applied sciences

MDPI

Article

Characterizing the Impact of Physical Activity on Patients with Type 1 Diabetes Using Statistical and Machine Learning Models

David Chushig-Muzo [1,*], Hugo Calero-Díaz [1], Himar Fabelo [2,3,4], Eirik Årsand [5], Peter Ruben van Dijk [6,7] and Cristina Soguero-Ruiz [1]

[1] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Fuenlabrada, 28943 Madrid, Spain; h.calero.2017@alumnos.urjc.es (H.C.-D.); cristina.soguero@urjc.es (C.S.-R.)
[2] Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC), 35012 Las Palmas de Gran Canaria, Spain; hfabelo@iuma.ulpgc.es
[3] Research Unit, Hospital Universitario de Gran Canaria Dr. Negrin, 35010 Las Palmas de Gran Canaria, Spain
[4] Institute for Applied Microelectronics, Universidad de Las Palmas de Gran Canaria, 35001 Las Palmas de Gran Canaria, Spain
[5] Department of Computer Science, Faculty of Science and Technology, UiT The Arctic University of Norway, 9019 Tromsø, Norway; eirik.arsand@uit.no
[6] Department of Internal Medicine, Divisions of Endocrinology, Isala, Diabetes Center, 8025 AB Zwolle, The Netherlands; p.r.van.dijk@umcg.nl
[7] University Medical Center Groningen, University of Groningen, 9712 CP Groningen, The Netherlands
* Correspondence: david.chushig@urjc.es

Abstract: Continuous glucose monitoring (CGM) represents a significant advancement in diabetes management, playing an important role in glycemic control for patients with type 1 diabetes (T1D). Despite their benefits, their performance is affected by numerous factors such as the carbohydrate intake, alcohol consumption, and physical activity (PA). Among these, PA could cause hypoglycemic episodes, which might happen after exercising. In this work, two main contributions are presented. First, we extend the performance evaluation of two glucose monitoring devices, Eversense and Free Style Libre (FSL), for measuring glucose concentrations during high-intensity PA and normal daily activity (NDA). The impact of PA is investigated considering (1) different glucose ranges (hypoglycemia, euglycemia, and hyperglycemia); and (2) four time periods throughout the day (morning, afternoon, evening, and night). Second, we evaluate the effectiveness of machine learning (ML) models, including logistic regression, K-nearest neighbors, and support vector machine, to automatically detect PA in T1D individuals using glucose measurements. The performance analysis showed significant differences between glucose levels obtained in the PA and NDA period for Eversense and FSL devices, specially in the hyperglycemic range and two time intervals (morning and afternoon). Both Eversense and FSL devices present measurements with large variability during strenuous PA, indicating that their users should be cautious. However, glucose recordings provided by monitoring devices are accurate for NDA, reaching similar values to capillary glucose device. Lastly, ML-based models yielded promising results to determine when an individual has performed PA, reaching an accuracy value of 0.93. The results can be used to develop an individualized data-driven classifier for each patient that categorizes glucose profiles based on the time interval during the day and according to if a patient performs PA. Our work contributes to the analysis of PA on the performance of CGM devices.

Keywords: continuous glucose monitoring; type 1 diabetes; physical activity; machine learning; TabPFN

Citation: Chushig-Muzo, D.; Calero-Díaz, H.; Fabelo, H.; Årsand, E.; van Dijk, P.R.; Soguero-Ruiz, C. Characterizing the Impact of Physical Activity on Patients with Type 1 Diabetes Using Statistical and Machine Learning Models. *Appl. Sci.* **2024**, *14*, 9870. https://doi.org/10.3390/app14219870

Academic Editors: Grigorios Beligiannis, Adnane Cabani and Kévin Bouchard

Received: 2 September 2024
Revised: 8 October 2024
Accepted: 20 October 2024
Published: 29 October 2024

1. Introduction

The prevalence of type 1 diabetes (T1D) is increasing about 2–3% per year worldwide [1]. T1D is a chronic disease caused by the destruction of insulin-producing beta cells in the pancreatic islets, leading to either absolute or partial insulin deficiency [2]. To

treat this disease and maintain normal glycemic levels, patients require exogenous insulin administration via either multiple injections or continuous subcutaneous insulin infusion such as pumps [2]. The insulin dose is adjusted according to metabolic needs defined mainly by glucose concentrations, carbohydrate intake, and physical activity (PA) [3]. The monitoring and control of glucose is essential to achieve target glycemic control, determining more accurate insulin doses, and diminishing insulin-related complications such as hypoglycemia [4].

Capillary blood glucose monitoring devices have historically been the standard and most established technology for the management of T1D. However, continuous glucose monitoring (CGM) devices have become one of the most relevant advancements, showing similar accuracy to capillary glucose devices [5]. CGM devices use subcutaneous tissue sensors that provide interstitial fluid glucose measurements at regular time intervals and without need for frequent capillary blood glucose measurements [6]. CGM devices provide periodic information about the glycemic levels and may serve as an alarming tool in the onset of hyperglycemic and hypoglycemic events. Several clinical trials have demonstrated that CGM devices help to better adjust insulin doses, improve glycemic control and maintain target glycated hemoglobin [7]. However, the glucose measurements and the performance of CGM devices can be affected by several factors such as administered insulin, carbohydrate intake, and PA [8].

A balanced diet and regular exercise are highly recommended by physicians to maintain a healthy lifestyle. Several studies have shown that PA improves blood lipid profiles, psychological well-being, and may reduce cardiovascular disease risk [9–11]. Despite these benefits, glucose concentrations during PA are highly variable, increasing the use of glucose storage and the risk of glycemic events [11]. For T1D patients, maintaining glycemic control during and after PA is challenging because the risk of hypoglycemia is increased for up to 24 h after the bout of exercise [12]. Additionally, strenuous exercise produces some metabolic changes, particularly in the microcirculation and oxygen tension, which may negatively affect the sensor accuracy of CGM devices [13]. Therefore, it is relevant to evaluate the performance of CGM devices in different scenarios with PA and compare glucose measurements during normal daily activity (NDA). This would help one to make better decisions about the amount of food to consume and amount of exogenous insulin required, thus mitigating the onset of acute events. Several studies have investigated the performance of CGM devices during different activities, including, but not limited to high-intensity interval training, long-distance running, and skiing [14–17]. However, the performance of CGM devices during strenuous PA is still under study, and it is crucial to perform studies that evaluate the accuracy of these devices in other scenarios with high-intensity PA.

Over recent years, machine learning (ML) methods have attracted great attention in both academia and industry for the outstanding predictive performance in multiple applications [18–20]. Many researchers have applied ML models in diabetes research for detecting glycemic events [21,22], predicting glucose levels [23,24], and identifying clinical patterns [25], among others. ML-based models are promising to automatically detect and classify whether a patient has performed PA or not, which can help T1D individuals in the management of exercise-induced glycemic changes, thus mitigating long-term health risks [26]. In the literature, most of the studies proposed to detect PA rely on the use of heart rate sensors, accelerometers, and other movement sensors [27–29]. This study makes sense given that many patients have easy access to these technologies and only few studies have considered glucose data to detect and classify PA [30,31].

In this work, two main contributions are presented. First, we extend the analysis of the impact of high-intensity PA (particularly, mountain biking) on glucose concentrations measured by two CGM devices, the FreeStyle Libre Flash (FSL) and Eversense, which are based on glucose oxidase and fluorescence, respectively. A statistical and comparative analysis were conducted by employing glucose measurements from each of these CGM devices against reference values obtained from capillary glucose measurements using two-group tests. The impact of PA is investigated during two periods, including PA and NDA, and

considering (i) different glucose ranges (hypoglycemia, euglycemia, and hyperglycemia); and (ii) four time intervals throughout the day (morning, afternoon, evening, and night). Second, we evaluate the effectiveness of using CGM data and supervised ML-based models to automatically detect PA in T1D individuals. To achieve this, we employed data from the Bas van de Goor Foundation challenge, where 23 T1D subjects measured glucose with devices during periods of strenuous PA and NDA. T-tests and Mann–Whitney U tests were performed to statistically measure differences in glucose ranges and time intervals during NDA and PA. Logistic regression (LR), K-nearest neighbors (KNN), and support vector machine (SVM) were used to classify PA using glucose measurements.

2. Materials and Methods

In this section, we present the study design and the dataset description and preprocessing. Then, we detail the parametric and nonparametric statistical tests to measure significant differences between glucose measurements in different glucose ranges and intervals during the day, and the ML models used to identify whether a T1D individual performed PA or not. Finally, we present the feature selection methods and post hoc interpretability methods used to identify the most relevant features involved in the detection of PA.

2.1. Study Design and Participants

In this study, we employed the data collected at the Bas van de Goor Foundation challenge, where 23 T1D subjects (10 men and 13 women) from Spain and the Netherlands performed high-intensity PA using mountain bikes in Sierra Nevada [32]. The participants covered a total distance of 263 km at altitudes between 4753 and 11,000 m. Data acquisition was conducted by the authors in [32], a prospective and observational study where the performance of CGM devices was also analyzed during strenuous PA. The study was approved by the Medical Ethical committees in Spain (Hospital Universitario Central de Asturias; 163/18) and the Netherlands (Isala Hospital; NL66388.075.18/180603) and registered in the Dutch trial register (https://www.onderzoekmetmensen.nl/en number NL7133). The authors in [32] obtained the written informed consent by the 23 T1D participants.

The study was conducted during 12 days (two weeks) in two separate periods. In the first period (week 1), subjects performed high-intensity PA, while in the second period (week 2) after mountain biking, subjects carried out NDA with no sports activities. Three different devices were used to record glucose concentrations during the periods of PA and NDA: (i) the capillary device named Free Style Libre Precision NeoPro strip (shortened to FSLCstrip); (ii) the fluorescence-based and subcutaneously implanted CGM device (shortened to Eversense); and (iii) the glucose-oxidase CGM device (shortened to FSL). In the current study, FSLCstrip was considered as the reference for obtaining glucose levels because in prior research its capillary measurements were comparable with National Institute of Standards and Technology standards to the gold reference method [33].

To extend the analysis of the performance of each CGM device against FSLCstrip during the periods of PA and NDA, we studied the glucose measurements in four time intervals throughout the day:

- Morning (M_{int}) from 6:00 to 12:00 h;
- Afternoon (A_{int}) from 12:00 to 18:00 h;
- Evening (E_{int}) from 18:00 to 24:00 h;
- Night (N_{int}) from 0:00 to 6:00 h.

We also studied the glucose measurements in three glucose ranges:

- Hypoglycemia (<70 mg/dL);
- Euglycemia (in the range [70, 180] mg/dL);
- Hyperglycemia (>180 mg/dL).

