



## Article

# Incorporating the Effect of Behavioral States in Multi-Step Ahead Deep Learning Based Multivariate Predictors for Blood Glucose Forecasting in Type 1 Diabetes

Mehrad Jaloli , William Lipscomb and Marzia Cescon \*

Department of Mechanical Engineering, University of Houston, Houston, TX 77004, USA

\* Correspondence: mcescon2@uh.edu; Tel.: +1-713-743-2037

**Abstract:** Behavioral factors can affect the blood glucose (BG) levels in people with type 1 diabetes (T1D), therefore, their effects need to be incorporated in blood glucose management for these individuals. Accordingly, in this work, we study the effect of two behavioral states, physical activity (PA) and stress state (SS), on BG fluctuations in individuals with T1D. We provide two methods for quantifying biomarkers related to PA and SS using raw acceleration (ACC) and electrodermal activity (EDA) data collected with a wearable device. We evaluate the impact of PA and SS on BG fluctuation by adding the derived behavior-related biomarkers in two cutting-edge deep learning-based glucose predictive models, a long short-term memory (LSTM) and a convolutional neural network (CNN)-LSTM network, for prediction horizons (PHs) of 30, 60 and 90 min. Through an ablation study, we demonstrate that incorporating the estimated behavior-related biomarkers improves the BG predictive model's performance obtaining mean absolute error (MAE)  $9.13 \pm 0.95$ ,  $17.75 \pm 1.93$  and  $31.85 \pm 2.88$  in [mg/dL], root mean square error (RMSE),  $12.35 \pm 1.06$ ,  $24.71 \pm 2.31$  and  $41.64 \pm 4.12$  in [mg/dL], and coefficient of determination (R<sup>2</sup>),  $95.34 \pm 3.34$ ,  $78.87 \pm 4.35$  and  $60.11 \pm 4.76$  in [%], for the LSTM model; and MAE  $9.37 \pm 0.88$ ,  $17.87 \pm 1.67$  and  $29.47 \pm 2.13$  in [mg/dL], RMSE  $12.51 \pm 1.40$ ,  $24.37 \pm 2.49$  and  $39.52 \pm 3.89$  in [mg/dL], and R<sup>2</sup>  $94.65 \pm 3.90$ ,  $78.37 \pm 4.11$  and  $61.12 \pm 4.30$  in [%], for the CNN-LSTM model, respectively, across all PHs. Additionally, we illustrate the generalizability of the proposed models by performing both population- and patient-wise.



**Citation:** Jaloli, M.; Lipscomb, W.; Cescon, M. Incorporating the Effect of Behavioral States in Multi-Step Ahead Deep Learning Based Multivariate Predictors for Blood Glucose Forecasting in Type 1 Diabetes. *Biomedinformatics* **2022**, *2*, 715–726. <https://doi.org/10.3390/biomedinformatics2040048>

Academic Editor: Jörn Lötsch

Received: 15 November 2022

Accepted: 8 December 2022

Published: 16 December 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** blood glucose management; continuous glucose monitoring (CGM); glucose forecasting; convolutional neural network (CNN); long short-term memory (LSTM); physical activity index (PAI); stress state index (SSI)

## 1. Introduction

Despite significant efforts devoted to the problem of blood glucose (BG) prediction in people with Type 1 Diabetes (T1D) over the last several decades [1–9] challenges associated with glycemic disturbances due to daily unrepresented inputs, such as behavioral states, are largely under researched. In this regard, perhaps the most crucial challenge is that of accounting for the effect of daily physical activity (PA), both structured and unstructured, and emotional or psychological stress state (SS), in the prediction of BG, as a stepping-stone toward the design of appropriate treatment decisions. To date, several studies have proposed various methodologies to investigate the impact of such glycemic disturbances on BG level fluctuations, enabled by the advent of continuous glucose monitoring (CGM) devices, and the growing use of consumer-grade wearable devices [10], which facilitated keeping track of BG levels as well as other physiological variables almost continuously and in a minimally invasive manner. Faccioli et al. [11] investigated the benefits of including steps count measured by an off-the-shelf wearable device, as a proxy for PA, together with insulin and meal information, in a linear black box model identified from actual patient data. The authors reported improved predictions accuracy when including PA in their model. Some studies included additional inputs related to physical exercise to the glucose

metabolism model: Refs. [12,13] used a quantized signal representing the level of exercise in an artificial neural network, while [14–16] employed an estimate of energy expenditure in linear multivariate models. Bertachi et al. [17] used multilayer perceptron (MLP) and support vector machine (SVM) to predict nocturnal hypoglycemia (NH) in the daily life of T1Ds including the effect of PA, however, they reported the manual feature extraction and high computational cost as the limitations of their methodology. More recently, Sevil et al. [18] demonstrated that by incorporating new features for the type and intensity of PA and acute psychological stress (APS), generated using ML techniques, as exogenous inputs into an adaptive system identification framework, the accuracy of one-hour-ahead BG prediction improved. De Paoli et al. [19] used a specific type of ANN, Jump Neural Network, to overcome the challenges associated with the prediction of abrupt changes in BG values caused by PA. However, as a result of testing on a small number of patients, they asserted that their results cannot be considered conclusive, and their method has an additional computational load that does not justify the insufficiently significant gain achieved.

Against this background, a more comprehensive model that is generalizable to a larger group of individuals with T1D, applicable to daily life events, computationally efficient, and has a robust performance, specifically for higher prediction horizons, is needed for improved BG management.

In a previous study [9], we have obtained encouraging preliminary results pertaining to BG prediction with a hybrid CNN-LSTM model using meal information, insulin intake and CGM data on a dataset [4] comprised of 59 individuals with T1D who had participated in a three-day in-hospital study. Our investigation demonstrated that our model achieved superior glucose forecasting for up to 90 min PH, compared to existing approaches in the literature. It is of interest, therefore, to study whether we can further improve the prediction accuracy and generalize our model to a wider range of people with T1D in the outpatient setting, by accounting for glycemic disturbances occurring in daily life including PA and SS. Our aim is therefore to characterize the effect of physical activity and stress on blood glucose fluctuations by using data collected with wearable devices, and incorporate it into state-of-the-art BG predictive models to achieve more accurate BG predictions.

To this end, we propose a 2-steps approach: in the first step, we derive biomarkers for PA and SS from raw accelerometer (ACC) and electrodermal activity (EDA) signals collected by a wearable device. In the second step, we combine the obtained biomarkers with the CGM, meal and insulin intakes in a multivariate dataset and feed it to our DL-based glucose predictive model to forecast the future BG values. We validate our novel approach on a publicly available dataset of 6 T1D patients whose data were recorded during an 8-weeks trial under free-living condition [20].

The remainder of the article is organized as follows. The Materials and Methods Section contains experimental data and a description of the data pre-processing pipeline followed by description of the proposed algorithms for quantifying PA and stress levels, as well as an explanation of the proposed prediction strategy. The Results Section presents the population-wise and patient-wise results, separately, as well as a comprehensive comparison with similar studies in the literature. Finally, the Discussion and Conclusion section contains a discussion of the proposed method's advantages and limitations, as well as the study's conclusion.

