative associations with BG change occur in different patients, or when the directionality is opposite to what is expected, e.g., carbohydrate intake showing a negative association with BG, we perform additional inspection of the data to provide explanations. Finally, we investigate the top lifestyle predictor for each patient, defined by the statistically significant lifestyle predictor having the largest absolute coefficient value. This top factor could point to the lifestyle behavior having the greatest effect on BG and therefore inform personalized lifestyle management for T1D. Figure 3 presents a summary of our regression analysis-based method for deriving personalized lifestyle insights.

## III. Results

### A. BG Change Prediction

We train multiple statistical and ML regression models to determine which is most effective at modeling the lifestyle-BG data, including support vector machine (SVM), random forest (RF), gradient boosting (GB), and multivariate linear regression (MLR). A dummy model that always predicts no change in BG is used as a benchmark. All ML models are trained using the final features in Table I, while for MLR we trained two versions - MLR-original and MLR-final - using either the original or final features as predictors to investigate how our feature engineering approach to avoid multicollinearity would affect the MLR's prediction performance. We performed 5-fold cross-validation on each patient's data, randomly picking 80% of the data as the training set and 20% as the testing set. This process was repeated 5 times, with the experimental results averaged. Mean absolute error (MAE) and root mean squared error (RMSE) are calculated as performance metrics. All model implementation and evaluation are done using the Scikit-learn library in Python. Hyperparameter tuning of the ML models is performed over the 5-fold cross-validation.

Patient-level and overall prediction results for each model are summarized in Table II. On an individual patient level, SVM and GB each achieved the best performance for Patients

TABLE II
Comparison of different models' patient-level and overall prediction performance (each cell is formatted as MAE / RMSE).

| | Patient 1 | Patient 2 | Patient 3 | Patient 4 | Patient 5 | Patient 6 | Patient 7 | Mean |
|---|---|---|---|---|---|---|---|---|
| SVM | 47.6/58.8 | 64.9/82.7 | 50.2/63.9 | **49.5/63.5** | 50.1/68.2 | 48.0/58.4 | 36.7/44.6 | 49.6/62.9 |
| RF | 32.8/42.8 | 68.0/91.9 | 47.2/62.4 | 50.5/64.5 | 45.8/57.5 | 45.6/57.9 | 34.0/41.8 | 46.3/59.8 |
| GB | 35.5/44.3 | 79.5/108.6 | 51.2/64.1 | 55.2/71.2 | 47.2/61.5 | **40.7/50.4** | 38.8/46.6 | 49.7/63.8 |
| MLR-original | 31.7/**42.6** | 62.3/82.0 | **44.5/59.5** | 50.6/65.2 | 39.7/52.5 | 43.7/52.9 | **32.7/38.5** | 43.6/56.2 |
| MLR-final | **31.5**/43.3 | **61.5/79.7** | 45.0/60.5 | 50.2/64.8 | **38.3/51.0** | 43.1/51.7 | 33.4/39.3 | **43.3/55.8** |
| Dummy Model | 51.7/63.5 | 64.8/82.8 | 50.8/64.5 | 49.5/63.6 | 51.6/69.7 | 47.7/58.2 | 36.4/44.7 | 50.4/63.9 |

TABLE III
Regression analysis results of each patient. Each cell displays the $\beta$ coefficient showing the association of the predictor with that patient's BG change. The top lifestyle predictor besides carbohydrate intake for each patient is bolded.

| | Patient 1 ($R^2 = 0.61$) | Patient 2 ($R^2 = 0.24$) | Patient 3 ($R^2 = 0.19$) | Patient 4 ($R^2 = 0.34$) | Patient 5 ($R^2 = 0.39$) | Patient 6 ($R^2 = 0.29$) | Patient 7 ($R^2 = 0.28$) |
|---|---|---|---|---|---|---|---|
| carb_window | 112.2** | 47.6** | 44.9** | 57.9** | 71.1** | 42.2** | 37.8** |
| bolus_window | -68.6** | -43.0** | -39.8** | -62.1** | -49.4** | -48.6** | -29.8** |
| basal_window | 16.3** | -10.3 | -1.8 | -8.0** | 2.9 | -18.7** | 0.5 |
| sleep_score_window | **-21.6**** | 3.6 | **18.0**** | 4.1 | 10.2* | 7.1 | -0.6 |
| exercise_load_window | -6.0 | -3.0 | -5.0 | -0.2 | **-19.5**** | **-10.6*** | **-9.7*** |
| work_load_window | 8.3* | **-11.9*** | -1.5 | 7.3 | -10.8** | n/a | n/a |
| steps_window | n/a | -17.7 | -9.7* | 6.2 | 0.4 | -1.8 | n/a |
| sleep_score_24hrs | -1.7 | 3.7 | -0.8 | 3.6 | 3.1 | -0.3 | -2.5 |
| exercise_load_24hrs | 2.4 | 3.5 | 0.9 | 3.0 | -0.4 | -2.2 | 6.0 |
| work_load_24hrs | -6.9 | 7.3 | -2.0 | 3.5 | 6.9* | n/a | n/a |
| steps_24hrs | n/a | -6.2 | 0.7 | 5.6 | -1.9 | -0.6 | n/a |

