Supplemental Material

Noninvasive hypoglycemia detection in people with diabetes using smartwatch data

Table of content

Table S1: Input features for the machine learning model

The Garmin vivoactive 4s was used to record heart rate, heart rate variability, motion and time signals. The Empatica E4 was used to record electrodermal activity signals. We applied identical aggregation functions for heart rate, electrodermal activity, and motion signals. For heart rate variability, we used aggregation functions to compute the features in the time- and frequency-domain. Additionally, time encoded as full hours from 0100 to 2400 was added as a feature. This resulted in a total of 37 input features.

		Term	Explanation
Data signals	**Garmin vivoactive 4s**	Heart rate	Number of heart beats within one minute
		Heart rate variability	Variation in the beat-to-beat intervals
		Motion	Zero-crossing (i.e., the total number of sign changes on the z-axis over a 30 second window)
		Time	Time of the day encoded in 24 hours
	Empatica E4	Electrodermal activity	Electrical conductance of the skin
Aggregation functions	**Aggregation functions for heart rate, electrodermal activity, and motion**	$Mean$	Arithmetic mean, measures the average value in a time series
		STD	Standard deviation, measures the amount of variation of the values in the time series
		IQR	Interquartile range, IQR between the 25th and 75th percentile of the signal
		IQR_{5-95}	Interquartile range, IQR between the 5th and 95th percentile of the signal
		P_5	5th percentile of the signal
		P_{95}	95th percentile of the signal
	Time-domain heart rate variability	$SDNN$	Standard deviation of all inter-beat (NN) intervals
		$SDSD$	Standard deviation of the differences between successive NN intervals
		$RMSSD$	The square root of the mean of the sum of the squares of differences between adjacent NN intervals
		pNN_{50}	Number of pairs of adjacent NN intervals differing by more than 50 milliseconds (ms) in the entire recording divided by the total number of all NN intervals
		pNN_{20}	Number of pairs of adjacent NN intervals differing by more than 20 ms in the entire recording divided by the total number of all NN intervals
		$CVNN$	Coefficient of variation equal to the ratio of SDNN divided by mean NN interval
		$CVSD$	Coefficient of variation of successive differences equal to the RMSSD divided by mean NN interval

	Frequency-domain heart rate variability	$Totalpower$	The variance of NN intervals over the temporal segment below 0.04 Hz
		vlf	Power in very low frequency range below or equal 0.04 Hz
		lf	Power in low frequency range 0.04 Hz and 0.15 Hz
		hf	Power in high frequency range 0.15 Hz and 0.4 Hz
		$lf/hf - ratio$	Ratio of lf to hf

Supplemental Methods 1: Sample size consideration

To estimate the sample size, we considered two scenarios. Today's consumer smartwatches include heart rate and motion sensors (scenario 1). The latest consumer smartwatches now entering the market also have electrodermal activity sensors (e.g., Fitbit Sense 2 and Versa 4, scenario 2, the focus of this paper). To estimate the number of participants for scenario 1 we assumed a moderate effect size of 0.5 and used GPower, which suggested a sample size of n=34 to achieve a power of 0.80 at an alpha level of 0.05. To estimate the number of participants for scenario 2 we assumed a higher effect size of 0.7 and used GPower, which suggested a sample size of n=19 to achieve a power of 0.80 at an alpha level of 0.05. However, we must acknowledge that estimating sample size for machine learning performance is challenging (1; 2), especially for observational studies without prior data on the main outcome (detecting hypoglycemia from smartwatch data). To address scenarios 1 and 2 while expecting ≈15% dropouts, we recruited 40 individuals. Ultimately, 31 participants used both wearables to capture heart rate, motion, and electrodermal activity (scenario 2), with 9 individuals having less than two hypoglycemic events, leaving 22 eligible for the final analysis.

To estimate the expected hypoglycemic events, we used retrospective continuous glucose measurement (CGM) data from a previous study performed at our clinic. Based on this data, we assumed a frequency of 0.2 hypoglycemic events per day in individuals on insulin treatment (3) and considered 30 days as the feasible study duration for our participants. Therefore, for scenario 1, we expected ≈200 hypoglycemic events (i.e., 34*30 days*0.2), for scenario 2 we expected ≈110 events (i.e., 19*30 days*0.2). Our final analysis comprised a total of 197 hypoglycemic events, thereby exceeding the assumptions.

Supplemental Methods 2: Detailed description of the pairing approach

For model building, we follow (4) and refer to it as the pairing approach. In brief, (4) proposed to segment the data (e.g., using a clustering algorithm), train individual models on each segment, and subsequently, find the best model for an unseen observation. We apply the same steps but adopted the approach slightly to the problem setting of hypoglycemia detection.

In the training stage (first step), we perform the subject segmentation and model training. Taking the size of the data set into account, we consider each subject as a segment instead of using an additional clustering algorithm. Then, we train a machine learning model on each subject, specifically a gradient-boosting decision tree. For the implementation, we leverage the open-source Python package LightGBM (version 3.3.2).

In the validation stage (second step), we find (pair) the best model for an unseen test subject based on a validation score (i.e., the mean area under the receiver operating curve [AUROC] calculated according to the event-based cross-validation).

Finally, in the testing stage (third step) we compute the mean performance of the paired model for the unseen test subject. For the implementation, we refer to our publicly available source code (https://github.com/im-ethz/radar).

Concerning our reported results, three subjects could not be paired. The three individuals were not paired during the validation process as their validation scores, returned by the machine learning models, were below a minimum threshold that is enforced to ensure a minimum or fair classification performance (AUROC >0.7, (1)). Most likely, the training dataset size for these three individuals is too small and the ML model was not able to learn generalizable patterns. Specifically, these three subjects

might be paired if more individuals were available in the training data set. Unfortunately, we could not empirically test this hypothesis.

References

1. Obuchowski NA, Lieber ML, Wians FH, Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 2004;50:1118-1125
2. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making 2012;12:8
3. Zueger T, Gloor M, Lehmann V, Melmer A, Kraus M, Feuerriegel S, Laimer M, Stettler C. White coat adherence effect on glucose control in adult individuals with diabetes. Diabetes Res Clin Pract 2020;168:108392
4. De Caigny A, Coussement K, De Bock KW. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 2018;269:760-772