2.2. Dataset Description and Preprocessing

The dataset was composed of 5166 glucose recordings belonging to 23 T1D subjects and from the devices FSL, Eversense, and FSLCstrip. The glucose measurements were acquired as follows. A total of 7 self-glucose measurements per day were obtained using the FSLCstrip, while interstitial glucose values using Eversense and FSL were taken within maximal 2 min of capillary measurement, ensuring the comparison of these technologies. To visualize the distribution of glucose concentrations for each period, PA and NDA, we show in Figure A1 of the Appendix A the box plots for each participant in the different periods. Note that the periods PA and NDA are represented in blue and orange, respectively, and using FSLCstrip (Figure A1a), Eversense (Figure A1b), and FSL (Figure A1c). In Figure A2, we show the time series associated with glucose concentration values for all subjects. Note that glucose values of each subject are depicted by a gray line and the mean value at each time is represented with a red dotted line. The green dotted line shows the limit for hyperglycemia (values above the line are in the hyperglycemic range), while the blue dotted line represents the limit for hypoglycemia (values below the line are in the hypoglycemic range).

2.3. Statistical Tests and Metrics to Measure Differences in Glucose Concentrations

To statistically measure the differences in glucose concentrations between PA and NDA periods, and to identify the time interval and glucose range more affected by high-intensity PA, several parametric and nonparametric tests were considered. Specifically, the following statistical tests were considered: (i) the Shapiro Wilk test was used to perform the normality test; (ii) the two-proportion z-test was used to determine whether two proportions from two independent samples are significantly different; (iii) the (parametric) two-sample t-test was used to compare and determine whether the difference between the means of two populations exists; and (iv) the (nonparametric) rank-based Mann–Whitney U test was used to identify differences between groups by considering their medians or means. The latter method is commonly used when sample sizes are small and for non-normally distributed data, being less sensitive to the non-homogeneity of the variance. For all these statistical tests, a level of significance $alpha = 0.05$ was considered.

To assess the differences between glucose concentrations obtained from CGM devices (Eversense and FSL) against the capillary glucose measurements (FSLCstrip), two approaches were considered: (i) the deviation metrics mean absolute deviation (MAD) and mean absolute relative deviation (MARD); and (ii) the proportion of glucose measurements in zone A of the Clarke error grid (CEG) plot [34]. Given $\mathbf{y} = [y_1, \ldots, y_N]$ as the vector representing the glucose measurements by FSLCstrip, and $\hat{\mathbf{y}} = [y_1, \ldots, y_N]$ as the vector indicating the glucose measurements of FSL and Eversense. MAD and MARD are defined as follows.

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

$$MARD = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| \tag{2}$$

where N is the total number of glucose measurements, y_i is the ith measure obtained by the reference device FSLCstrip, and $\hat{y}_i$ is the ith measure obtained using FSL or Eversense.

The CEG is a visual tool used to quantify the accuracy of measured glucose values compared to reference values, represented as a grid divided into five zones: A, B, C, D, and E. Figure 1 shows an example of a CEG plot with the different zones. Zone A contains values considered clinically accurate, indicating that the predicted glucose values would lead to the correct treatment of the patient. Zone B indicates altered clinical action with little or no effect on clinical outcome. The values in zone C indicate overcorrecting acceptable blood glucose levels and those in zone D suppose dangerous failures since they are predicted as being in an acceptable range when they are outside normoglycemia. Zone E values would

also lead to erroneous treatment since the predicted values are opposite to the accurate BG levels. Values in zones A and B are deemed clinically acceptable, while those in zones C, D, and E are considered potentially unsafe. Since it is the most representative and contains the largest amount of information, the analysis in this work is focused on studying the values that fall in zone A.

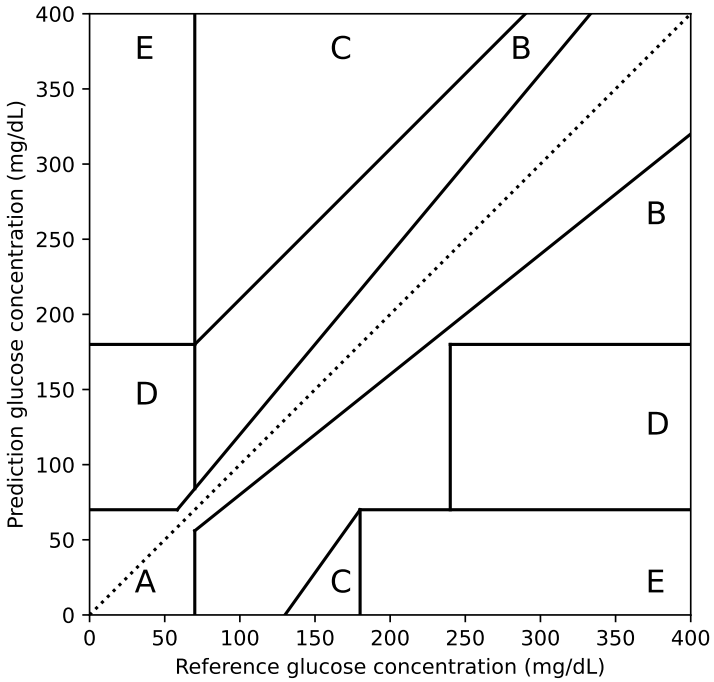


Figure 1. Example of a Clarke error grid plot.

2.4. Feature Selection Methods

Feature selection (FS) methods aim to select a subset of features from an original feature set, identifying the most relevant features and discarding the irrelevant ones [35]. FS becomes crucial in predictive tasks because decreasing the number of features could lead to maximizing generalization capacity and reaching better predictive results [36]. Additionally, FS may help in better data understanding and interpretability, which is crucial in certain areas, such as healthcare. FS methods are generally split into three categories [36]: filter, wrapper, and embedded methods.

Filter methods select features that present a strong relationship with the target and are performed independently of any predictive model. These methods are computationally efficient and fast because they do not involve a training stage of predictive models [36]. In this work, chi-squared (chi2) [37] and mutual information (MI) [37] were considered as filter methods.

Wrapper methods iteratively train a predictive model to evaluate and choose the feature subsets [38]. Depending on whether a forward or backward selection approach is followed, the features are added or discarded until the best subset is identified [39]. Since these methods use several predictive models to select the feature subsets, they can be computationally complex and time-consuming [38]. In this work, a backward selection approach was considered, where the ML algorithm is initially trained with the whole set of features to which one feature is discarded at each repetition until the algorithm is trained with the empty set of features (backward selection approach) [40]. For performing wrapper methods, a filter-based method using ML was employed to set the hierarchy of importance of the features and select which feature to discard at each iteration.

Embedded methods intrinsically perform feature selection during the ML algorithm training. We selected two tree-based embedded methods: decision tree (DT) and random forest (RF), where the importance is used to select the features. Also, the regularized

method least absolute shrinkage and selection operator (LASSO) [41] was included in the analysis.

2.5. Supervised Machine Learning Models and Performance Metrics

In this paper, we employed different linear and nonlinear supervised ML-based models for identifying if a patient has performed PA. In particular, we used four state-of-the-art models, including DT, logistic regression (LR), K-nearest neighbors (KNN), and support vector machine (SVM). We selected these ML models based on different factors, including ease of implementation, extensive use in benchmarking studies, and because they have been used and internationally validated in multiple clinical studies, reaching great performance. Additionally, aiming to improve the predictive performance, we considered two ensemble models RF and the extreme gradient boosting (XGB), the neural network-based model named multilayer perceptron (MLP), and a novel predictive transformer-based model for tabular data named TabPFN.

LR is a linear model used for classification that uses a logistic function that finds the best fitting coefficients to describe the relationship between the independent variables (input features) and the dependent variable (target/label vector) [42]. Once the LR model is trained, it can be used to predict the probability of a sample belonging to a positive class based on its input variables. A common threshold is set at 0.5, where probabilities above the threshold are classified as positive, and those below are classified as negative.

KNN is a nonlinear model that classifies an unlabeled sample according to the class that is most frequent within its K-nearest neighbors [43]. First, these neighbors are found by a similarity measure (e.g., Euclidean distance), and then the unlabeled sample is assigned the class that is most frequent within its K-nearest neighbors. KNN mainly depends on two hyperparameters: the similarity measure and the number of neighbors K. The former is used to measure how similar a sample is from others, being the Euclidean distance one of the most used. The latter is representative of the neighborhood size. Choosing an adequate value of K is highly related to predictive performance, with larger values generating more complex model decision boundaries and smaller values creating simpler ones [43].

SVM is a model that allows both linear and nonlinear approximations [44]. SVM aims to find an optimal hyperplane in a high-dimensional space (called feature space) that separates the samples into a discrete number of predefined classes. This hyperplane determines the margin between the classes, and when data are not linearly separable, different kernel functions (e.g., polynomial, radial basis, sigmoid) can be used to maximize margins between hyperplanes [45].

RF is a model that generates an ensemble of simpler models, particularly training multiple DTs [46]. Each DT is trained using a subsample of the training set, and then, a combination of the DTs are used to make the final predictions. This approach with different samples and models helps to reduce overfitting and improve generalization.

MLP is a type of artificial neural network and it consists of an input layer, various hidden layers, and an output layer, where each neuron in each layer is fully connected to neurons in the subsequent layers. MLP uses nonlinear activation functions to learn complex patterns and relationships from data, and the back-propagation algorithm is used to update the weights of the connections by aiming to minimize errors.

TabPFN is a transformer-based model specifically designed for supervised classification on tabular datasets [47]. It combines a transformer architecture called prior-data fitted network (PFN) [47] and Bayesian inference to solve classification tasks over tabular data. TabPFN does not require training from scratch on new data, and presents excellent classification results and low computational complexity.

XGB is an algorithm based on a gradient boosting tree that employs an ensemble approach to integrate several weak models and improve model predictions. It also involves a better regularization strategy that can effectively overcome overfitting and improve learning performance [48]. XGB offers scalability due to several algorithmic optimizations

and handling sparse data with new tree learning scheme. It is one of the most used models in benchmarking studies on tabular data.

To quantitatively evaluate the predictive performance, we used the following figures of merit: accuracy, sensitivity, specificity, and F1-score [49], which take into account how the model correctly or incorrectly predicts the positive class and negative class: true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). Note that the positive and negative class, in our case, correspond to 'PA' and 'NDA', respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{F1-score} = 2\frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{7}$$

2.6. Post Hoc Interpretability Methods

During the last decade, the use of ML models has exponentially grown in multiple areas, achieving great acceptance and popularity due to their remarkable results in supervised tasks [50]. Despite their benefits, most ML models are characterized by a lack of interpretability (known as black-box models). Interpretability can be defined as the ability of a human to understand the cause of a decision in computational models [51]. This is essential to be considered and adopted in clinical practice, and several methods have been proposed for extracting knowledge and gaining interpretability [52,53]. Many interpretability methods have been proposed, with the post hoc and model-agnostic approaches the most used [53]. These approaches that provide explanations through both inspection of learned features and feature importance could support the identification of risk factors for several diseases. In this work, the post hoc method called Shapley additive explanations (SHAP) [54] was considered.