* indicates significance at the 0.05 level.
** indicates significance at the 0.01 level.
n/a indicates that no lifestyle data is recorded for that patient.

4 and 6, respectively, while MLR-original or MLR-final yielded the best performance for all other patients. Out of the three ML models (SVM, RF, GB), RF achieved the best overall prediction performance with an MAE of 46.3 and RSME of 59.8 averaged across all patients. MLR-original achieved a mean MAE of 43.6 and RMSE of 56.2 across all patients. MLR-final resulted in the best overall performance, achieving an MAE of 43.3 and RMSE of 55.8. One possible reason for the ML models' worse performance is overfitting due to the small dataset. In this case, MLR is better suited for smaller datasets due to its simplicity and lower risk of overfitting. All methods outperformed the dummy model, indicating their success in capturing meaningful patterns within the data.

Moreover, MLR-final achieved better patient-level and overall performance than MLR-original, confirming that the feature engineering we performed to enhance the model's statistical power did not compromise the model's predictive power. These observations led us to use MLR-final in our regression analysis for determining personalized insights about lifestyle's impact on BG.

### *B. Regression Analysis of Individualized Lifestyle Impact*

The regression analysis results for each patient are presented in Table III. These results include the $R^2$ value, the $\beta$ coefficients, and their statistical significance under the 0.05 and 0.01 levels for each of the predictors.

Across all patient models, the $R^2$ values ranged from 0.19 to 0.61, averaging 0.33. The number of predictors significant under the 0.05 significance level ranged from 3 to 6, averaging 4.7. These observations suggest that T1D patients vary in terms of how much their BG changes are explained by lifestyle factors. For example, Patient 3's model resulted in an $R^2$ value of 0.19, indicating that this patient's BG change may not be well explained by the insulin and lifestyle factors included in the dataset. However, Patient 3's model still resulted in statistically significant coefficients for four predictors (*carb_window*, *bolus_window*, *sleep_score_window*, *steps_window*), indicating some associations exist between the lifestyle factors and BG changes. On the other hand, Patient 1's model achieved an $R^2$ value of 0.61, demonstrating a strong fit of the model. Patient 1's model also resulted in five statistically significant predictor variables including *carb_window*, *bolus_window*, *sleep_score_window*, *exercise_load_window*, and *work_load_window*, suggesting that multiple lifestyle factors might play important roles in managing BG for this patient.

Having an overall understanding of lifestyles' effect on the patients, we proceed to investigate the coefficients of each predictor to study if variance exists among patients. For all patients, *carb_window* and *bolus_window* show significance under the 0.01 significance level and achieved the overall largest absolute values of coefficients, where coefficients of *carb_window* are all positive, and coefficients of *bolus_window* are all negative. These findings are consistent with the strong rising and reducing effects of carbohydrates and bolus insulin on BG, respectively. This indicates that our method can correctly capture and model the dynamics of carbohydrate intake, bolus insulin, and BG.

*basal_window* is significant in Patients 1, 4, and 6, two of which have negative coefficients and the other with positive coefficients. Compared with *bolus_window*, the lower

frequency of significance and smaller absolute values of coefficients may be explained by basal insulin's slower and milder effect on BG. It is interesting to note that for Patient 1, the basal insulin rate has a positive association with BG change, despite its usual BG-lowering effect. It is possible that an increase in BG change may require a higher dosage of basal insulin to stabilize BG levels, leading to a positive association. Therefore, it is important to take caution when interpreting regression analysis results as there may be other underlying reasons contributing to the associations.