## 2. Materials and Methods

### 2.1. Experimental Condition

The dataset used in this work was the OH1OT1DM dataset [20] comprises of eight weeks of data collected from 12 people with T1D. The data consists of periodic blood glucose measurements, information regarding administered insulin, different biometric data, and self-reported meals and exercise information which was released in 2020 for the second edition of the BG Level Prediction (BGLP) Challenge [21], a competition in which the participating researchers were expected to produce BG predictions for 30 and 60 min

PH using a common dataset and evaluate the performances of their algorithms according to predefined metrics. The dataset is distributed in two separate files for the train and test sets. Each file contains data from six individuals with T1D identified anonymously as 540, 544, 552, 567, 584, and 596 who were on insulin pump therapy (Medtronic 530G and 630G [22]) and wore Medtronic Enlite [23] CGM sensors during the eight-week data collection period. The dataset includes BG concentration collected by a CGM device with a 5 min sampling rate, basal and bolus insulin doses, self-reported ingested carbohydrate (CHO) intake estimates, self-reported exercise, sleep, work, stress, and illness. Additionally, all subjects were required to wear an Empatica Embrace [24] fitness band throughout the study, which recorded one-minute aggregations of EDA, skin temperature, and magnitude of ACC.

The ACC and EDA signals presented many missing samples, as it is common in dataset collected in real-life situations. Previous studies that used this dataset omitted indeed these two signals because of this problem, and thus ignored the effect of PA and stress level on BG level prediction [25–29]. In our case, we used a linear interpolation algorithm to estimate the missing data samples when the gaps between samples were less than 60 min, while for gaps greater than 60 min, corresponding data were discarded.

All sensors' data were synchronized, and the variables of interest, namely BG, ACC, and EDA, were uniformly resampled at 15 min sampling intervals. Next, each feature was normalized and standardized independently to ensure that the model was not biased toward any particular variable.

## 2.2. Physical Activity Intensity Estimation

In a previous work [30] we proposed to detect PA and grade its intensity with a novel threshold-based algorithm. The method exploited the magnitude of the average third-order time derivative of the 3D displacement retrieved from three-axis accelerometer data low-pass filtered at 1[Hz] to isolate the static component due to gravity. User-determined activity intensity thresholds from sedentary to vigorous activity, were tuned to patient self-reported low, moderate and high fitness conditions, using more than 5000 h of outpatient subject data available. More detailed information regarding the methodology, are provided in the paper [30].

In the current study, we apply the same methodology, with appropriate modifications accounting for the differences in the sampling frequency of the raw ACC signals, to obtain the biomarker for physical activity intensity, which we call Physical Activity Intensity (PAI) from now on.

## 2.3. Stress State Estimation

In this work, we focus on detecting and monitoring changes in the sympathetic arousal state triggered by the fight-or-flight response to emotional stress. EDA signal has two components: tonic component which is a slow varying signal and phasic component which incorporates the sudden neural spikes corresponding to physiological changes. It has been demonstrated that an increase in the frequency of these spikes relates to an increase in patients' stress levels [31]. Accordingly, in [32], Wickramasuriya et al. proposed to deconvolve EDA signal to obtain neural impulses that stimulate sweat glands using a two-step optimization formulation that simultaneously recovers sparse neural stimuli and estimates physiological system parameters. Then, in a previous study [33], we proposed that sympathetic arousal states can be quantified from skin conductance measurements collected with a wrist-worn device.

For each stress event, the corresponding hidden cognitive state  $\hat{x}_k$  was defined as a function of previous cognitive states  $\hat{x}_{k-1}$ , state defining variable  $n_{k-1}$ , external stimuli variable,  $env_{k-1}$  and  $u_k$ , and noise component,  $w_k$ :

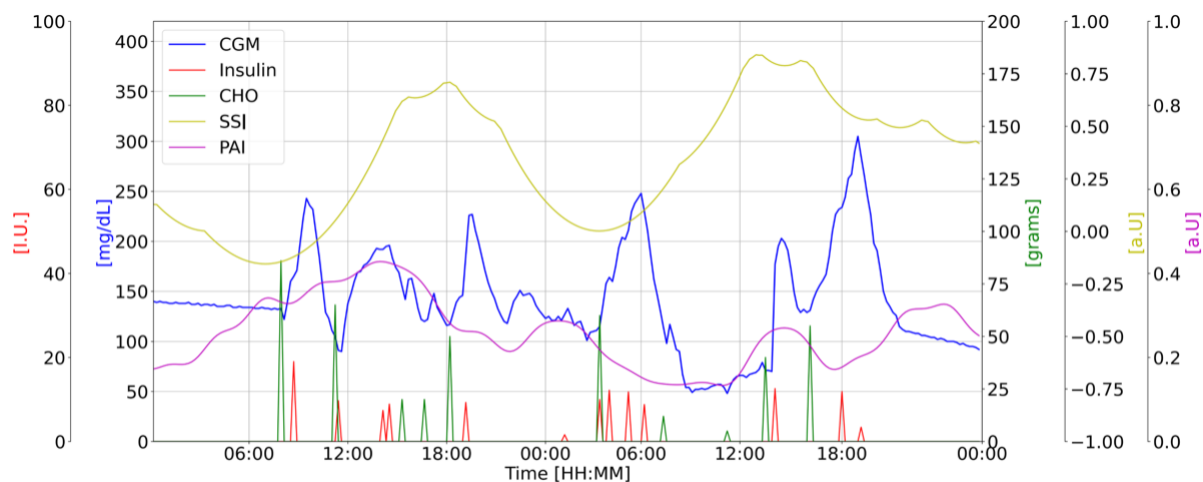
$$n_k = \begin{cases} 0 & \text{if no peak detected} \\ 1 & \text{if peak detected} \end{cases} \quad (1)$$

$$\hat{x}_k = f(x_{k-1}, n_{k-1}, u_k(n_{k-1}, env_{k-1}), w_k), \quad (2)$$

where  $u_k$  and  $env_k$  are obtained through assessing the stress event and modeling the increase in it at any timestamp. Therefore, the hidden state  $x_k$  was computed for each time point through an iterative process. More detailed information and formulation can be found in [33].

Correspondingly, in this study we leverage this methodology to obtain a biomarker representing the quantified estimation of stress level obtained from raw EDA signal in the OH1OT1DM dataset. We call this feature Stress State Index (SSI) from now on.

Figure 1 shows the available data along with the proposed biomarkers for a representative subject in the study.



**Figure 1.** Representative subject's data: CGM [mg/dL] (blue); CHO intake [g] (green); insulin doses [u] (red); PAI [a.u.] (purple) and SSI [a.u.] (yellow) vs. Time of the day [HH:MM].