SHAP is a post hoc interpretability method based on the aggregation of local interpretations and allows one to explain the general behavior of a model by analyzing the prediction of several samples [54]. Each feature value of a sample is treated as a player in a coalitional game, (i.e., a game where several players cooperate towards the same objective). In this case, the objective is the prediction for a particular sample. In this way, SHAP divides the final payout (the output of the model) between all players (the feature values of the instance to be analyzed). The Shapley value is computed as the mean of the marginal contribution of the player in all the possible coalitions.

3. Results

In this work, the experiments are divided into two main parts. In the first part, we analyze the accuracy of glucose measurements obtained from the devices Eversense and FSL against the capillary measurements obtained by FSLCstrip during the period of PA and NDA. The comparison was carried out considering (1) three glucose ranges (hypoglycemia, euglycemia, and hyperglycemia); and (2) four time intervals throughout the day (M_{int}, A_{int}, E_{int}, and N_{int}). In the second part, several ML-based models are used to identify whether a patient has performed PA or not using statistical information extracted from glucose measurements.

3.1. Analysis of the Impact of Physical Activity on CGM Devices' Performance

3.1.1. Analysis of the Impact of Physical Activity on CGM Devices' Performance in Different Glucose Ranges

Figure 2 shows the deviation metrics MAD and MARD for both Eversense and FSL devices by considering different glucose ranges (hypoglycemia, euglycemia, hyperglycemia) and for PA and NDA, respectively. Both MAD and MARD were calculated between pairs of glucose values recorded for each CGM device and capillary glucose measurements (FSLCstrip). In general, both devices showed significantly higher deviations in PA in all glucose ranges. However, Eversense in the hypoglycemic range seems to perform better in PA than NDA, showing smaller MAD and MARD values. For both Eversense and FSL, comparing the glucose ranges, MAD values are the highest in the hyperglycemic range, and for MARD, it occurs for hypoglycemia. It is worth noting that standard deviations (represented by a line in the center of bars) are also high in hyperglycemia (MAD) and hypoglycemia (MARD). This suggested that CGM devices are more sensitive to these glucose ranges, with less accurate glucose measurements compared to FSLCstrip.

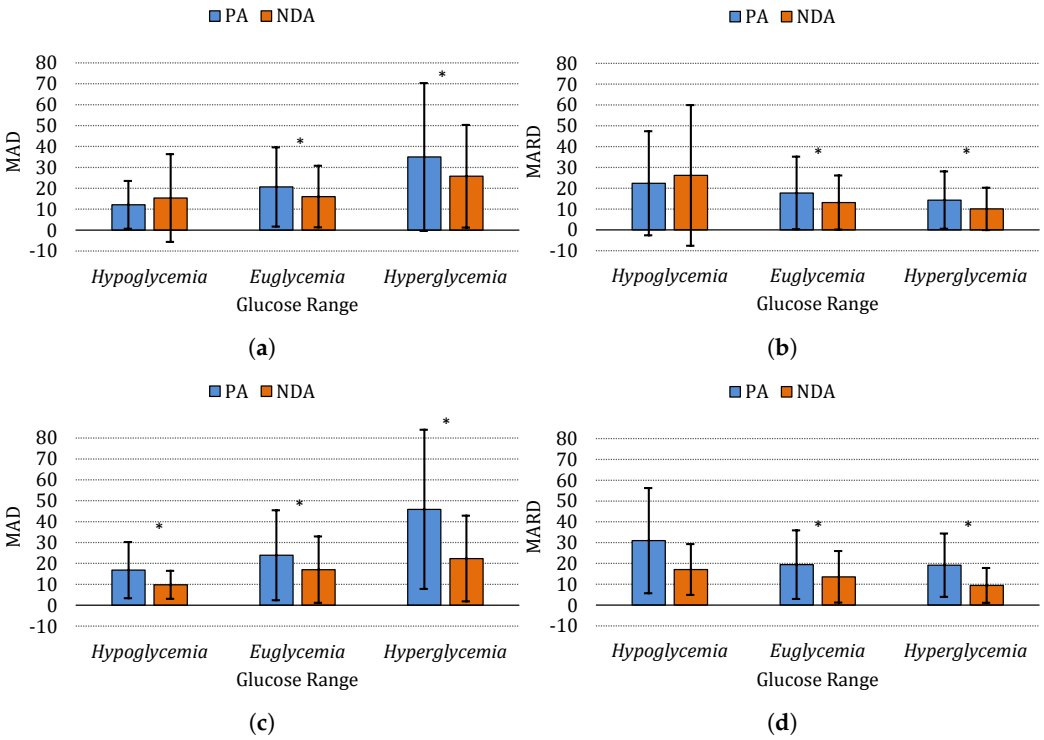


Figure 2. MAD and MARD values during PA and NDA for different glycemic ranges for (**a**,**b**) Eversense against FSLCstrip and (**c**,**d**) FSL against FSLCstrip. MAD and MARD were calculated based on the differences between glucose values recorded by the CGM device and FSLCstrip for PA and NDA, respectively. The comparison highlights whether there are significant differences between the PA and NDA conditions across glucose ranges. * indicates statistically significant differences in the impact of physical activity according to the two-proportion z-test (with level of significance of 0.05).

To evaluate the quality of the glucose measures, we also studied the CEG. The higher concentration of points placed in zone A of the CEG plot, the greater performance of CGM devices. Figure 3 shows the proportion of values in zone A of CEG for both CGM devices during NDA and PA when considering hypoglycemic, euglycemic, and hyperglycemic ranges. In Figure 3a, it is observed that Eversense presents a high proportion of measures in zone A (88.5%) during PA when considering the hypoglycemic range. However, the proportions of points in zone A for the PA period are lower than the NDA in the other ranges: 66.4% versus 79.6% (euglycemic range) and 75.9% versus 86.2% (hyperglycemic range). According to the two-proportion test, all differences are statistically significant (represented by *) except for the hypoglycemic range. For FSL (see Figure 3b), the scenario is

different. The proportion of points in zone A for NDA is greater than those obtained in the PA period in all glucose ranges. The differences in proportions are statistically significant in all cases by considering the two-proportion test. Note that the highest proportion in zone A occurs in the hyperglycemic range (92%). In Eversense, the differences in proportions between PA and NDA are smaller compared to values of FSL. This suggests that glucose measurements from Eversense are more accurate, performing better than FSL.

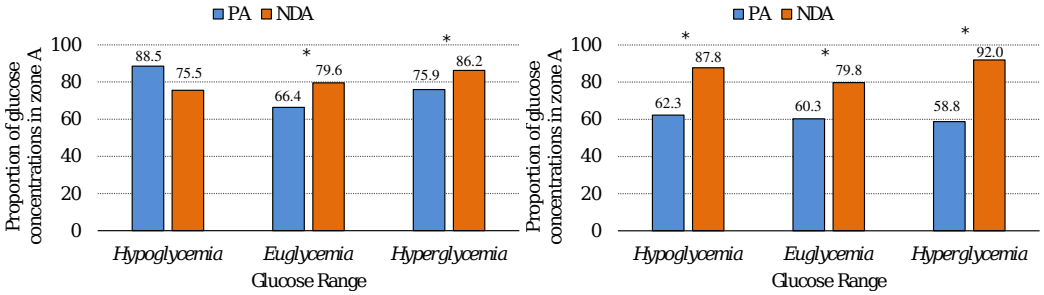


Figure 3. Proportion of glucose concentrations in zone A of CEG by considering different glucose ranges (hypoglycemia, euglycemia, hyperglycemia) and PA and NDA for (**a**) Eversense and (**b**) FSL. * indicates statistically significant differences in the impact of physical activity (PA versus NDA) according to the two-proportion z-test (with level of significance of 0.05).

3.1.2. Analysis of the Impact of Physical Activity on CGM Devices' Performance at Different Time Intervals Throughout the Day

To study the performance of CGM devices (Eversense and FSL) during the PA and NDA in different time intervals (M_{int}, A_{int}, E_{int}, N_{int}), the deviation metrics and the CEG were considered. Figure 4 shows the MAD and MARD values for different time intervals by distinguishing between the periods PA and NDA. In Figure 4a,b, it is observed that there exist differences between PA and NDA for Eversense, with greater MAD and MARD values during the PA period, except for N_{int}. The MAD and MARD values differ significantly during M_{int} and A_{int} for both Eversense and FSL in PA and NDA (see Figure 4a–d). FSL also presented significant differences in the case of E_{int} for both MAD and MARD (see Figure 4a–d), while Eversense only presented differences during evening for MARD.

The CEG was also analyzed for every time interval independently. Figure 5 depicts the proportion of glucose concentration values falling in zone A for each glucose monitoring device and period (PA in blue and NDA in orange). For Eversense (see Figure 5a), the proportion of points in zone A is greater for the NDA period in all time intervals during the day. The two-proportion test suggests that these differences are statistically significant in all cases except for N_{int}. In FSL (Figure 5b), the proportions of points in zone A in the NDA are greater than values associated with the PA period in M_{int}, A_{int}, and E_{int}, being statistically significant for M_{int} and A_{int} (following the two-proportion test results). Note that the highest values are obtained in the NDA, for Eversense, during M_{int} (86.0%) and N_{int} (84.1%), while in FSL, it occurs in the intervals E_{int} (84.5%) and A_{int} (85.7%).

3.2. Prediction of Physical Activity Through Machine Learning Models

In this subsection, we present the results obtained using several ML-based models for identifying when a T1D patient has performed high-intensity PA. In particular, DT, KNN, LR, MLP, RF, SVM, TabPFN, and XGB were considered in this study. The source code for the reproducibility of results is available at https://github.com/cdchushig/exercise-cgm (accessed on 21 October 2024).

Before applying ML models, information from glucose time series was extracted using a feature extraction approach based on statistics and the number of hypoglycemic and hyperglycemic episodes. As stated, glucose measurements were represented by time series; however, since these data were scarce and irregularly sampled, we performed a feature extraction process based on statistics. In particular, four statistics were used, including

sum, median, variance, and entropy. We also considered the number of hypoglycemic and hyperglycemic episodes and the total number of adverse events (sum of hypoglycemic and hyperglycemic episodes). A brief description of these features is shown in Table 1.

