BEYOND GLUCOSE-ONLY ASSESSMENT: ADVANCING NOCTURNAL HYPOGLYCEMIA PREDICTION IN CHILDREN WITH TYPE 1 DIABETES

Marco Voegeli* ETH Zurich
Sonia Laguna* ETH Zurich
Heike Leutheuser University of Bayreuth
Marc Pfister University Children's Hospital Basel
Marie-Anne Burckhardt University Children's Hospital Basel
Sara Bachmann University Children's Hospital Basel
Julia E. Vogt ETH Zurich

ABSTRACT

The *dead-in-bed* syndrome describes the sudden and unexplained death of young individuals with Type 1 Diabetes (T1D) without prior long-term complications. One leading hypothesis attributes this phenomenon to nocturnal hypoglycemia (NH), a dangerous drop in blood glucose during sleep. This study aims to improve NH prediction in children with T1D by leveraging physiological data and machine learning (ML) techniques. We analyze an in-house dataset collected from 16 children with T1D, integrating physiological metrics from wearable sensors. We explore predictive performance through feature engineering, model selection, architectures, and oversampling. To address data limitations, we apply transfer learning from a publicly available adult dataset. Our results achieve an AUROC of 0.75 ± 0.21 on the in-house dataset, further improving to 0.78 ± 0.05 with transfer learning. This research moves beyond glucose-only predictions by incorporating physiological parameters, showcasing the potential of ML to enhance NH detection and improve clinical decision-making for pediatric diabetes management.

1 INTRODUCTION

Type 1 Diabetes (T1D) is a chronic autoimmune disease characterized by the destruction of pancreatic beta cells (American Diabetes Association, 2009). In 2021, approximately 8.4 million individuals worldwide had T1D, 18% being under 20 years. By 2040, the number of cases is projected to increase by 60-107%, particularly in low-income and lower-middle-income countries (G. A. Gregory et al., 2022). A major complication of insulin-treated diabetes is hypoglycemia, which occurs when blood glucose levels drop below 70 mg/dL (3.9 mmol/L) (Mathew & Thoppil, 2024). Hypoglycemia occurring at night, mainly during sleep, is known as nocturnal hypoglycemia (NH) and poses a severe threat to individuals with T1D (Kalra et al., 2013; Edelman & Blose, 2014). Because individuals are often unaware of hypoglycemic events while asleep, they are unable to take corrective actions in real time. Beyond the immediate physiological risks, NH can impair sleep quality, reduce daytime cognitive function, and increase the likelihood of cardiovascular complications, psychological distress, and fear of future episodes (Allen & Frier, 2003; Brod et al., 2012). An early-warning system for NH could help mitigate these risks by allowing individuals to take necessary precautions in advance.

Glycemic variability, a key factor influencing NH, is affected by several daily lifestyle habits, including physical activity (Zhu et al., 2021). Managing glycemic variability in children is particularly challenging due to the physiological changes during puberty and their limited understanding of the

*Equal contribution. Correspondence to slaguna@inf.ethz.ch

topic (Nadella et al., 2017; Franzese et al., 2004). Wearable health monitoring devices have gained widespread adoption (Piwek et al., 2016; Lu et al., 2020), presenting an opportunity to integrate physiological data into NH prediction models. By leveraging these devices in combination with machine learning (ML) models, the prediction and prevention of NH could be significantly improved, reducing life-threatening complications.

ML has shown promise in the medical domain for predictive modelling, diagnostics, and automation of complex decision-making tasks (Alex et al., 2012; He et al., 2015). However, challenges such as data sparsity, limited patient-specific training data, and out-of-domain distributions complicate the development of robust predictive models (Javaid et al., 2021; Araújo et al., 2016). To address these challenges, this study employs tailored ML techniques to predict NH in children with T1D.