## 2.4. Input-Output Partitioning

We concatenated and aligned the obtained PAI and SSI with BG concentration measured by CGM, self-reported ingested CHO and insulin intakes, to create the multivariate dataset to be used for prediction. Then, a sliding window with a step size of one was rolled over the multivariate dataset to generate sequences of corresponding input,  $\mathcal{X}$ , and output,  $\mathcal{Y}$ , sets for training the proposed DL-based BG prediction model depending on the user-specified PH. By construction, each windowed sample  $X_i \in \mathcal{X}$ ,  $X_i \in \mathbb{R}^{n \times 3PH}$  is a retrospective snapshot of all variables corresponding to a sequence of future CGM values,  $Y_i \in \mathcal{Y}$ ,  $Y_i \in \mathbb{R}^{1 \times PH}$ , such that:

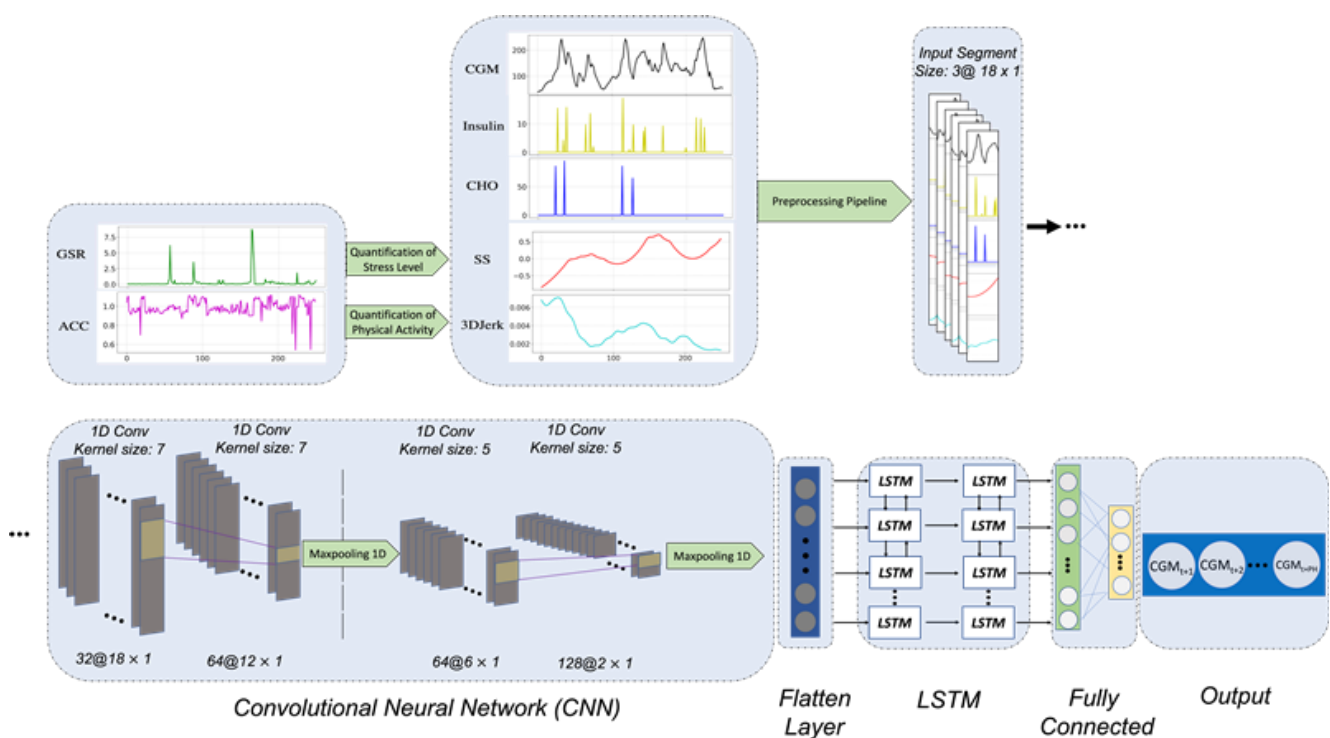
$$X_i(t) = \begin{bmatrix} x_1(t-3PH) & \cdots & x_1(t) \\ \vdots & \ddots & \vdots \\ x_n(t-3PH) & \cdots & x_n(t) \end{bmatrix} \quad (3)$$

$$Y_i(t) = [CGM(t+1), \cdots, CGM(t+PH)], \quad (4)$$

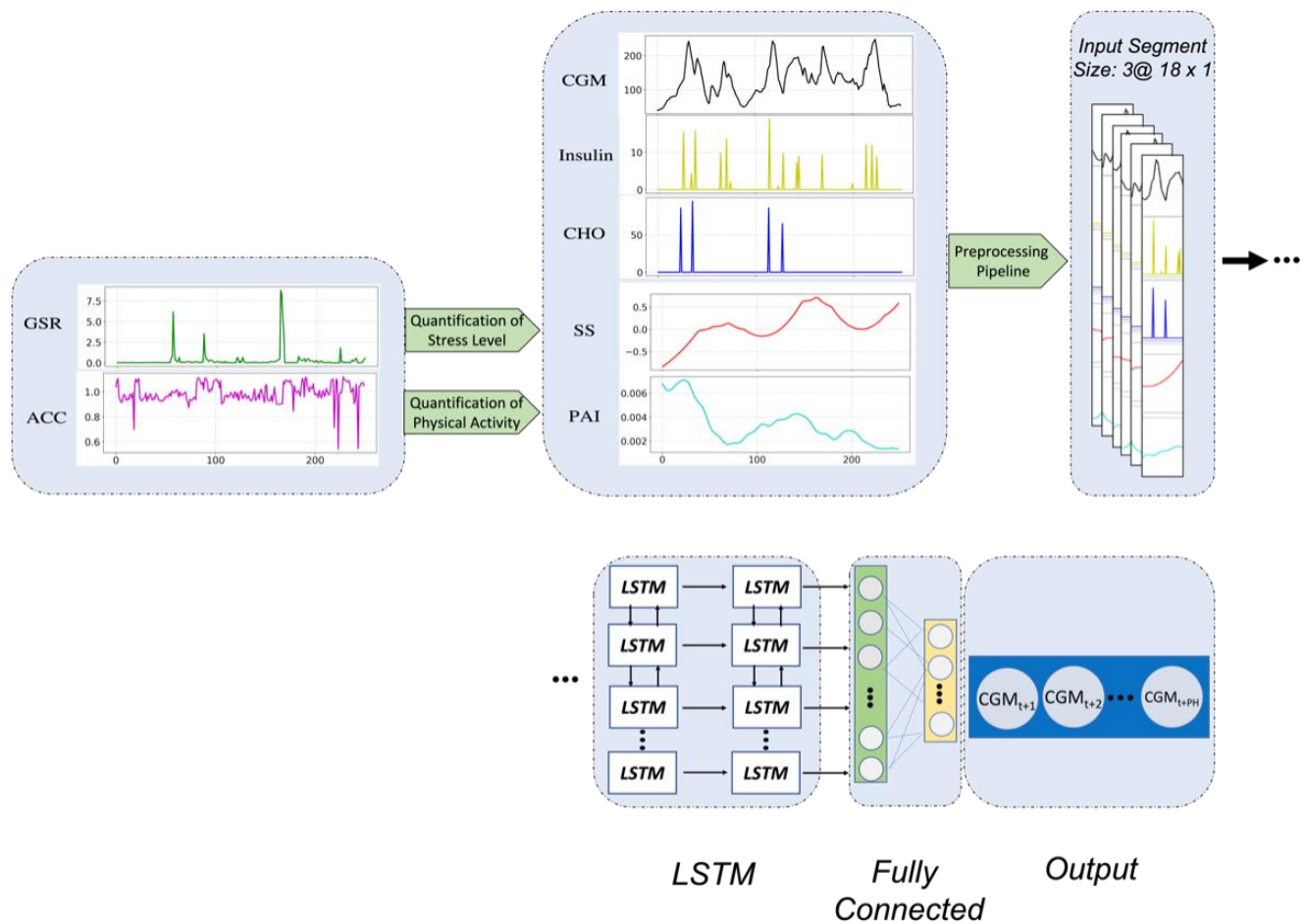
where  $i = [1, \cdots, n]$ , stands for the window index with  $n$  the total number of input windows generated by rolling the sliding window all over the multivariate time series.  $x_j, j = [1, \cdots, n]$  is a vector representing the  $j$ th feature with  $n$  showing the number of features used for training the model. We investigate the effect of PAI and SSI on the performance of the GC predictive model via an ablation study; thus,  $n$  varies depending on the scenario. For example, when CGM, CHO, and insulin are used as inputs to the model,  $n = 3$ , whereas when PAI and SSI are added to those three variables,  $n = 5$ .