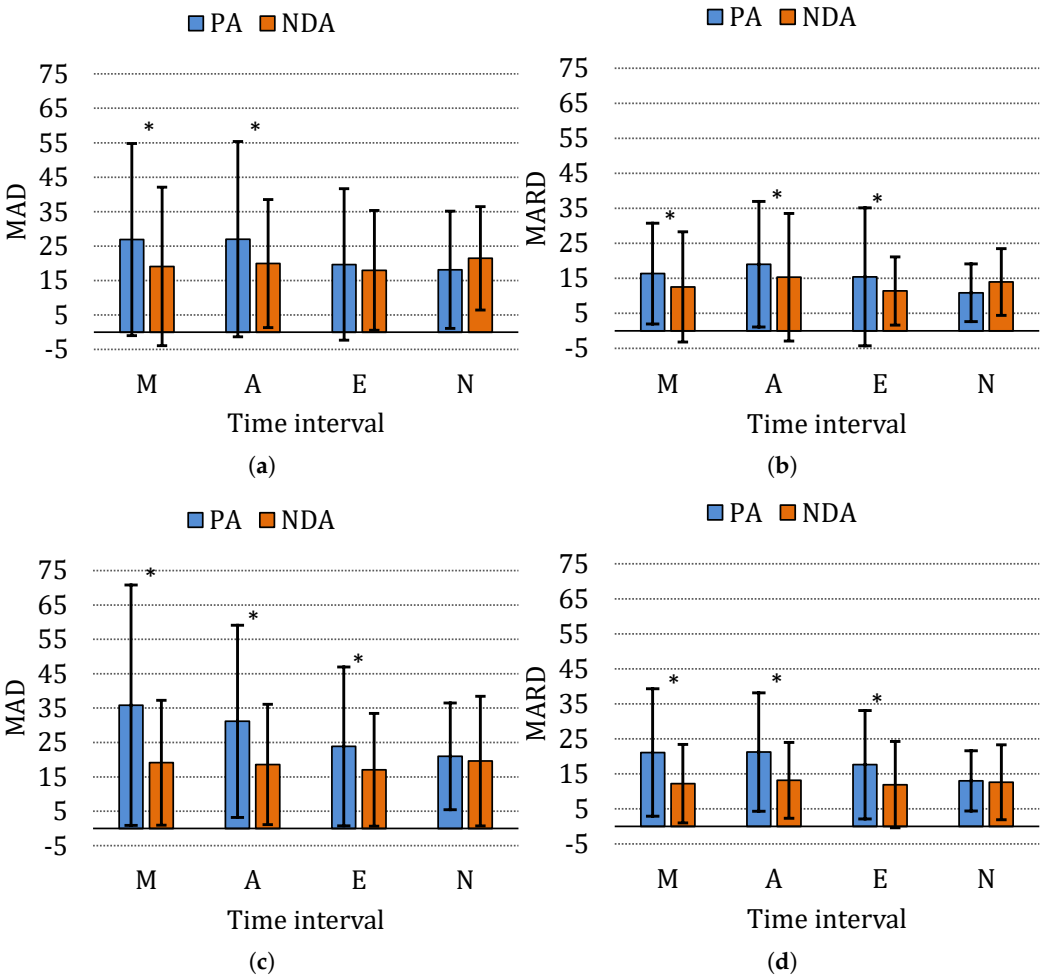


Figure 4. MAD and MARD values during PA and NDA for different time intervals for (**a**,**b**) Eversense against FSLCstrip and (**c**,**d**) FSL against FSLCstrip. MAD and MARD were calculated based on the differences between glucose values recorded by the CGM device and FSLCstrip for PA and NDA, respectively. The comparison highlights whether there are significant differences between the PA and NDA conditions across glucose ranges. * indicates statistically significant differences in the impact of physical activity according to the two-proportion z-test (with level of significance of 0.05).

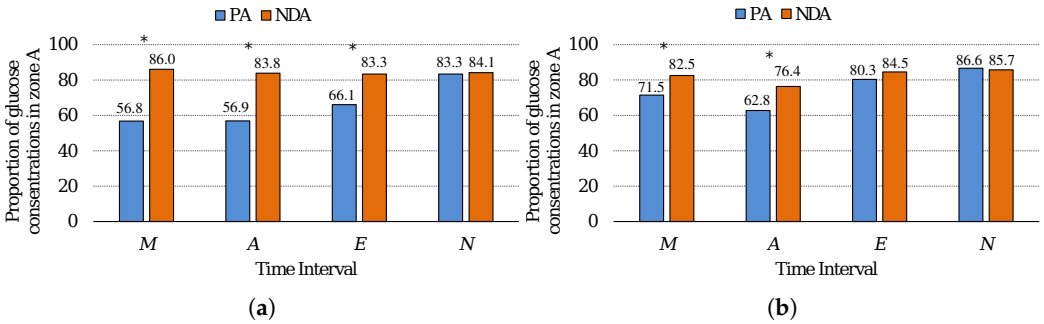


Figure 5. Proportion of glucose concentrations in zone A of CEG by considering different time intervals (M_{int}, A_{int}, E_{int}, N_{int}) and the periods PA and NDA for (**a**) Eversense and (**b**) FSL. * indicates statistically significant differences in the impact of physical activity (PA versus NDA) according to the two-proportion test (with level of significance of 0.05).

Table 1. Description of the variables extracted from glucose measurements.

Feature	Identifier	Description
Sum	Sum	Sum of the observed glucose concentration values.
Median	Median	Value lying at the midpoint of the frequency distribution of observed glucose concentration values.
Variance	Var	Measure of dispersion of the observed glucose concentration values.
Entropy	Entropy	Measure of amount of uncertainty within the observed glucose values.
Hyperglycemic events	Hyper	Total number of observed glucose concentration values within the hyperglycemic range.
Hypoglycemic events	Hypo	Total number of observed glucose concentrations values within the hypoglycemic range.
Adverse events	Adverse	Sum of hypoglycemic and hyperglycemic events.

The extraction process results in a total of 35 features, 7 features for each time interval (M_{int}, A_{int}, E_{int}, N_{int}), and 7 considering glucose data during all day. We also conducted a correlation analysis using the Pearson correlation coefficient (PCC), aiming to discard those features highly correlated (with a PCC over 0.7). As a result, 15 features were kept for FSLCstrip, 16 for Eversense, and 15 for FSL. Then, we used several filter, wrapper, and embedded FS methods to select the most important features. Given that several FS methods were considered, a voting strategy was conducted for selecting the most representative features, keeping those features chosen by at least three of the eight FS methods. Figure 6 shows the voting results of the eight FS methods, which indicates the frequency of selection of each feature for different BG monitoring devices studied in the paper. A total of 11 features were selected for FSLCstrip (F_Adverse, N_Median, N_Sum, N_Var, E_Median, A_Var, E_Hyper, N_Hypo, M_Entropy, M_Hypo, E_Hypo), 14 for Eversense (F_Adverse, N_Median, N_Var, N_Adverse, M_Var, M_Entropy, E_SumValues, E_Var, F_SumValues, A_Var, M_Hypo, E_Hypo, M_Median, E_Median), and 10 for FSL (M_SumValues, A_Var, F_Adverse, M_Var, N_Var, A_Hypo, N_Hyper, N_Median, E_Hypo, N_Entropy). As shown in Figure 6, F_Adverse, N_Median, N_Var, A_Var, and E_Hypo are the features most voted for. F_Adverse indicates the presence of a glycemic event by considering all glucose values, while E_hypo denotes the episodes of hypoglycemia during the evening (E_{int}). N_Median, N_Var, and A_Var are statistics from the time series that point out the variance and median of glucose values in the interval N_{int} and A_{int}.

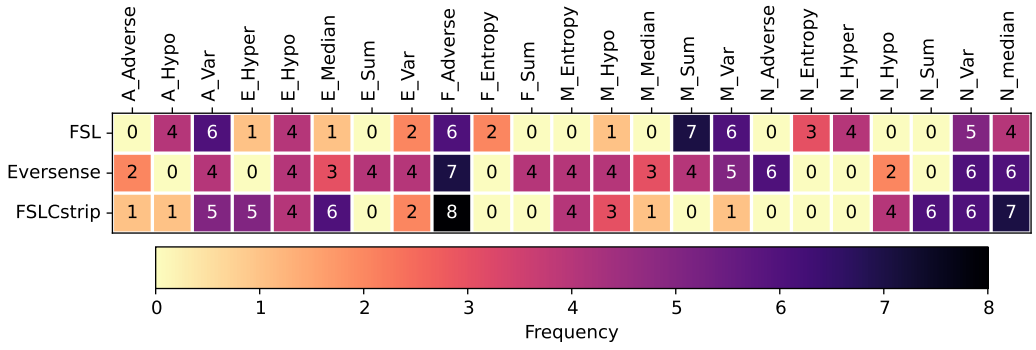


Figure 6. Voting results of eight FS methods by indicating the frequency of selection of each feature for each BG monitoring device.

Data extracted from time series were split into training and test subsets, with 70% and 30% of samples, respectively. The train subset was used for training the ML methods,

whereas the test subset was only used to assess the performance of the trained models. Both subsets were balanced to ensure the same number of samples for each class (PA and NDA). To avoid bias, we used five independent partitions, each of which has a training and test subset. To find the best hyperparameters, we considered a leave-one-out cross-validation approach, a special case of k-fold cross-validation, in which the number of folds equals the number of samples. This method is adopted to obtain reliable predictive results when the number of samples of a dataset is scarce. Several hyperparameters were explored for the supervised models, which are shown in Table 2.

Table 2. Hyperparameters explored for the supervised ML-based models DT, KNN, LR, MLP, RF, SVM, TabPFN, and XGB.

Model	Hyperparameter	Values/Options
DT	maximum depth	$[3, 20]$
	split criterion	Gini, entropy
	minimum samples per split	$[2, 8]$
KNN	number of neighbours	$[1, 20]$
LR	penalty	L2
	C	$C = \{0.01, 0.001, 0.1, 1, 10\}$
MLP	hidden layer sizes	$\{(\text{num\_features}, 4), (\text{num\_features}, 3), (\text{num\_features},)\}$
	activation function	Tanh, RELU
	optimizer	SGD, Adam
	alpha	$\{0.0001, 0.05\}$
	learning rate	constant, adaptive
RF	# estimators	$\{10, 20, 30, 40\}$
	maximum depth	$[1, 16]$
SVM	kernel	RBF
	γ	$\{0.01, 0.001, 0.0001, 0.00001\}$
	C	$\{0.01, 0.001, 0.0001, 0.00001\}$
TabPFN	number of ensemble	$\{16, 32\}$
	batch size	$\{20, 30\}$
XGB	# estimators	$\{10, 20, 30, 40\}$
	maximum depth	$[1, 16]$
	learning rate	$\{0.1, 0.01, 0.001\}$
	subsample	$\{0.5, 0.7, 1.0\}$

To study the impact of PA during the day, we consider the four time intervals defined in the previous section (M_{int}, A_{int}, E_{int}, and N_{int}). In this way, the features mentioned previously are extracted independently for the whole glucose values F_{int} as well as for each of the four time intervals. In the remainder of this paper, the feature names will include a prefix to denote what time interval each mentioned feature represents. For instance, A_Var refers to the variance in A_{int}. By summarizing, the extraction process results in a total of 35 features (7 features for each time interval and 7 considering the full day).