Overall, our main contributions are: (i) We introduce a novel NH prediction approach in children, going beyond traditional glucose-only methods by integrating physiological signals from wearable sensors and focusing on a challenging long prediction horizon, not generally tackled in the literature. (ii) We explore advanced feature selection, tailored preprocessing, and optimized model architectures to tackle the challenges of high-feature, low-cardinality data. (iii) We demonstrate the power of transfer learning by effectively leveraging adult data to enhance predictive performance for pediatric diabetes management.

2 Related Work

Several studies have explored ML approaches for predicting NH, leveraging different datasets, prediction horizons, and modelling techniques. Vu et al. (2020) investigated NH classification for two prediction windows: 0:00-3:00 am and 3:00-6:00 am, achieving AUROC scores of 0.90 and 0.75, respectively. Their model, a Random Forest Classifier (RFC), was trained on a large dataset comprising 1 million continuous glucose monitoring (CGM) data points from adults aged 45.34 ± 16.38 years. Mosquera-Lopez et al. (2020) focused on predicting minimum nocturnal glucose concentration across an entire night using a dataset of 22,804 nights from donors aged 31 ± 19 years. The authors extracted features from CGM, insulin intake, and meal data and employed a Support Vector Regressor (SVR) to estimate glucose concentration. The model then predicted NH events, achieving 94.1% accuracy in correctly identifying NH nights and an AUROC score of 0.86 (95% CI, 0.75–0.98). Berikov et al. (2022) took a different approach, using a dataset of 36,900 CGM data points from adults aged 18–70 years, along with 23 clinical and laboratory parameters. They employed an RFC model to predict NH at shorter prediction horizons of 15 and 30 minutes, achieving AUROC scores of 0.97 and 0.942, respectively. This study incorporated glucose metric extraction and additional physical parameters influencing NH risk. Lastly, Bertachi et al. (2020) examined NH prediction using physiological metrics from wearable sensors in addition to CGM data. The study collected data over 12 weeks from 10 adult participants, yielding approximately 840 nights of data. The authors trained an SVM model to predict NH with a 6-hour prediction horizon, beginning at sleep onset, achieving sensitivity and specificity scores of 78.75% and 82.15%, respectively. Despite the smaller dataset, our work stands out by integrating physiological features from wearable devices and exploring transfer learning across pediatric and adult datasets, providing a more robust evaluation framework and addressing key gaps in NH prediction research.

3 Study Framework: Datasets for NH

3.1 In-house Dataset

Study Setup The in-house dataset originates from a one-week sports day camp for children with T1D, approved by the local ethics committee. The study ran from 7:00 am on day 1 to 10:00 am on day 7, with pediatric endocrinologists supervising from 9:00 am to 5:00 pm. Activities, insulin treatment, and nutrition throughout the study were standardized. The first day included climbing, while days 2–6 featured structured sports. The final day, mainly concluding the study, was excluded from the analysis due to missing overnight data.

Participants 16 children aged 7–16 years, diagnosed with T1D for at least six months, participated. They used either multiple daily injections (MDI) or continuous subcutaneous insulin infusion

(CSII) for insulin therapy. Written informed consent was obtained from children and/or caregivers before the study. Finally, data from 11 children were used; the remaining 5 were excluded due to recording errors.

Hardware Devices used for this study consisted of a physiological wearable sensor Everion (Biofourmis, Boston, US) and a continuous subcutaneous glucose sensor. The Everion devices continuously recorded vital signs throughout the study (described here: 3.1) and were typically charged or replaced each morning when the children arrived at the camp. Glucose was measured continuously by a continuous subcutaneous glucose sensor; low glucose values were confirmed by a fingerprick (self-monitoring blood glucose SMBG) measurement. These recordings throughout the study were stored as distinct databases: an Everion database, consisting of the vital signs recorded by the Everion sensor; and a glucose database, storing the glucose readings, CGM, and SMBG. The Everion sensor is a CE-certified research device with a sampling rate of 1 Hz that captures 22 vital signs in real time, with corresponding quality measures. The Everion sensor was fitted on the upper part of the participant's arm (right or left). The glucose sensors for this study are intermittently scanned continuous glucose monitoring (isCGM), Freestyle libre 2 (Abbott Diabetes Care Inc., Alameda, US), CGM, Dexcom (Dexcom, San Diego, US) or Guardian 3 (Minimed Medtronic, Northridge, US) with a sampling rate of 5 minutes for the CGM devices and 15 minutes for the isCGM system.