### 2.5. Glucose Predictive Model

To investigate the feasibility of our hypothesis, we build on our previously published hybrid CNN-LSTM model [9]. The input samples are historical windows of data, containing BG values, meal, and insulin intakes information, as well as PAI and SSI samples, biomarkers as described in Sections 2.2 and 2.3, for time interval of  $(t - (3 \times PH), \dots, t)$  as the input, such that  $\hat{Y}_i(t) = f(X_i(t), W_k)$  for  $i = [1, \dots, n]$  with  $n$ , total number of windowed training samples,  $W_k$  with  $k = [1, \dots, m]$  model parameters and  $m$  number of parameters. These samples are sent into the CNN component so that significant features are extracted. The produced feature vectors are then input into two LSTM layers, which learn the variables' temporal dynamics and causal relationships. Finally, the output of the final LSTM layer is embedded into two fully connected layers, where each neuron is connected to every neuron in the preceding layer, in order to interpret the non-linear combination of these features and predicted future BG values  $\hat{Y}_i(t)$ , for up to a user specified PH, in the output layer. The proposed model architecture is depicted in Figure 2. For comparison purposes, we designed a second model comprised of two layers of LSTM network followed by three fully connected layers. The third fully connected layer is the output layer which contains predicted future BG values  $\hat{Y}_i(t)$ , for up to a user specified PH, given the input samples, structured in the same as what explained for the CNN-LSTM model. The proposed LSTM model architecture is depicted in Figure 3. It should be noted that for model architecture and hyperparameters selection, we run grid search to make sure the best possible model is designed.



**Figure 2.** The architecture of the proposed CNN-LSTM model. Raw ACC and EDA signals are used to estimate PAI and SSI features, respectively. Then, these generated features, along with the CGM, insulin, and CHO signals are passed through the preprocessing pipeline.



**Figure 3.** The architecture of the proposed LSTM model. Raw ACC and EDA signals are used to estimate PAI and SSI features, respectively. Then, these obtained features, are combined with the CGM, insulin, and CHO timeseries and are passed through the preprocessing.

The training procedure is carried out by backpropagating the error via the network layers and changing the weights  $W_k$  in such a way that the loss function is minimized, i.e., the model learns to predict the future BG value as close to the actual value as possible. Following each convolutional layer, a batch normalization operation is performed to re-center and scale the input features for the subsequent layer. By standardizing the parameters, this stabilizes the learning process and minimizes internal covariance shift. Additionally, in both models, a dropout layer is utilized following each LSTM layer to minimize overfitting by randomly setting certain input units to zero during training.

We used the TensorFlow framework [34] in Python 3.7 programming language to build and implement our models. The training set was divided into train and validation subsets with an 80:20 ratio, respectively. To minimize loss and optimize the cost function, the root mean square propagation (RMSProp) approach was adopted with a moving average parameter of 0.9 and an initial learning rate of 0.0001. The model was trained for 300 epochs with a batch size of 128 while an early stopping point strategy was utilized to minimize over-training by monitoring changes in validation loss with a patience of 50 epochs.

### 3. Results

In this section, we present results obtained on eight weeks of real-life data for six T1D patients from the OHIO-T1DM 2020 dataset [20].

### 3.1. Evaluation Metrics

Several metrics were utilized to evaluate the proposed model's performance and the influence of incorporating PAI and SSI in our computations: the mean absolute error (MAE), the root-mean-square error (RMSE), and coefficient of determination (R2). Equations (5)–(7) report the formulas used for their calculation, where  $n$  is the total number of windowed samples,  $Y_i(t)$  and  $\hat{Y}_i(t)$  are the actual and predicted GC values at time  $t + PH$  and  $\hat{Y}_i$  in R2 is the mean value of all samples:

$$MAE = \frac{\sum_{i=1}^n |Y_i(t) - \hat{Y}_i(t)|}{n} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i(t) - \hat{Y}_i(t))^2}{n}} \quad (6)$$

$$R2 = 1 - \frac{\sum_{i=1}^N (Y_i(t) - \hat{Y}_i(t))^2}{\sum_{i=1}^N (Y_i(t) - \bar{Y}_i(t))^2} \quad (7)$$

Using these metrics will enable us to compare our results to those of other research studies in the literature.

### 3.2. Evaluation Scenarios

We evaluated our proposed models with seven distinct scenarios, summarized in Table 1. Scenario 1 represents the baseline case, containing CGM, CHO and insulin intakes, while in Scenario 2 and 3, raw EDA and ACC signals, representing the effect of SS and PA, respectively, are included into computations, to investigate the effectiveness of incorporating PA and SS in BG predictive model's performance. In Scenario 4 and 5 the biomarkers for PAI and SSI are added to the baseline case, to be compared with scenarios 2 and 3, respectively, in terms evaluating the changes in results obtained as a result of the two quantification methodologies. Finally, Scenarios 6 and 7 incorporate the raw ACC and EDA, and PAI and SSI, respectively, to the baseline. The rationale behind defining these scenarios is to differentiate the influence that the raw EDA and ACC signals may have on BG prediction compared to that of the estimated biomarkers for SSI and PAI.

**Table 1.** Scenarios considered in our ablation study to understand the contribution of different variables on BG prediction performance.

| Scenario   | Variables                   |
|------------|-----------------------------|
| Scenario 1 | CGM, CHO, Insulin           |
| Scenario 2 | CGM, CHO, Insulin, EDA      |
| Scenario 3 | CGM, CHO, Insulin, ACC      |
| Scenario 4 | CGM, CHO, Insulin, SSI      |
| Scenario 5 | CGM, CHO, Insulin, PAI      |
| Scenario 6 | CGM, CHO, Insulin, ACC, EDA |
| Scenario 7 | CGM, CHO, Insulin, SSI, PAI |

### 3.3. Population-Wise Analysis

We employed training sets of all patients for population-wise analysis, and then test sets of all six individuals for performance evaluation in order to forecast future BG values for an unknown data sample. The model was trained on three distinct PHs, namely 30, 60, and 90 min. Tables 2 and 3 show the performance of the proposed LSTM and CNN-LSTM algorithms for different scenarios in terms of MAE [mg/dL], RMSE [mg/dL] and R2 [%] metrics, respectively.