The classification results using the three glucose monitoring devices are presented in Table 3, showing the mean and standard deviation obtained over five test partitions. As shown, the highest predictive results are achieved using the ML models KNN, LR, SVM, TabPFN, and XGB, reaching accuracy mean values over 0.88. The best predictive results are reached using Eversense, with LR and XGB obtaining an F1-score of 0.91 ± 0.05 and 0.92 ± 0.04, respectively. Although it was seen in the previous section that the deviation metrics by Eversense and FSL are higher in PA than NDA, these differences could potentially aid in the identification of high-intensity PA.

Table 3. Mean and standard deviation of classification metrics when considering test subset partitions and the selected features for FSLCstrip, Eversense, and FSL. The best results for each figure of merit are in bold.

ML Model	Glucose Monitoring Device	Accuracy	Sensitivity	Specificity	F1-Score
DT	FSLCstrip	0.74 ± 0.11	0.73 ± 0.20	0.76 ± 0.26	0.73 ± 0.12
DT	Eversense	0.73 ± 0.16	0.70 ± 0.25	0.76 ± 0.16	0.70 ± 0.20
DT	FSL	0.65 ± 0.17	0.58 ± 0.17	0.73 ± 0.31	0.63 ± 0.16
KNN	FSLCstrip	0.87 ± 0.05	0.87 ± 0.12	0.88 ± 0.16	0.87 ± 0.05
KNN	Eversense	0.90 ± 0.08	$\mathbf{0.93 \pm 0.13}$	0.88 ± 0.16	0.91 ± 0.08
KNN	FSL	0.85 ± 0.07	0.87 ± 0.06	0.82 ± 0.16	0.85 ± 0.06
LR	FSLCstrip	0.83 ± 0.06	0.81 ± 0.11	0.85 ± 0.22	0.83 ± 0.04
LR	Eversense	$\mathbf{0.91 \pm 0.06}$	0.90 ± 0.07	0.91 ± 0.17	0.91 ± 0.05
LR	FSL	0.82 ± 0.16	0.82 ± 0.27	0.82 ± 0.22	0.80 ± 0.21
MLP	FSLCstrip	0.83 ± 0.11	0.81 ± 0.11	0.85 ± 0.22	0.83 ± 0.09
MLP	Eversense	0.86 ± 0.10	0.85 ± 0.15	0.88 ± 0.16	0.86 ± 0.10
MLP	FSL	0.80 ± 0.12	0.76 ± 0.19	0.85 ± 0.22	0.79 ± 0.13
RF	FSLCstrip	0.83 ± 0.12	0.76 ± 0.21	0.91 ± 0.06	0.80 ± 0.15
RF	Eversense	0.86 ± 0.09	0.82 ± 0.20	0.91 ± 0.11	0.84 ± 0.13
RF	FSL	0.78 ± 0.14	0.70 ± 0.25	0.85 ± 0.22	0.74 ± 0.19
SVM	FSLCstrip	0.83 ± 0.06	0.79 ± 0.19	0.88 ± 0.16	0.82 ± 0.08
SVM	Eversense	0.88 ± 0.08	0.85 ± 0.15	0.91 ± 0.17	0.87 ± 0.09
SVM	FSL	0.82 ± 0.12	0.79 ± 0.19	0.85 ± 0.22	0.81 ± 0.13
TabPFN	FSLCstrip	0.82 ± 0.06	0.81 ± 0.14	0.82 ± 0.20	0.82 ± 0.05
TabPFN	Eversense	0.91 ± 0.10	0.91 ± 0.17	$\mathbf{0.91 \pm 0.17}$	0.91 ± 0.11
TabPFN	FSL	0.77 ± 0.11	0.73 ± 0.24	0.82 ± 0.20	0.74 ± 0.16
XGB	FSLCstrip	0.79 ± 0.07	0.76 ± 0.19	0.82 ± 0.20	0.77 ± 0.08
XGB	Eversense	0.92 ± 0.04	0.90 ± 0.07	0.93 ± 0.07	$\mathbf{0.92 \pm 0.04}$
XGB	FSL	0.77 ± 0.05	0.70 ± 0.20	0.84 ± 0.13	0.74 ± 0.09

With the goal of providing interpretability to ML-based models, SHAP was used alongside the models built with the selected features obtained by the FS methods. The SHAP analysis is commonly used for gaining interpretability of the importance value assigned to each feature. This leads to identifying how they contribute to the prediction model. Figure 7 shows the SHAP bar plots associated with the ML models with best predictive results for each device (i.e., SVM for FSLCstrip, and LR for both Eversense and FSL). These plots show the order of relevance of the features for the models' predictions in decreasing order (vertical axis). It is remarkable to see that the total number of adverse events (both hypoglycemic and hyperglycemic) during the day (F_Adverse) has the most important influence on all models' decisions, receiving the maximum importance value in all cases. The rest of the features appear to have very different impacts on every model.

Since the Eversense device provided the best predictive results, we used glucose measures from this device and conducted an analysis at different times of the day. The ML-based models were trained using the most relevant features associated with each time interval during the day (M_{int}, A_{int}, E_{int}, N_{int}) and F_{int}. Following a voting FS strategy, we selected the most relevant features for each time interval:

- M_{int}: M_Sum, M_Var, M_Adverse, M_Hypo, and M_Median;
- A_{int}: A_Var, A_Entropy, A_Hyper, A_Hypo and A_Median;
- E_{int}: E_Entropy, E_Median, E_Hyper, E_Var and E_Hypo;
- N_{int}: N_Hypo, N_Hyper, N_Var, N_Median and N_Entropy;
- F_{int}: F_Adverse, N_Median, N_SumValues, N_Var, E_Median, A_Var, E_Hyper, N_Hypo, M_Entropy, M_Hypo, E_Hypo.

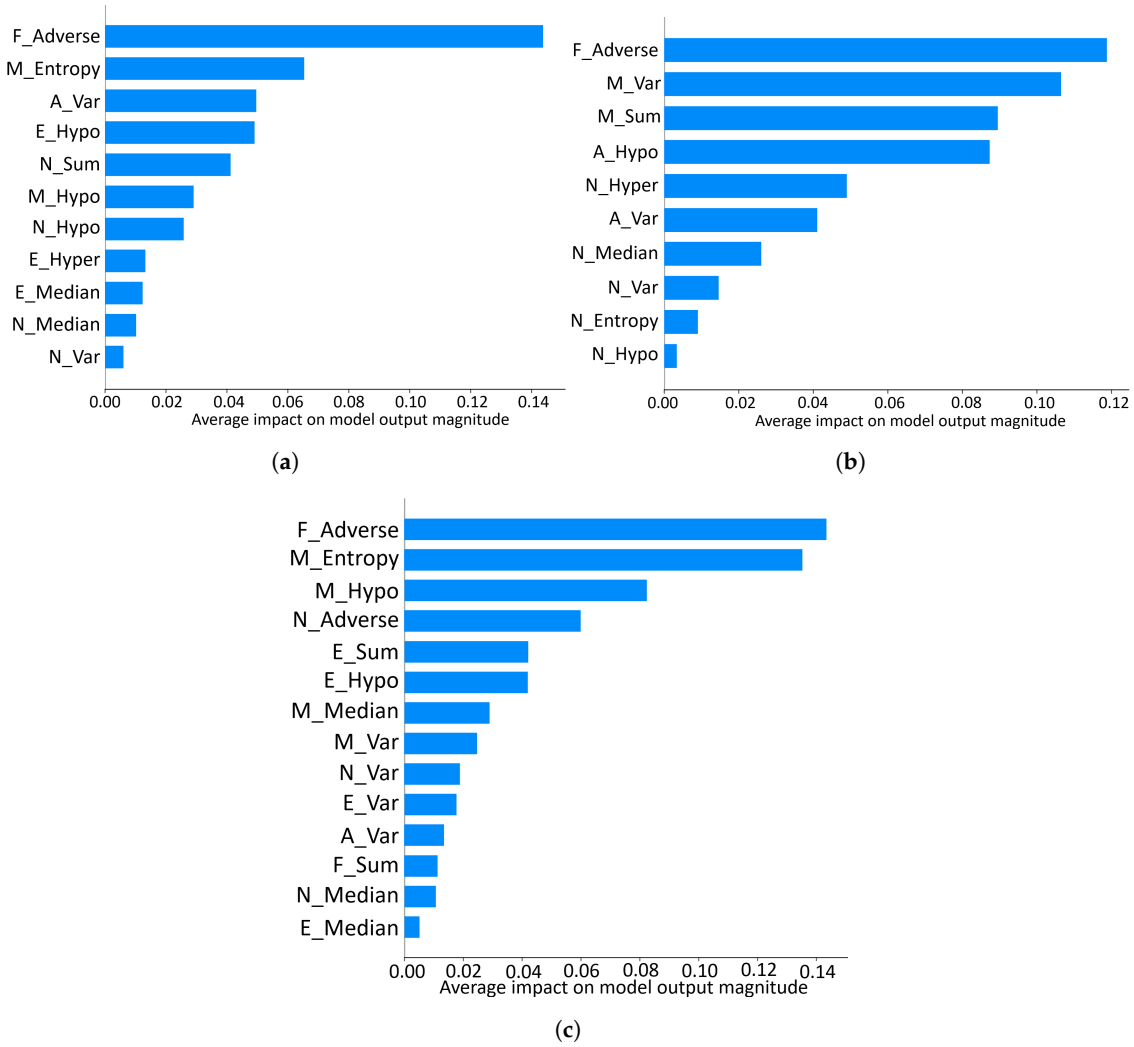


Figure 7. SHAP bar plots showing the importance of each feature for predicting PA and NDA, using (**a**) FSLCstrip and SVM; (**b**) Eversense and LR; and (**c**) FSL and LR.

Table 4 shows the mean and standard deviation of the classification metrics for the five test subset partitions by using only the features relative to each 6-h time interval. As shown, there is a substantial decrease in the predictive performance compared to the results obtained considering features from all time periods (see values in Table 3). This is remarkable in the cases of the afternoon, evening, and night (A_{int}, E_{int}, N_{int}), where accuracy values are the lowest, in some cases below 0.52, indicating that the models randomly assign classes to the samples. Furthermore, the better time interval to predict if an individual performed high-intensity PA was (M_{int}).

Table 4. Mean and standard deviation of the figures of merit were calculated across 5 test partitions, considering Eversense measurements at different time intervals. The best results for each figure of merit are highlighted in bold.