Manual records The glucose dataset was completed with the SMBG records. The SMBG measurements were taken each time hyperglycemic or hypoglycemic symptoms were observed, i.e. sensor measurements were below 3.9 mmol/l or above 15 mmol/l, before and after physical activity and hourly during physical activity. Insulin doses in type, time, and units, carbohydrate intake, type and duration of physical activity, symptoms of hypoglycemia, and SMBG were noted in a logbook by the study team. The children continued the measurements and logbook entries at home in the evenings, nights, and mornings. Children's metadata, including morphological information such as age, weight, height, and body mass index (BMI) for each participant, were also recorded.

Features The dataset encompasses three distinct categories of features: time-dependent features detailed in Section A, logbook-recorded features presented in Table 1, and patient-specific metadata attributes presented in Table 6.

Table 1: List and description of features in the logbook dataset.

Logbook Feature	**Description**
Date	Date of the data entry
Time	Time of the data entry
Blood Sugar	Glucose level in the blood
Sensor Glucose	Sensor measured glucose level
SGL Trend	Time trend of sensor glucose level
Basal Insulin	Baseline insulin dose
Rapid-Acting Insulin Meals	Insulin dose for meal times
Rapid-Acting Insulin Correction	Dose to correct high blood sugar
Carbohydrates (g) Mixed	Mixed carbohydrates intake in grams
Carbohydrates (g) Fast	Fast-absorbing carbohydrates intake
Carbohydrates (g) Slow	Slow-absorbing carbohydrates intake
Hypo Correction (yes/no)	Whether a hypo correction was made
Type of Carbohydrates	The type of carbohydrates consumed
Duration of Exercise	Duration of physical activity
Duration of Exercise (estimate)	Estimated duration of physical activity
Type of Sport/Activity	Type of physical activity performed
Hypo Symptoms	Presence of hypoglycemia symptoms
Remarks	Additional notes

3.2 OHIOT1DM DATASET

The OhioT1DM dataset by Marling & Bunescu (2020) is a comprehensive and well-curated dataset specifically designed for research in T1D management. It includes detailed physiological and be-

havioural data collected from adults with T1D, providing valuable insights for developing and testing predictive models and treatment strategies.

Study Setup The dataset comprised two 8-week studies, one released in 2018 and one in 2020. The studies used different sensor bands with varying sampling rates and sensors.

Participants The dataset released in 2018 involved six participants, two males and four females, aged 20 to 40, and thus in 2020, six participants, five males and one female, aged 20 to 80.

Hardware The participants wore Medtronic 530G or 630G insulin pumps and Medtronic Enlite CGM sensors throughout the 8-week data collection. They reported life-event data via a custom smartphone app and physiological data from a fitness band. For the 2018 dataset, Basis Peak fitness bands were used. The six individuals used from the 2020 set wore the Empatica Embrace device.

Features The OhioT1DM dataset's full list of features is listed in Marling & Bunescu (2020). This study worked on the following features: Insulin type, glucose level (CGM data), finger stick (blood glucose values obtained through self-monitoring by the patient), hypo event (time of self-reported hypoglycemic episode), basis heart rate (heart rate, aggregated every 5 minutes), basis GSR (galvanic skin response, aggregated every 5 minutes), basis steps (step count, aggregated every 5 minutes), basis sleep (times when the sensor band reported that the subject was asleep), and acceleration (magnitude of acceleration, aggregated every 1 minute). A 5-minute aggregation of heart rate data is only available when participants wore the Basis Peak band (2018 cohort).

3.3 Labels

Labels for both datasets were calculated using overnight CGM and SMBG recordings (10 pm - 7 am). A night was classified as hypoglycemic if it met either of the following criteria: (1) CGM readings fell below the hypoglycemic threshold (3.9 mmol/L) for at least 15 consecutive minutes, or (2) any SMBG measurement was below 3.9 mmol/L. In the OhioT1DM dataset, CGM represents glucose levels, while SMBG refers to the fingerstick measurements.