**Table 2.** Population results for LSTM model. RMSE [mg/dL], MAE [mg/dL] and R2 [%] for 30, 60 and 90 min ahead BG prediction, for different scenarios.

| Scenario | 30 min              |                    |                     | 60 min              |                     |                     | 90 min              |                     |                     |
|----------|---------------------|--------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
|          | RMSE [mg/dL]        | MAE [mg/dL]        | R2 [%]              | RMSE [mg/dL]        | MAE [mg/dL]         | R2 [%]              | RMSE [mg/dL]        | MAE [mg/dL]         | R2 [%]              |
| 1        | 18.32 ± 2.53        | 12.98 ± 1.99       | 91.54 ± 4.32        | 33.26 ± 3.13        | 22.98 ± 2.90        | 66.92 ± 6.83        | 48.76 ± 5.61        | 37.11 ± 3.56        | 49.76 ± 6.93        |
| 2        | 17.64 ± 1.45        | 12.54 ± 1.15       | 92.11 ± 5.11        | 32.12 ± 3.45        | 21.44 ± 2.53        | 69.61 ± 6.13        | 47.21 ± 5.44        | 35.11 ± 3.12        | 52.09 ± 6.42        |
| 3        | 17.98 ± 1.85        | 12.81 ± 1.54       | 91.12 ± 4.21        | 31.07 ± 2.93        | 20.98 ± 1.99        | 69.30 ± 6.12        | 46.98 ± 6.01        | 34.60 ± 3.42        | 53.22 ± 5.94        |
| 4        | 12.47 ± 1.04        | 9.93 ± 0.93        | 94.35 ± 3.45        | 26.27 ± 1.96        | 18.05 ± 1.89        | 77.12 ± 4.56        | 42.90 ± 4.98        | 32.65 ± 2.91        | 55.74 ± 5.30        |
| 5        | 16.88 ± 1.56        | 11.97 ± 1.08       | 93.12 ± 5.12        | 32.15 ± 3.11        | 19.91 ± 2.80        | 70.34 ± 5.11        | 45.65 ± 5.24        | 33.65 ± 4.02        | 54.33 ± 5.65        |
| 6        | 17.11 ± 1.91        | 12.10 ± 1.34       | 92.43 ± 4.11        | 31.11 ± 2.89        | 19.59 ± 2.56        | 71.54 ± 5.67        | 45.11 ± 4.66        | 32.88 ± 3.76        | 55.08 ± 5.11        |
| 7        | <b>12.35 ± 1.06</b> | <b>9.13 ± 0.95</b> | <b>95.34 ± 3.34</b> | <b>24.71 ± 2.31</b> | <b>17.75 ± 1.93</b> | <b>78.87 ± 4.35</b> | <b>41.64 ± 4.12</b> | <b>31.85 ± 2.88</b> | <b>60.11 ± 4.76</b> |

**Table 3.** Population results for CNN-LSTM model. RMSE [mg/dL], MAE [mg/dL] and R2 [%] for 30, 60 and 90 min ahead BG prediction, for different scenarios.

| Scenario | 30 min              |                    |                     | 60 min              |                     |                     | 90 min              |                     |                     |
|----------|---------------------|--------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
|          | RMSE [mg/dL]        | MAE [mg/dL]        | R2 [%]              | RMSE [mg/dL]        | MAE [mg/dL]         | R2 [%]              | RMSE [mg/dL]        | MAE [mg/dL]         | R2 [%]              |
| 1        | 19.35 ± 2.67        | 13.31 ± 1.89       | 90.87 ± 5.53        | 32.68 ± 3.11        | 22.32 ± 3.11        | 68.82 ± 6.13        | 46.12 ± 5.10        | 34.20 ± 3.11        | 54.12 ± 5.53        |
| 2        | 18.65 ± 2.35        | 13.23 ± 1.82       | 90.98 ± 5.59        | 31.48 ± 2.90        | 21.12 ± 2.66        | 70.02 ± 5.73        | 45.56 ± 5.53        | 33.45 ± 3.21        | 56.34 ± 5.11        |
| 3        | 18.43 ± 2.53        | 13.02 ± 1.89       | 91.21 ± 4.51        | 31.12 ± 2.89        | 20.23 ± 2.11        | 70.34 ± 5.23        | 45.23 ± 5.11        | 33.30 ± 3.09        | 56.72 ± 5.56        |
| 4        | 12.63 ± 1.78        | 10.05 ± 1.01       | 93.75 ± 4.11        | 26.48 ± 2.83        | 18.79 ± 1.90        | 76.50 ± 4.41        | 40.59 ± 4.83        | 30.82 ± 2.20        | 59.82 ± 4.06        |
| 5        | 17.36 ± 2.56        | 12.37 ± 1.61       | 91.82 ± 4.90        | 30.15 ± 3.13        | 19.80 ± 2.22        | 72.34 ± 4.75        | 44.67 ± 4.92        | 31.90 ± 2.53        | 58.34 ± 5.03        |
| 6        | 17.28 ± 2.87        | 12.30 ± 1.53       | 91.90 ± 4.34        | 31.02 ± 3.05        | 19.59 ± 2.27        | 72.14 ± 4.90        | 43.67 ± 4.11        | 30.75 ± 2.90        | 59.12 ± 4.58        |
| 7        | <b>12.51 ± 1.40</b> | <b>9.37 ± 0.88</b> | <b>94.65 ± 3.90</b> | <b>25.37 ± 2.49</b> | <b>17.87 ± 1.67</b> | <b>78.37 ± 4.11</b> | <b>39.52 ± 3.89</b> | <b>29.47 ± 2.13</b> | <b>61.12 ± 4.30</b> |

Although adding raw EDA and ACC in Scenarios 2, 3 and 6 improved results slightly comparing to the baseline, using the computed biomarkers SSI and PAI in Scenarios 4, 5 and 7 significantly improved model performance, demonstrating the validity of our suggested methodologies for producing these two biomarkers and including them in the prediction model. Overall, the best results for both models are obtained in Scenario 7, as highlighted in bold face characters in the tables.

### 3.4. Patient-Wise Analysis

We evaluate the suggested models patient-by-patient, training and testing a tailored model on each patient's data, in order to address the inter-subject variability inherent in T1D. Tables 4 and 5 illustrate the proposed LSTM and CNN-LSTM model's performance in terms of MAE [mg/dL], RMSE [mg/dL], and R2 [%] metrics in predicting future BG for each patient individually, respectively, using the same patient's data as the training and test sets and Scenario 7 as the training scenario.