ML Method	Interval	Accuracy	Sensitivity	Specificity	F1-Score
DT	F_{int}	0.87 ± 0.07	0.90 ± 0.07	0.84 ± 0.09	0.88 ± 0.07
	M_{int}	0.79 ± 0.13	0.73 ± 0.24	0.84 ± 0.15	0.75 ± 0.18
	A_{int}	0.52 ± 0.07	0.59 ± 0.14	0.45 ± 0.15	0.55 ± 0.09
	E_{int}	0.59 ± 0.02	0.58 ± 0.13	0.59 ± 0.14	0.57 ± 0.05
	N_{int}	0.62 ± 0.07	0.45 ± 0.19	0.79 ± 0.11	0.51 ± 0.16

Table 4. *Cont.*

ML Method	Interval	Accuracy	Sensitivity	Specificity	F1-Score
KNN	F_{int}	0.80 ± 0.08	0.85 ± 0.15	0.76 ± 0.10	0.80 ± 0.09
	M_{int}	0.75 ± 0.15	0.70 ± 0.19	0.79 ± 0.26	0.73 ± 0.14
	A_{int}	0.63 ± 0.12	0.63 ± 0.11	0.63 ± 0.18	0.63 ± 0.11
	E_{int}	0.67 ± 0.13	0.82 ± 0.13	0.52 ± 0.14	0.71 ± 0.11
	N_{int}	0.65 ± 0.13	0.57 ± 0.22	0.73 ± 0.13	0.60 ± 0.18
LR	F_{int}	0.88 ± 0.05	0.88 ± 0.10	0.88 ± 0.10	0.87 ± 0.05
	M_{int}	0.73 ± 0.11	0.61 ± 0.24	0.85 ± 0.15	0.67 ± 0.17
	A_{int}	0.58 ± 0.05	0.60 ± 0.14	0.56 ± 0.19	0.59 ± 0.05
	E_{int}	0.65 ± 0.08	0.72 ± 0.10	0.58 ± 0.22	0.67 ± 0.05
	N_{int}	0.65 ± 0.07	0.60 ± 0.24	0.70 ± 0.15	0.60 ± 0.15
MLP	F_{int}	0.85 ± 0.10	0.85 ± 0.22	0.85 ± 0.09	0.83 ± 0.14
	M_{int}	0.63 ± 0.03	0.48 ± 0.17	0.79 ± 0.17	0.55 ± 0.09
	A_{int}	0.57 ± 0.06	0.69 ± 0.08	0.44 ± 0.18	0.62 ± 0.03
	E_{int}	0.65 ± 0.07	0.72 ± 0.10	0.58 ± 0.15	0.67 ± 0.06
	N_{int}	0.62 ± 0.03	0.50 ± 0.20	0.73 ± 0.16	0.54 ± 0.14
RF	F_{int}	0.90 ± 0.05	0.84 ± 0.10	**0.97 ± 0.05**	0.90 ± 0.06
	M_{int}	0.74 ± 0.08	0.64 ± 0.18	0.85 ± 0.15	0.70 ± 0.14
	A_{int}	0.54 ± 0.03	0.66 ± 0.11	0.42 ± 0.10	0.59 ± 0.05
	E_{int}	0.64 ± 0.09	0.60 ± 0.10	0.67 ± 0.20	0.62 ± 0.07
	N_{int}	0.63 ± 0.10	0.57 ± 0.19	0.70 ± 0.11	0.59 ± 0.14
SVM	F_{int}	0.82 ± 0.06	0.79 ± 0.15	0.85 ± 0.09	0.80 ± 0.08
	M_{int}	0.75 ± 0.10	0.64 ± 0.24	0.85 ± 0.15	0.69 ± 0.17
	A_{int}	0.60 ± 0.03	0.60 ± 0.14	0.59 ± 0.19	0.59 ± 0.04
	E_{int}	0.65 ± 0.08	0.79 ± 0.06	0.51 ± 0.20	0.69 ± 0.04
	N_{int}	0.69 ± 0.05	0.60 ± 0.11	0.79 ± 0.11	0.66 ± 0.08
TabPFN	F_{int}	0.91 ± 0.08	**0.91 ± 0.11**	0.91 ± 0.11	0.91 ± 0.08
	M_{int}	0.73 ± 0.11	0.58 ± 0.19	0.88 ± 0.16	0.67 ± 0.16
	A_{int}	0.57 ± 0.04	0.60 ± 0.16	0.53 ± 0.17	0.57 ± 0.09
	E_{int}	0.62 ± 0.04	0.69 ± 0.12	0.54 ± 0.15	0.64 ± 0.04
	N_{int}	0.61 ± 0.05	0.69 ± 0.18	0.54 ± 0.16	0.63 ± 0.10
XGB	F_{int}	**0.92 ± 0.04**	0.90 ± 0.07	0.93 ± 0.07	**0.92 ± 0.04**
	M_{int}	0.77 ± 0.07	0.72 ± 0.10	0.82 ± 0.13	0.76 ± 0.08
	A_{int}	0.66 ± 0.12	0.74 ± 0.22	0.57 ± 0.16	0.67 ± 0.15
	E_{int}	0.66 ± 0.06	0.58 ± 0.15	0.75 ± 0.14	0.62 ± 0.10
	N_{int}	0.63 ± 0.03	0.57 ± 0.11	0.70 ± 0.17	0.60 ± 0.03

4. Discussion

In this work, we studied the performance of CGM devices (Eversense and FSL) compared to the capillary glucose device (FSLCstrip) during periods of high-intensity PA and NDA, when considering (i) three different glucose ranges (hypoglycemia, euglycemia, and hyperglycemia), and (ii) four time periods throughout the day (morning, afternoon, evening, and night).

CGM devices have increased their popularity for tracking and controlling glycemic levels because they are less invasive compared to capillary BG measurements that require frequent punctures for obtaining blood samples. Despite these characteristics, in several studies CGM devices have been shown to be less accurate than traditional capillary glucose devices. It is for this reason that the evaluation of CGM performance against gold standard measures becomes crucial to guarantee reliability and confidence in their use, and more particularly during PA.

The comparison of the performance of Eversense and FSL against FSLCstrip was performed following two different approaches. Firstly, we analyzed MAD and MARD with respect to the FSLCStrip. Secondly, we visually analyzed how glucose concentrations values are distributed in different zones of CEG plots. We can conclude that these CGM devices present a clinically meaningful deviation compared to FSLCstrip based on MAD/MARD and on the proportion of

values in zone A, both providing a worse performance whenever the subjects perform PA. This is an important insight to consider, since diabetic patients need strict control of their glucose levels and performing strenuous PA can lead to less accurate glucose measurements when using such type of sensors. Also, when comparing the values of both periods (NDA and PA), it was observed that the distributions of glucose values were significantly different for Eversense and FSL (both in the total and the hyperglycemic ranges of glucose), whereas this was not the case for FSLCstrip. These experimental results prove that FSL and Eversense are less accurate for measuring glucose under high-intensity PA scenarios which is in line with previous studies. Moreover, the fact that the differences appear in the hyperglycemic range is remarkable since it implies that the sensors work inappropriately when trying to detect glycemic events. This can be critical for T1D subjects because prolonged periods of hyperglycemia are one of the underlying causes of complications related to diabetes. The rapid change in glucose levels during PA continues to be a challenging situation and it is also expected to affect the performance of sensor systems.

To extend this work, we studied glucose levels during four different time intervals (M_{int}, A_{int}, E_{int}, N_{int}) throughout the day. No significant differences were observed among the distributions of each time interval in the NDA period, with stable glucose concentration values during the day. The opposite occurred during the PA period, obtaining significant differences between morning and afternoon, as well as between morning and evening. In a similar way to this previous analysis, the frequency of onset of glucose outside the target range (between 70 and 180 mg/dL) was studied, in particular for hypoglycemic and hyperglycemic episodes. Hypoglycemic events appeared to be stable at both periods (PA and NDA). However, hyperglycemic events vary significantly at different time intervals during the PA period compared to the NDA period. This makes sense considering that it is known that performing exercise diminishes the glucose storage of the body, therefore avoiding hyperglycemia. Although CGM devices play an important role in the management of T1D, we confirmed (extending the analysis of the study [32]) that accuracy of these devices is reduced during PA at different time intervals and glucose ranges.

The literature has shown that CGM devices support the reduction in hypoglycemia and hyperglycemia, improving the management of T1D. However, most of these studies were not conducted during high-intensity activities. In a similar way to this previous analysis, the frequency of onset of glucose outside the target range, in particular hypoglycemic/hyperglycemic episodes, was studied at different time intervals. Hypoglycemic events appeared to be stable at both periods (PA and NDA). However, hyperglycemic events vary significantly at different time intervals during the PA period compared to the NDA period. This makes sense considering that it is known that performing exercise diminishes the glucose storage of the body, therefore avoiding hyperglycemia. Although CGM devices play an important role in the management of T1D, we confirmed that accuracy of these devices are reduced during PA. The reviewed literature has shown that CGM devices support reducing hypoglycemia and hyperglycemia. However, most of these studies were not conducted during high-intensity activities.

The use of ML models to determine when an individual has performed PA yielded promising results with an accuracy value of 0.93. The predictive models developed using supervised ML techniques demonstrate high accuracy in distinguishing between PA and NDA, with Eversense showing the best performance. The SHAP analysis provides insights into the importance of different features in the predictive models. It highlights the significant influence of the total number of adverse events (F_Adverse) on the model's predictions, suggesting that the occurrence of glycemic events is a key indicator of PA. This finding is consistent across all devices studied. These predictive results can be used to develop an individualized data-driven classifier for each patient that categorizes glucose profiles based on the time interval during the day and according to if a patient performs PA.

Despite the extensive comparative analysis between Eversense and FSL devices against FSLCstrip, it is important to address the limitations of this study. One of the major limitations is related to the dataset size, where only glucose measurements from 23 T1D participants were considered. We acknowledge that the small sample size may affect

the generalization of our results since the glycemic variability of individuals may impact both the statistical analyses and extracted features used in the classification of PA. This could limit the broader applicability of our findings to the larger T1D population and other scenarios with PA since our work has been evaluated with glucose measurements in high-intensity PA. In this line, we also have to mention the potential bias in participant selection, and how the specific demographic, clinical, and glycemic characteristics of the participants might limit the extent to which these results can be applied to other studies or cohorts. However, it is worth noting that studies that collect data from individuals performing high-intensity PA are scarce, and our study is one of the few that has conducted an analysis on this, and more particularly using data from mountain biking.

To overcome these limitations, future work will focus on expanding the analysis by incorporating larger and more diverse datasets related to glucose, thereby enhancing the reliability of the findings. Additionally, to mitigate potential selection bias, future studies will include not only high-intensity PA but also various forms of aerobic and anaerobic exercises. This approach will allow for a comprehensive evaluation of glucose monitoring devices across a broader range of PA scenarios. Another valuable direction for future research involves the evaluation of other measures to capture the glycemic dynamics, and fusing them with those proposed in this study. Finally, to identify the most relevant features for classification, we combined several types of FS methods with reasonable results, but recent studies have shown that ensemble learning techniques might improve the FS, and this is a potential line of research for extending our study.