Among other short-term lifestyle predictors, Patients 1, 3, and 5 showed significance in *sleep_score_window* (two positive and one negative association), Patients 5, 6, and 7 showed significance in *exercise_load_window* (all three negative associations), Patients 1, 2, and 5 showed significance in *work_load_window* (two negative and one positive association), and Patient 3 showed significance in *steps_window* (one negative association). For the long-term lifestyle predictors, absolute values of coefficients are low, and only one patient had *work_load_24hrs* as a significant predictor. It could be inferred that the long-term effects of lifestyle factors on T1D patients' current BG change are less significant than the short-term effects. For the top lifestyle predictor defined in subsection II-D, Patients 1 and 3 have *sleep_score_window*, Patients 5, 6, and 7 have *exercise_load_window*, and Patient 2 has *work_load_window* as their top lifestyle predictors. Patient 4 doesn't have significant lifestyle predictors. The variety in top lifestyle predictors and their different directionality with BG change implies interpersonal differences exist in lifestyle's impact on BG.

With a closer look at each lifestyle predictor, the coefficients of significant *exercise_load_window* and *steps_window* predictors are all negative, consistent with physical activity's immediate lowering effect on BG levels. For significant *sleep_score_window* predictors, two patients (Patients 3 and 5) have positive coefficients, implying longer sleep or higher sleep quality is associated with higher BG raise. This finding may be explained by a rising BG trend during sleep for these two patients, suggesting sleep disturbances, potentially caused by sleep apnea, may be taking place [18]. Furthermore, the self-reported sleep quality rated from 1 to 3 has a limited scale and could be inaccurate as patients might consistently report the same value out of convenience or their subjective estimation of the quality could be biased. As a result, longer sleep with sleep disturbance could lead to higher sleep scores and is associated with higher BG raises. For *work_load_window*, the distinction in the directionality of coefficients might be attributed to different work types such as physical work and sedentary work which are not recorded in the OhioT1DM dataset. Overall, the insights gained from the regression analysis point towards interpersonal differences in the relationship between lifestyle factors and BG in T1D patients.

## IV. Discussion and Future Work

Based on the insights generated by our method, a personalized recommendation system could be developed. Lifestyle factors having the top associations with BG changes could be reported to T1D patients so that they can prioritize managing these lifestyle factors for optimal glycemic control. Guidelines on how to manage these factors could be provided based on the directionality of the coefficients. For instance, if the coefficient of a lifestyle factor is consistent with its established effect in the literature, we can guide patients on how to increase or reduce the conduct of that lifestyle factor to achieve lower BG levels. On the other hand, if the coefficient of a lifestyle factor is contrary to its established effect in the literature, further analysis should be done before recommendations are given. For instance, as mentioned in subsection III-B, positive associations exist between sleep score and BG changes in two T1D patients. A deeper analysis of the data rationalizes this finding as the BG raises during sleep, suggesting the occurrence of sleep disturbances. In this case, a warning could be given to patients to help them be aware and diagnose their potential sleep problems. The implementation of such a system needs future endeavors and is the topic of our future research. It is also essential to note that association doesn't necessarily equate to causal effect, and the actual effects of such recommendation systems should be further studied through controlled trials in the future.

With minor modifications, our method can be adapted to other T1D datasets collecting different sets of lifestyle factors. Just like what we do on the OhioT1DM dataset, a new set of events indicating the emergence of BG-affecting factors could be defined to segment the data into windows. Lifestyle and insulin data are then aggregated inside each window, with a distinct calculation of the ingested amount of carbohydrates and bolus insulin as described in subsection II-C. Subsequently, VIF values and the correlation matrix are used to guide the elimination or combination of factors that have a high correlation with other factors to avoid multicollinearity issues before generating the regression analysis results.

There are two limitations of this study. Firstly, though the performance of the MLR model outperformed other ML models significantly as shown in Table II, the MAE and RMSE values are still high. A more accurate model could potentially increase the $R^2$ values of MLR models and yield a more accurate estimation of each factor's association with the BG changes. Secondly, the data size is relatively small, limiting the analysis of the long-term effects of lifestyle factors. In our future work, we plan to study the effect of sample size on personalized analysis of lifestyle factors on BG by (a) collecting larger datasets and (b) exploring using smaller window lengths. A larger dataset could increase the model's fit to the complex T1D data and reduce prediction errors, also opening up possibilities to advanced ML models for better data modeling and lifestyle insights.

## V. Conclusion

In this study, we propose the first data-driven system to analyze the personalized effects of lifestyle factors on T1D patients. Through carefully designed data segmentation, feature engineering, and MLR-based regression analysis, the impact of insulin and lifestyle factors on T1D patients' BG change can be isolated to derive personalized lifestyle insights. Regression analysis results on the OhioT1DM dataset demonstrate significant interpersonal differences in lifestyle factors' effect on BG changes, underscoring the need for future research in personalized T1D lifestyle intervention.