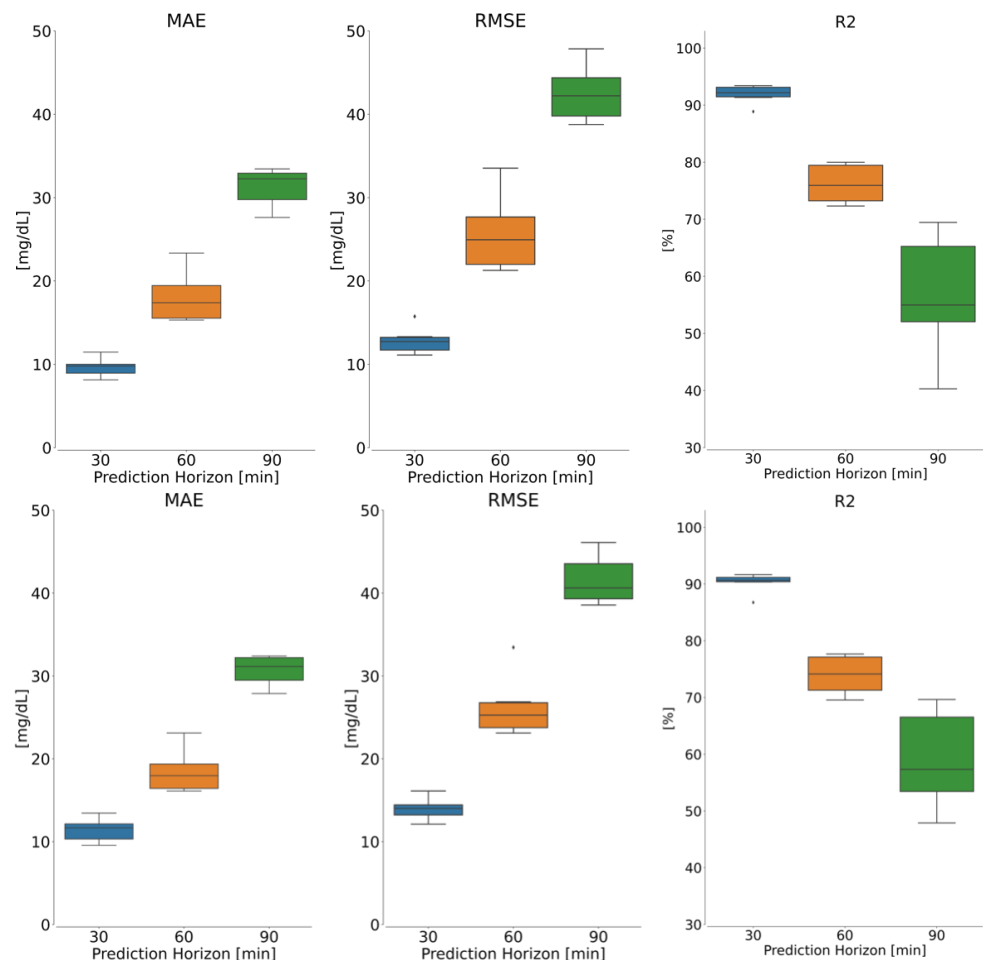
**Table 4.** Patient-wise analysis results with LSTM model, given scenario 7. MAE, RMSE and R2 of BG prediction for personalized training for each patient separately, for 30 to 90 min PH.

| Patient ID   | 30 min       |             |              | 60 min       |              |              | 90 min       |              |              |
|--------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|              | RMSE [mg/dL] | MAE [mg/dL] | R2 [%]       | RMSE [mg/dL] | MAE [mg/dL]  | R2 [%]       | RMSE [mg/dL] | MAE [mg/dL]  | R2 [%]       |
| 540          | 15.74        | 11.46       | 88.86        | 33.54        | 23.35        | 72.34        | 41.73        | 31.61        | 56.12        |
| 544          | 11.47        | 8.14        | 93.41        | 21.28        | 15.37        | 79.99        | 38.77        | 29.15        | 68.30        |
| 552          | 11.12        | 8.71        | 92.56        | 21.79        | 15.31        | 79.92        | 39.15        | 27.62        | 69.45        |
| 567          | 12.47        | 9.61        | 91.36        | 27.8         | 18.76        | 73.78        | 47.85        | 33.44        | 40.30        |
| 584          | 13.32        | 10.01       | 91.78        | 27.28        | 19.68        | 73.06        | 44.94        | 32.95        | 51.45        |
| 596          | 12.99        | 9.98        | 93.32        | 22.61        | 16.03        | 78.12        | 42.71        | 32.90        | 53.87        |
| Mean ± (STD) | 12.85 ± 1.50 | 9.65 ± 1.05 | 91.88 ± 1.54 | 25.71 ± 4.37 | 18.08 ± 2.89 | 76.20 ± 3.22 | 42.52 ± 3.17 | 31.27 ± 2.16 | 56.58 ± 9.01 |

**Table 5.** Patient-wise analysis results with CNN-LSTM model, given scenario 7. MAE, RMSE and R2 of BG prediction for personalized training for each patient separately, for 30 to 90 min PH.

| Patient ID       | 30 min           |                  |                  | 60 min           |                  |                  | 90 min           |                  |                  |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                  | RMSE [mg/dL]     | MAE [mg/dL]      | R2 [%]           | RMSE [mg/dL]     | MAE [mg/dL]      | R2 [%]           | RMSE [mg/dL]     | MAE [mg/dL]      | R2 [%]           |
| 540              | 16.12            | 13.45            | 86.76            | 33.45            | 23.11            | 69.56            | 40.12            | 30.61            | 58.11            |
| 544              | 12.98            | 9.98             | 91.65            | 23.65            | 16.11            | 77.67            | 38.56            | 29.11            | 69.35            |
| 552              | 12.12            | 9.56             | 90.56            | 23.11            | 16.23            | 77.34            | 39.05            | 27.88            | 69.65            |
| 567              | 13.98            | 11.34            | 90.36            | 26.44            | 18.87            | 71.12            | 46.11            | 32.41            | 47.89            |
| 584              | 14.56            | 12.18            | 90.78            | 26.87            | 19.54            | 71.76            | 44.35            | 31.66            | 52.41            |
| 596              | 14.06            | 12.01            | 91.32            | 24.11            | 17.05            | 76.54            | 41.12            | 32.40            | 56.57            |
| Mean $\pm$ (STD) | 13.97 $\pm$ 1.25 | 11.42 $\pm$ 1.32 | 90.23 $\pm$ 1.61 | 26.27 $\pm$ 3.50 | 18.48 $\pm$ 2.43 | 73.99 $\pm$ 3.26 | 41.55 $\pm$ 2.77 | 30.67 $\pm$ 1.69 | 58.99 $\pm$ 8.10 |

Additionally, the box plots in Figure 4 depicts the distribution of the resulting metrics for the patient-wise analysis for both models. Although the model performance degrades when compared to the population-wise analysis, the distribution of the findings across all individuals remains favorable, indicating the generalizability of the suggested methods across a diverse collection of T1D patients.

**Figure 4.** Box plot showing the patient-wise results for (top) LSTM and (bottom) CNN-LSTM models in terms of (left) MAE, (middle) RMSE and (right) R2 metrics, given scenario 7 as dataset. In each panel, blue, orange and green are boxplots for 30, 60 and 90 min.

### 3.5. Comparison with Existing Methods

Table 6 compares our results to the top five ranked research works that participated in the 2020 BGLP challenge [21]. Studies were evaluated based on MAE [mg/dL] and RMSE

[mg/dL] for 30 and 60 min PH and ranked according to the overall accuracy defined as the summation of MAE and RMSE in [mg/dL]. We see that our work outperforms the previous research across all metrics and overall performance. Contributions are identified by their Paper ID, as syndicated on the official BGLP challenge ranking page [21].