5. Conclusions

Our study demonstrated that CGM devices, particularly Eversense, can effectively monitor glucose levels during high-intensity PA, albeit with some limitations in accuracy. Our findings revealed significant differences in glucose measurements between these periods, with a noticeable impact on the performance of CGM devices. The results indicate that both Eversense and FSL show higher deviations in glucose values during PA, particularly in hyperglycemic and hypoglycemic ranges. This suggests that strenuous PA affects the accuracy of CGM devices, making them less reliable compared to capillary measurements. Under scenarios with high-intensity PA, the use of CGM devices should be cautious since glucose measures could not be sufficiently accurate. In addition, we developed an automated approach using ML techniques yielding promising results and achieving an accuracy value of 0.93. This approach could help to reduce the workload of clinicians by tracking PA in T1D patients and controlling medical recommendations. The findings underscore the importance of considering the impact of PA on glucose monitoring and the potential of ML methods to improve diabetes management.

Author Contributions: Conceptualization, D.C.-M. and C.S.-R.; methodology, D.C.-M. and C.S.-R.; software, H.C.-D. and D.C.-M.; validation, E.Å., H.F. and C.S.-R.; formal analysis, D.C.-M. and C.S.-R.; investigation, H.C.-D. and D.C.-M.; resources, C.S.-R.; data curation, P.R.v.D. and H.C.-D.; writing—original draft preparation, H.C.-D. and D.C.-M.; writing—review and editing, C.S.-R., E.Å., H.F., P.R.v.D. and C.S.-R.; visualization, H.C.-D. and D.C.-M.; supervision, C.S.-R.; project administration, C.S.-R.; funding acquisition, C.S.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Commission through the H2020-EU.3.1.4.2, European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under the Grant Agreement 101017385; by the Spanish Government by the Grant AAVis-BMR PID2019-107768RA-I00/AEI/10.13039/50110 0011033; by the Community of Madrid with the European Social Fund through the Youth Employment Operational Program and the Youth Employment Initiative (YEI) under the grant TIC-11649; and by Rey Juan Carlos University (grant 2023/SOLCON-132212).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Medical Ethical committees in Spain (Hospital Universitario Central de Asturias; 163/18) and the Netherlands (Isala Hospital; NL66388.075.18/180603) and registered

in the Dutch trial register (https://onderzoekmetmensen.nl/en/trial/45979, accessed on 8 August 2018, number NL7133).

Informed Consent Statement: Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

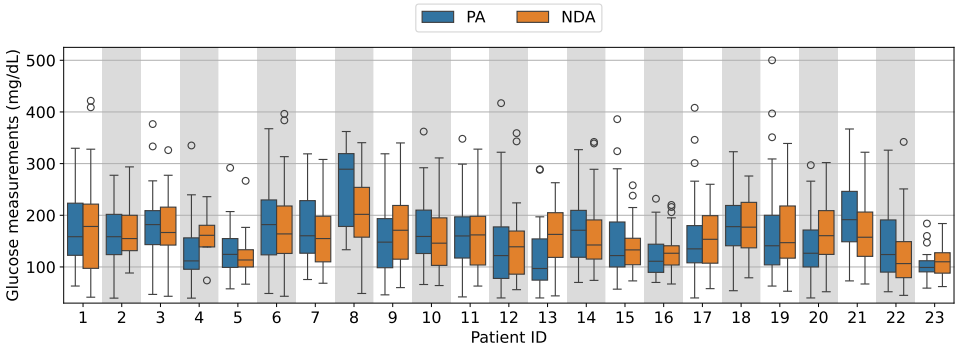
Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

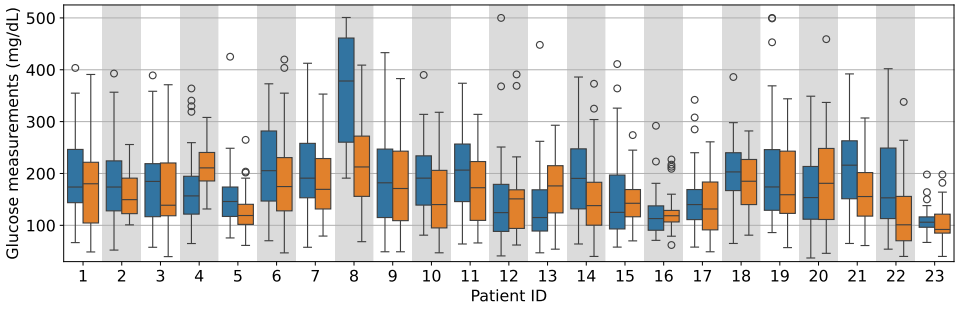
The following abbreviations are used in this manuscript:

CEG	Clarke error grid
CGM	Continuous glucose monitoring
DT	Decision tree
KNN	K-nearest neighbors
LR	Logistic regression
PA	Physical activity
MAD	Mean absolute deviation
MARD	Mean absolute relative deviation
NDA	Normal daily activity
ML	Machine learning
MLP	Multilayer perceptron
RF	Random forest
SHAP	Shapley additive explanations
SVM	Support vector machine
T1D	Type 1 diabetes
XGB	Extreme gradient boosting

Appendix A

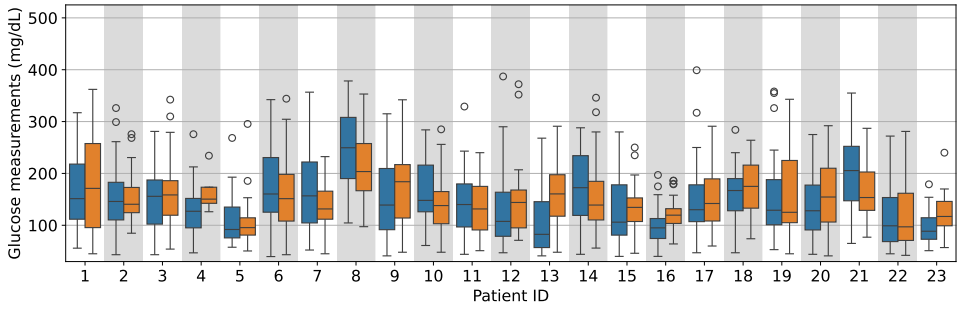


(**a**)



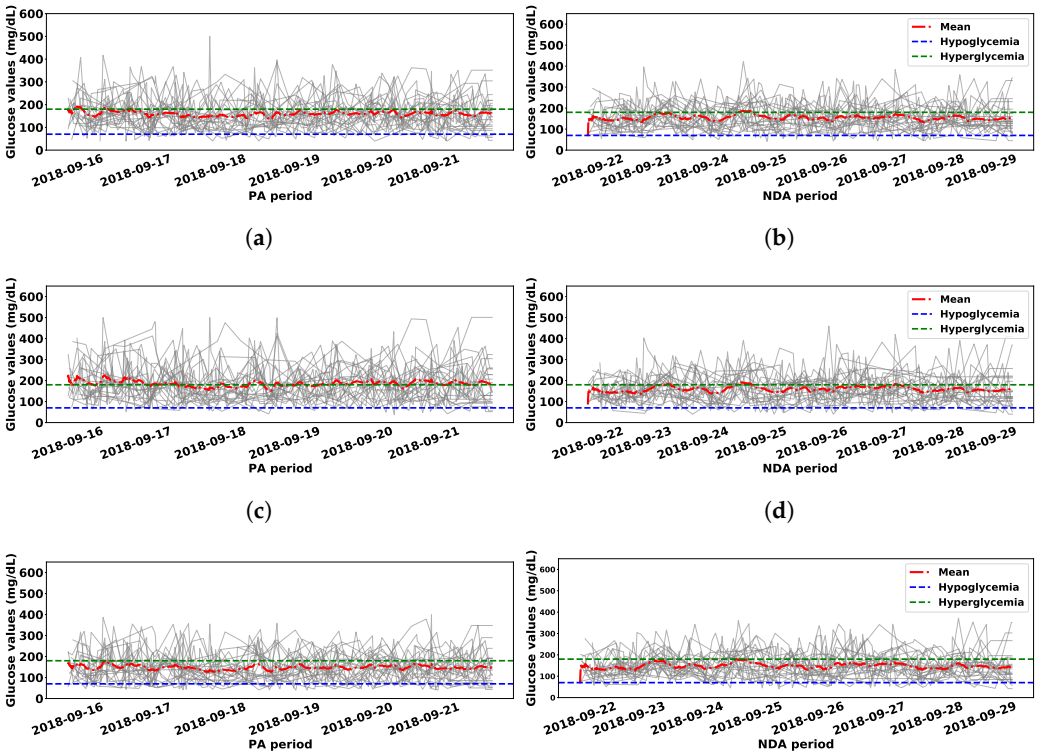
(**b**)

Figure A1. *Cont.*



(**c**)

Figure A1. Box plots of glucose concentrations values considering the periods of PA and NDA and the devices (**a**) FSLCstrip; (**b**) FSL; and (**c**) Eversense.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure A2. Temporal evolution of glucose concentrations values using (**a**,**b**) FSLCstrip; (**c**,**d**) FSL; and (**e**,**f**) Eversense. Figures in the first column correspond to the PA period, whereas those in the second column correspond to NDA.