## References


[1] D. Daneman, "Type 1 diabetes," *The Lancet*, vol. 367, no. 9513, pp. 847–858, 2006.

[2] A. D. Association, "Standards of medical care in diabetes—2022 abridged for primary care providers," *Clinical Diabetes*, vol. 40, no. 1, pp. 10–38, 2022.

[3] J. V. Nielsen, C. Gando, E. Joensson, and C. Paulsson, "Low carbohydrate diet in type 1 diabetes, long-term improvement and adherence: A clinical audit," *Diabetology & metabolic syndrome*, vol. 4, no. 1, pp. 1–5, 2012.

[4] A. E. Goebel-Fabbri, "Disturbed eating behaviors and eating disorders in type 1 diabetes: clinical significance and treatment recommendations," *Current diabetes reports*, vol. 9, no. 2, pp. 133–139, 2009.

[5] M. Chimen, A. Kennedy, K. Nirantharakumar, T. Pang, R. Andrews, and P. Narendran, "What are the health benefits of physical activity in type 1 diabetes mellitus? a literature review," *Diabetologia*, vol. 55, pp. 542–551, 2012.

[6] M. C. Riddell, I. W. Gallen, C. E. Smart, C. E. Taplin, P. Adolfsson, A. N. Lumb, A. Kowalski, R. Rabasa-Lhoret, R. J. McCrimmon, C. Hume, *et al.*, "Exercise management in type 1 diabetes: a consensus statement," *The lancet Diabetes & endocrinology*, vol. 5, no. 5, pp. 377–390, 2017.

[7] G. Nefs, E. Bazelmans, E. Donga, C. Tack, and B. de Galan, "Sweet dreams or bitter nightmare: a narrative review of 25 years of research on the role of sleep in diabetes and the contributions of behavioural science," *Diabetic Medicine*, vol. 37, no. 3, pp. 418–426, 2020.

[8] S. Reutrakul, A. Thakkinstian, T. Anothaisintawee, S. Chontong, A.-L. Borel, M. M. Perfect, C. C. P. S. Janovsky, R. Kessler, B. Schultes, I. A. Harsch, *et al.*, "Sleep characteristics in type 1 diabetes and associations with glycemic control: systematic review and meta-analysis," *Sleep medicine*, vol. 23, pp. 26–45, 2016.

[9] K. Sharif, A. Watad, L. Coplan, H. Amital, Y. Shoenfeld, and A. Afek, "Psychological stress and type 1 diabetes mellitus: what is the link?," *Expert review of clinical immunology*, vol. 14, no. 12, pp. 1081–1088, 2018.

[10] D. M. Minich, J. S. Bland, *et al.*, "Personalized lifestyle medicine: relevance for nutrition and lifestyle recommendations," *The Scientific World Journal*, vol. 2013, 2013.

[11] B. Zhu, G. M. Abu Irsheed, P. Martyn-Nemeth, and S. Reutrakul, "Type 1 diabetes, sleep, and hypoglycemia," *Current Diabetes Reports*, vol. 21, pp. 1–19, 2021.

[12] C. Marling and R. Bunescu, "The ohiot1dm dataset for blood glucose level prediction: Update 2020," in *CEUR workshop proceedings*, vol. 2675, p. 71, NIH Public Access, 2020.

[13] F. Brouns, I. Bjorck, K. Frayn, A. Gibbs, V. Lang, G. Slama, and T. Wolever, "Glycaemic index methodology," *Nutrition research reviews*, vol. 18, no. 1, pp. 145–171, 2005.

[14] C. Tonoli, E. Heyman, B. Roelands, L. Buyse, S. S. Cheung, S. Berthoin, and R. Meeusen, "Effects of different types of acute and chronic (training) exercise on glycaemic control in type 1 diabetes mellitus: a meta-analysis," *Sports medicine*, vol. 42, pp. 1059–1080, 2012.

[15] "Understanding Insulin on Board (IOB) Calculations — OpenAPS 0.0.0 documentation."

[16] A. Alin, "Multicollinearity," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 3, pp. 370–374, 2010.

[17] N. Shrestha, "Detecting multicollinearity in regression analysis," *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39–42, 2020.

[18] B. Zhu, G. M. Abu Irsheed, P. Martyn-Nemeth, and S. Reutrakul, "Type 1 Diabetes, Sleep, and Hypoglycemia," *Current Diabetes Reports*, vol. 21, p. 55, Dec. 2021.