**Table 6.** Comparison of our results and the top five ranked research works that participated in the 2020 BGLP challenge. Population mean  $\pm$  (standard deviation) of accuracy metrics (MAE and RMSE) for 30 and 60 min ahead BG prediction.

| Paper ID              | 30 Min       |             | 60 Min       |              | Overall [mg/dL] |
|-----------------------|--------------|-------------|--------------|--------------|-----------------|
|                       | RMSE [mg/dL] | MAE [mg/dL] | RMSE [mg/dL] | MAE [mg/dL]  |                 |
| 13                    | 18.22        | 12.83       | 31.66        | 23.60        | 86.31           |
| 6                     | 19.21        | 13.08       | 31.77        | 23.09        | 87.15           |
| 16                    | 18.34        | 13.37       | 32.21        | 24.20        | 88.12           |
| 15                    | 19.05        | 13.50       | 32.03        | 23.83        | 88.41           |
| 1                     | 18.23        | 14.37       | 31.10        | 25.75        | 89.45           |
| CNN-LSTM (Scenario 7) | <b>12.51</b> | <b>9.37</b> | <b>25.37</b> | <b>17.87</b> | <b>65.30</b>    |
| LSTM (Scenario 7)     | <b>12.35</b> | <b>9.13</b> | <b>24.71</b> | <b>17.75</b> | <b>63.94</b>    |

#### 4. Discussion and Conclusions

In this work, we analyzed whether the incorporation of our proposed biomarkers for PA and SS, computed from raw ACC and EDA signals collected with a commercially available wearable device, improved the BG prediction accuracy for PH ranging from 30 to 90 min. We compare the performances of our previously published CNN-LSTM prediction model [9] with those of a novel LSTM-based predictor. At a population level, as far as the CNN-LSTM model is concerned, our results demonstrate a decrease in RMSE of  $6.84 \pm 3.01$  [mg/dL],  $7.31 \pm 2.08$  [mg/dL] and  $6.6 \pm 6.41$  [mg/dL] for 30, 60, 90 min PH, respectively; a decrease in MAE of  $3.96 \pm 2.08$  [mg/dL],  $4.45 \pm 3.53$  [mg/dL] and  $4.73 \pm 3.77$  [mg/dL] for 30, 60, 90 min PH, respectively; and an increase in R2 of  $3.78 \pm 6.77$  [%],  $9.55 \pm 7.38$  [%] and  $7 \pm 7.01$  [%] for 30, 60, 90 min PH, respectively, from Scenario 1 to Scenario 7. While considering the LSTM model, instead, the reduction in RMSE is  $5.97 \pm 2.74$  [mg/dL];  $8.55 \pm 2.33$  [mg/dL] and  $7.12 \pm 6.96$  [mg/dL] for 30, 60, 90 min PH, respectively; a reduction in MAE of  $3.85 \pm 2.21$  [mg/dL],  $5.23 \pm 3.48$  [mg/dL] and  $5.26 \pm 4.58$  [mg/dL] for 30, 60, 90 min PH, respectively; and an increase in R2 of  $3.8 \pm 5.46$  [%],  $11.95 \pm 8.10$  [%] and  $10.35 \pm 8.41$  [%], respectively, when comparing Scenario 1 with Scenario 7. These findings validate our hypothesis that adding our proposed biomarkers is indeed beneficial for improved BG prediction, especially on longer PH. When comparing the 2 proposed models, we report similar accuracy for the CNN-LSTM and the LSTM for shorter PHs (30 and 60 min), however for 90 min PH, CNN-LSTM has a minor superiority over LSTM. This acknowledges the capacity of the CNN model in extracting significant features from more complicated datasets, given the fact that for 90 min PH, we trained the model with longer data windows as input sample, making it difficult for LSTM to deal with it. Additionally, we demonstrated that our technique outperforms all prior research that utilized the OHIOT1DM 2020 dataset and were participants in the second BGLP competition.

We acknowledge several limitations in our study. First of all, since data recording occurred in a free-living environment, we reported missed data points for a variety of reasons, including the patient's discomfort with the sensor being attached to his/her body at all times, particularly during sleep or activity. This may have a substantial impact on the model's performance. For example, during the preprocessing steps, due to the presence of numerous missing samples in the OHIOT1DM dataset, specifically in variables recoded by the Empatica E4 sensor, we had to use a variety of mathematical techniques to compensate for the missing samples, such as interpolation and extrapolation for the training and test sets, respectively, or even discarding some parts, where the gap was greater than one hour

in the dataset. This undoubtedly has a detrimental effect on the model's performance and liability.

Finally, we assert that, adding the proposed biomarkers, obtained by the state-of-the-art signal processing and mathematical approaches, to the physiological information recorded from individuals with T1D improved the performance of BG predictive models. Such models, if integrated into a decision support systems or artificial pancreas (AP) system, can extend the time in the euglycemic range for T1D patients, which is one of the most crucial concerns in diabetes management.

**Author Contributions:** Conceptualization, M.J. and M.C.; Formal analysis, M.J., W.L. and M.C.; Methodology, M.J. and M.C.; Software, M.J.; Supervision, M.C.; Validation, M.J. and M.C.; Investigation, M.J., W.L. and M.C.; Data curation, M.J. and W.L.; Writing original draft, M.J.; Writing—review editing, M.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the University of Houston-National Research University Fund (NRUF): R0504053.

**Data Availability Statement:** The dataset used in this work is publicly available and can be obtained at: <http://smarthealth.cs.ohio.edu/OhioT1DM-download.html> (accessed on 15 August 2022). Moreover, Ref. [20] is the paper describing the dataset.

**Conflicts of Interest:** Marzia Cescon serves on the advisory board for Diatech Diabetes, Inc. Mehrad Jaloli declares no conflict of interest relevant to this project.