References

1. DiMeglio, L.A. ; Evans-Molina, C.; Oram, R.A. Type 1 diabetes. *Lancet* **2018**, *391*, 2449–2462. [CrossRef] [PubMed]
2. Janež, A.; Guja, C.; Mitrakou, A.; Lalic, N.; Tankova, T.; Czupryniak, L.; Tabák, A.G.; Prazny, M.; Martinka, E.; Smircic-Duvnjak, L. Insulin therapy in adults with type 1 diabetes mellitus: A narrative review. *Diabetes Ther.* **2020**, *11*, 387–409. [CrossRef] [PubMed]
3. Nimri, R.; Dassau, E.; Segall, T.; Muller, I.; Bratina, N.; Kordonouri, O.; Bello, R.; Biester, T.; Dovc, K.; Tenenbaum, A.; et al. Adjusting insulin doses in patients with type 1 diabetes who use insulin pump and continuous glucose monitoring: Variations among countries and physicians. *Diabetes Obes. Metab.* **2018**, *20*, 2458–2466. [CrossRef] [PubMed]
4. Dunn, T.C.; Xu, Y.; Hayter, G.; Ajjan, R.A. Real-world flash glucose monitoring patterns and associations between self-monitoring frequency and glycaemic measures: A European analysis of over 60 million glucose tests. *Diabetes Res. Clin. Pract.* **2018**, *137*, 37–46. [CrossRef]
5. Lee, I.; Probst, D.; Klonoff, D.; Sode, K. Continuous glucose monitoring systems—Current status and future perspectives of the flagship technologies in biosensor research. *Biosens. Bioelectron.* **2021**, *181*, 113054. [CrossRef]

6. Alfian, G.; Syafrudin, M.; Anshari, M.; Benes, F.; Atmaji, F.T.D.; Fahrurrozi, I.; Hidayatullah, A.F.; Rhee, J. Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features. *Biocybern. Biomed. Eng.* **2020**, *40*, 1586–1599. [CrossRef]
7. Wong, J.C.; Foster, N.C.; Maahs, D.M.; Raghinaru, D.; Bergenstal, R.M.; Ahmann, A.J.; Peters, A.L.; Bode, B.W.; Aleppo, G.; Hirsch, I.B.; et al. Real-time continuous glucose monitoring among participants in the T1D Exchange clinic registry. *Diabetes Care* **2014**, *37*, 2702–2709. [CrossRef] [PubMed]
8. Yu, X.; Yang, T.; Lu, J.; Shen, Y.; Lu, W.; Zhu, W.; Bao, Y.; Li, H.; Zhou, J. Deep transfer learning: A novel glucose prediction framework for new subjects with type 2 diabetes. *Complex Intell. Syst.* **2022**, *8*, 1875–1887. [CrossRef]
9. Colberg, S.R.; Sigal, R.J.; Yardley, J.E.; Riddell, M.C.; Dunstan, D.W.; Dempsey, P.C.; Horton, E.S.; Castorino, K.; Tate, D.F. Physical activity/exercise and diabetes: A position statement of the American Diabetes Association. *Diabetes Care* **2016**, *39*, 2065–2079. [CrossRef]
10. Codella, R.; Terruzzi, I.; Luzi, L. Why should people with type 1 diabetes exercise regularly? *Acta Diabetol.* **2017**, *54*, 615–630. [CrossRef]
11. Zaharieva, D.P.; Messer, L.H.; Paldus, B.; O'Neal, D.N.; Maahs, D.M.; Riddell, M.C. Glucose control during physical activity and exercise using closed loop technology in adults and adolescents with type 1 diabetes. *Can. J. Diabetes* **2020**, *44*, 740–749. [CrossRef] [PubMed]
12. Tonoli, C.; Heyman, E.; Roelands, B.; Buyse, L.; Cheung, S.S.; Berthoin, S.; Meeusen, R. Effects of different types of acute and chronic (training) exercise on glycaemic control in type 1 diabetes mellitus. *Sports Med.* **2012**, *42*, 1059–1080. [CrossRef] [PubMed]
13. Bally, L.; Zueger, T.; Pasi, N.; Carlos, C.; Paganini, D.; Stettler, C. Accuracy of continuous glucose monitoring during differing exercise conditions. *Diabetes Res. Clin. Pract.* **2016**, *112*, 1–5. [CrossRef]
14. Biagi, L.; Bertachi, A.; Quirós, C.; Giménez, M.; Conget, I.; Bondia, J.; Vehí, J. Accuracy of continuous glucose monitoring before, during, and after aerobic and anaerobic exercise in patients with type 1 diabetes mellitus. *Biosensors* **2018**, *8*, 22. [CrossRef]
15. Larose, S.; Taleb, N.; Roy-Fleming, A.; Suppere, C.; Messier, V.; Rabasa-Lhoret, R. Comparison of Continuous Glucose Monitoring with Capillary Glucose Levels and Dynamics of Accuracy Changes during Moderate-Intensity Aerobic Exercise in Patients with Type 1 Diabetes. *Can. J. Diabetes* **2018**, *42*, S51–S52. [CrossRef]
16. Zaharieva, D.P.; Turksoy, K.; McGaugh, S.M.; Pooni, R.; Vienneau, T.; Ly, T.; Riddell, M.C. Lag time remains with newer real-time continuous glucose monitoring technology during aerobic exercise in adults living with type 1 diabetes. *Diabetes Technol. Ther.* **2019**, *21*, 313–321. [CrossRef]
17. Cuerda del Pino, A.; Martín-San Agustín, R.; José Laguna Sanz, A.; Díez, J.L.; Palanca, A.; Rossetti, P.; Gumbau-Gimenez, M.; Ampudia-Blasco, F.J.; Bondia, J. Accuracy of Two Continuous Glucose Monitoring Devices During Aerobic and High-Intensity Interval Training in Individuals with Type 1 Diabetes. *Diabetes Technol. Ther.* **2024**, *26*, 411–419. [CrossRef] [PubMed]
18. Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A.L.; Chen, I.Y.; Ranganath, R. A review of challenges and opportunities in machine learning for health. *AMIA Summits Transl. Sci. Proc.* **2020**, *2020*, 191.
19. Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M.A.; Shambour, M.K.Y.; Alsalibi, A.I.; Gandomi, A.H. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* **2022**, *145*, 105458. [CrossRef]
20. Chushig-Muzo, D.; Calero-Díaz, H.; Lara-Abelenda, F.J.; Gómez-Martínez, V.; Granja, C.; Soguero-Ruiz, C. Interpretable data-driven approach based on feature selection methods and GAN-based models for cardiovascular risk prediction in diabetic patients. *IEEE Access* **2024**, *12*, 84292–84305. [CrossRef]
21. Munoz-Organero, M. Deep physiological model for blood glucose prediction in T1DM patients. *Sensors* **2020**, *20*, 3896. [CrossRef] [PubMed]
22. Kodama, S.; Fujihara, K.; Shiozaki, H.; Horikawa, C.; Yamada, M.H.; Sato, T.; Yaguchi, Y.; Yamamoto, M.; Kitazawa, M.; Iwanaga, M.; et al. Ability of current machine learning algorithms to predict and detect hypoglycemia in patients with diabetes mellitus: Meta-analysis. *JMIR Diabetes* **2021**, *6*, e22458. [CrossRef]
23. Woldaregay, A.Z.; Årsand, E.; Walderhaug, S.; Albers, D.; Mamykina, L.; Botsis, T.; Hartvigsen, G. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artif. Intell. Med.* **2019**, *98*, 109–134. [CrossRef] [PubMed]
24. Woldaregay, A.Z.; Årsand, E.; Botsis, T.; Albers, D.; Mamykina, L.; Hartvigsen, G. Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes. *J. Med. Internet Res.* **2019**, *21*, e11030. [CrossRef]
25. Chushig-Muzo, D.; Soguero-Ruiz, C.; De Miguel-Bohoyo, P.; Mora-Jiménez, I. Learning and visualizing chronic latent representations using electronic health records. *BioData Min.* **2022**, *15*, 18. [CrossRef]
26. Contreras, I.; Vehi, J. Artificial intelligence for diabetes management and decision support: Literature review. *J. Med. Internet Res.* **2018**, *20*, e10775. [CrossRef]
27. 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]
28. van Doorn, W.P.; Foreman, Y.D.; Schaper, N.C.; Savelberg, H.H.; Koster, A.; van der Kallen, C.J.; Wesselius, A.; Schram, M.T.; Henry, R.M.; Dagnelie, P.C.; et al. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. *PLoS ONE* **2021**, *16*, e0253125. [CrossRef]

29. Cescon, M.; Choudhary, D.; Pinsker, J.E.; Dadlani, V.; Church, M.M.; Kudva, Y.C.; Doyle, F.J., III; Dassau, E. 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]
30. Dénes-Fazakas, L.; Siket, M.; Szilágyi, L.; Kovács, L.; Eigner, G. Detection of Physical Activity Using Machine Learning Methods Based on Continuous Blood Glucose Monitoring and Heart Rate Signals. *Sensors* **2022**, *22*, 8568. [CrossRef]
31. Cho, S.; Aiello, E.M.; Ozaslan, B.; Riddell, M.C.; Calhoun, P.; Gal, R.L.; Doyle, F.J., III. Design of a real-time physical activity detection and classification framework for individuals with type 1 diabetes. *J. Diabetes Sci. Technol.* **2023**, *18*, 1146–1156. [CrossRef] [PubMed]
32. Fokkert, M.; van Dijk, P.R.; Edens, M.A.; Hernández, A.D.; Slingerland, R.; Gans, R.; Álvarez, E.D.; Bilo, H. Performance of the Eversense versus the Free Style Libre Flash glucose monitor during exercise and normal daily activities in subjects with type 1 diabetes mellitus. *BMJ Open Diabetes Res. Care* **2020**, *8*, e001193. [CrossRef]
33. Fokkert, M.; Van Dijk, P.; Edens, M.; Abbes, S.; De Jong, D.; Slingerland, R.; Bilo, H. Performance of the FreeStyle Libre Flash glucose monitoring system in patients with type 1 and 2 diabetes mellitus. *BMJ Open Diabetes Res. Care* **2017**, *5*, e000320. [CrossRef] [PubMed]
34. Clarke, W.L.; Anderson, S.; Farhy, L.; Breton, M.; Gonder-Frederick, L.; Cox, D.; Kovatchev, B. Evaluating the clinical accuracy of two continuous glucose sensors using Continuous glucose–error grid analysis. *Diabetes Care* **2005**, *28*, 2412–2417. [CrossRef]
35. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [CrossRef]
36. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]
37. Bommert, A.; Sun, X.; Bischl, B.; Rahnenführer, J.; Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* **2020**, *143*, 106839. [CrossRef]
38. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
39. He, Z.; Li, L.; Huang, Z.; Situ, H. Quantum-enhanced feature selection with forward selection and backward elimination. *Quantum Inf. Process.* **2018**, *17*, 154. [CrossRef]
40. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
41. Battineni, G.; Sagaro, G.G.; Chinatalapudi, N.; Amenta, F. Applications of machine learning predictive models in the chronic disease diagnosis. *J. Pers. Med.* **2020**, *10*, 21. [CrossRef] [PubMed]
42. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.
43. Kramer, O. *Dimensionality Reduction with Unsupervised Nearest Neighbors*; Springer: Berlin/Heidelberg, Germany, 2013.
44. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
45. Williams, G. Support vector machines. In *Data Mining with Rattle and R*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 293–304.
46. Tama, B.A.; Rhee, K.H. Tree-based classifier ensembles for early detection method of diabetes: An exploratory study. *Artif. Intell. Rev.* **2019**, *51*, 355–370. [CrossRef]
47. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv* **2022**, arXiv:2207.01848.
48. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
49. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
50. Ding, W.; Abdel-Basset, M.; Hawash, H.; Ali, A.M. Explainability of Artificial Intelligence Methods, Applications and Challenges: A Comprehensive Survey. *Inf. Sci.* **2022**, *615*, 238–292. [CrossRef]
51. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
52. Chou, Y.L.; Moreira, C.; Bruza, P.; Ouyang, C.; Jorge, J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inf. Fusion* **2022**, *81*, 59–83. [CrossRef]
53. Moreira, C.; Chou, Y.L.; Velmurugan, M.; Ouyang, C.; Sindhgatta, R.; Bruza, P. LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decis. Support Syst.* **2021**, *150*, 113561. [CrossRef]
54. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.