## References

1. Cescon, M.; Ståhl, F.; Landin-Olsson, M.; Johansson, R. Subspace-based model identification of diabetic blood glucose dynamics. *IFAC Proc. Vol.* **2009**, *42*, 233–238. [\[CrossRef\]](#)
2. Cescon, M.; Johansson, R. Glycemic trend prediction using empirical model identification. In Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, Shanghai, China, 16–18 December 2009; pp. 3501–3506.
3. Percival, M.W.; Bevier, W.C.; Zisser, H.; Jovanović, L.; Seborg, D.E.; Doyle, F.J., III. Prediction of dynamic glycemic trends using optimal state estimation. *IFAC Proc. Vol.* **2008**, *41*, 4222–4227. [\[CrossRef\]](#)
4. Cescon, M. Modeling and Prediction in Diabetes Physiology. Ph.D. Thesis, Lund University, Lund, Sweden, 2013. Available online: <http://archive.control.lth.se/Research/medicalProjects/diadvistorm.html> (accessed on 15 February 2022).
5. Bunescu, R.; Struble, N.; Marling, C.; Shubbrook, J.; Schwartz, F. Blood glucose level prediction using physiological models and support vector regression. Proceedings of 2013 12th International Conference on Machine Learning and Applications, Washington, DC, USA, 4–7 December 2013; Volume 1, pp. 135–140.
6. Oviedo, S.; Vehí, J.; Calm, R.; Armengol, J. A review of personalized blood glucose prediction strategies for T1DM patients. *Int. J. Numer. Method Biomed. Eng.* **2017**, *33*, e2833. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Li, K.; Daniels, J.; Liu, C.; Herrero, P.; Georgiou, P. Convolutional recurrent neural networks for glucose prediction. *IEEE J. Biomed. Health Inf.* **2019**, *24*, 603–613. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Kushner, T.; Breton, M.D.; Sankaranarayanan, S. Multi-hour blood glucose prediction in type 1 diabetes: A patient-specific approach using shallow neural network models. *Diabetes Technol.* **2020**, *22*, 883–891. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Jaloli, M.; Cescon, M. Predicting Blood Glucose Levels Using CNN-LSTM Neural Networks. In *2020 Diabetes Technology Meeting Abstracts*, 2nd ed.; SAGE Publications: Los Angeles, CA, USA, 2020; Volume 15, p. 432.
10. ECRI. The Growing Use of Consumer-Grade Medical Devices: Advice for Physicians and Their Patients. *Health Devices*. 2019. Available online: <https://www.ecri.org/components/HDJournal/Pages/Advice-on-consumer-grade-medical-devices.aspx> (accessed on 14 November 2022).
11. Faccioli, S.; Ozaslan, B.; Garcia-Tirado, J.F.; Breton, M.; del Favero, S. Black-box model identification of physical activity in type-1 diabetes patients. Proceedings of 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 3910–3913.
12. De Canete, J.F.; Gonzalez-Perez, S.; Ramos-Diaz, J.C. Artificial neural networks for closed loop control of in silico and ad hoc type 1 diabetes. *Comput. Methods Programs Biomed.* **2012**, *106*, 55–66. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Fong, S.; Zhang, Y.; Fiaidhi, J.; Mohammed, O.; Mohammed, S. Evaluation of stream mining classifiers for real-time clinical decision support system: A case study of blood glucose prediction in diabetes therapy. *Biomed. Res. Int.* **2013**, *2013*, 4193. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Turksoy, K.; Hajizadeh, I.; Hobbs, N.; Kilgus, J.; Littlejohn, E.; Samadi, S. Multivariable artificial pancreas for various exercise types and intensities. *Diabetes Technol. Ther.* **2018**, *20*, 662–671. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Turksoy, K.; Bayrak, E.S.; Quinn, L.; Littlejohn, E.; Rollins, D.; Cinar, A. Hypoglycemia early alarm systems based on multivariable models. *Ind. Eng. Chem. Res.* **2013**, *52*, 12329–12336. [\[CrossRef\]](#) [\[PubMed\]](#)

16. Zarkogianni, K.; Mitsis, K.; Litsa, E.; Arredondo, M.-T.; Fico, G.; Fioravanti, A.; Nikita, K.S. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Med. Biol. Eng. Comput.* **2015**, *53*, 1333–1343. [[CrossRef](#)] [[PubMed](#)]
17. Bertachi, A.; Viñals, C.; Biagi, L.; Contreras, I.; Vehí, J.; Conget, I.; Giménez, M. Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor. *Sensors* **2020**, *20*, 1705. [[CrossRef](#)] [[PubMed](#)]
18. Sevil, M.; Rashid, M.; Hajizadeh, I.; Park, M.; Quinn, L.; Cinar, A. Physical activity and psychological stress detection and assessment of their effects on glucose concentration predictions in diabetes management. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 2251–2260. [[CrossRef](#)] [[PubMed](#)]
19. De Paoli, B.; D'Antoni, F.; Merone, M.; Pieralice, S.; Piemonte, V.; Pozzilli, P. Blood Glucose Level Forecasting on Type-1-Diabetes Subjects during Physical Activity: A Comparative Analysis of Different Learning Techniques. *Bioengineering* **2021**, *8*, 72. [[CrossRef](#)] [[PubMed](#)]
20. Marling, C.; Bunesco, R. The OhioT1DM dataset for blood glucose level prediction: Update 2020. *NIH Public Access* **2020**, 2675, 71.
21. Blood Glucose Level Prediction Challenge 2020. Available online: <http://smarthealth.cs.ohio.edu/bglp/bglp-results.html> (accessed on 8 December 2021).
22. Medtronic Insulin Pump Systems. Available online: <https://www.medtronic.com/us-en/healthcare-professionals/products/diabetes/insulin-pump-systems.html> (accessed on 27 January 2022).
23. Medtronic ENLITETM Glucose Sensor. Available online: <https://www.medtronic.com/ca-en/diabetes/home/products/cgm-systems/enlite-sensor.html> (accessed on 27 January 2022).
24. Empatica Embrace. Available online: <https://www.empatica.com/index.html> (accessed on 27 January 2022).
25. Bevan, R.; Coenen, F. Experiments in non-personalized future blood glucose level prediction. *CEUR Workshop Proc.* **2020**, 2675, 100–104.
26. Hameed, H.; Kleinberg, S. Investigating potentials and pitfalls of knowledge distillation across datasets for blood glucose forecasting. In Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data, Santiago de Compostela, Spain and Virtually, 29–30 August 2020.
27. Rubin-Falcone, H.; Fox, I.; Wiens, J. Deep Residual Time-Series Forecasting: Application to Blood Glucose Prediction. *KDH@ECAI 2020*. Available online: <https://ceur-ws.org/Vol-2675/paper18.pdf> (accessed on 14 November 2022).
28. Yang, T.; Wu, R.; Tao, R.; Wen, S.; Ma, N.; Zhao, Y. Multi-Scale Long Short-Term Memory Network with Multi-Lag Structure for Blood Glucose Prediction. *KDH@ECAI 2020*, *45*, 136–140.
29. Zhu, T.; Yao, X.; Li, K.; Herrero, P.; Georgiou, P. Blood glucose prediction for type 1 diabetes using generative adversarial networks. *CEUR Workshop Proc.* **2020**, 2675, 90–94.
30. Cescon, M.; Choudhary, D.; Pinsker, J.E.; Dadlani, V.; Church, M.M.; Kudva, Y.C. Activity detection and classification from wristband accelerometer data collected on people with type 1 diabetes in free-living conditions. *Comput. Biol. Med.* **2021**, *135*, 104633. [[CrossRef](#)] [[PubMed](#)]
31. Wickramasuriya, D.S.; Qi, C.; Faghih, R.T. A state-space approach for detecting stress from electrodermal activity. Proceedings of 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 3562–3567.
32. Wickramasuriya, D.S.; Amin, M.; Faghih, R.T. Skin conductance as a viable alternative for closing the deep brain stimulation loop in neuropsychiatric disorders. *Front. Neurosci.* **2019**, *13*, 780. [[CrossRef](#)] [[PubMed](#)]
33. Choudhary, D.; Cescon, M. EDA-sense: Dynamic Feedback Control of Sympathetic Arousal. *IFAC PapersOnLine* **2020**, *53*, 238–243. [[CrossRef](#)]
34. Martín, A.; Paul, B.; Jianmin, C.; Tensorflow, C.Z. A system for largescale machine learning. In Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.