4 Methods

4.1 Data Preprocessing

Prior to feeding the data to the predictive models, we preprocess it for homogeneity. The in-house dataset and OhioT1DM contained 60 and 308 labels, respectively. The sensors recorded data 24 hours a day, however for the in-house dataset the Everion devices were often swapped out in the mornings due to battery levels. This meant that a substantial amount of the recordings between 7am and 10am were missing. To reduce bias and minimize missing data, we restricted our analysis to the period between 10am and 10pm. This approach excludes the 7am to 10am window, which contained minimal useful information, ensuring that data closer to the prediction horizon—critical for accurate predictions—are prioritized. To further remove missing values from the sensor recordings we applied thresholding for each signal, the thresholds were taken from the technical ranges specified in Everion manual. The remaining missing values were substituted with zeros. Imputation methods such as polynomial and linear interpolation, excluding days with excessive missing values, and forward fill were considered. However zero imputation yielded the best results for our model.

When working on medical datasets, class imbalance is a common problem (Suresh et al., 2023). Especially for our study, the severe health risks posed by false negatives made us explore the distribution of our labels. The in-house dataset has a label imbalance ratio of 1:5.5 for hypoglycemic nights to normal nights, while the OhioT1DM dataset has a ratio of 1:2.1 respectively. Hence, we had to apply modifications to avoid overfitting.

We used the oversampling technique, ADASYN (Haibo et al., 2008) (derived from the SMOTE (Chawla et al., 2002) algorithm) with a 1-to-1 resampling ratio generating new data points close to the decision boundary. We chose ADASYN to tackle overfitting by generating challenging, hard-to-classify data points. Exposing our models to these tough examples helped them become more robust to outliers and less prone to overfitting. Generating synthetic samples for the minority class balanced

our labels and increased our dataset size. In Fig. 1, we observe the principal component analysis (PCA) illustration of the in-house dataset between the ADASYN augmented data and the raw data, where the generated samples are close to the decision boundary. This confirms that ADASYN generates hard-to-classify examples, enhancing our models. However, the drawback of oversampling is the introduction of synthetic data points, which can lead to biases in the data.

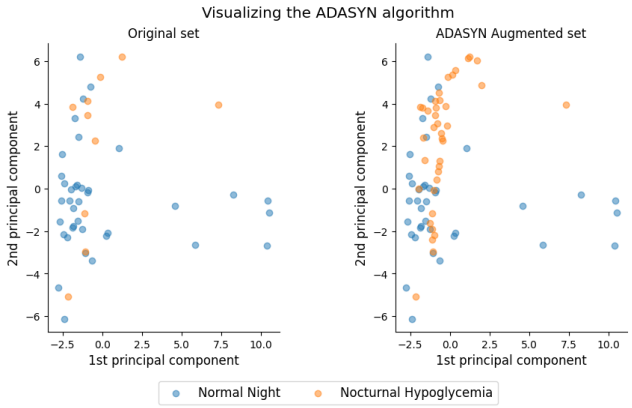

Figure 1: Two component principal component analysis (PCA) representation of the in-house dataset before and after ADASYN (Haibo et al., 2008) augmentation.

4.2 Feature Engineering

Features in the study were categorized into two types: temporal (time-varying) and static (unchanging). We resampled the temporal features to a 15-minute interval. Resampling the data to a 15-minute interval effectively reduces noise while preserving the sensitivity needed to detect critical events in a patient's glucose trajectory, such as cardiovascular variations during and after exertion (**?**Barak et al., 2010). The in-house dataset's temporal features consisted of 31 vital signs from the Everion device and 24 static features from logbook data and metadata, totalling 55 raw features for our model. We elaborate on the feature sets in the appendix A.

To improve the information extracted from key features like glucose and heart rate, we computed static daily features that capture the daily trends of these variables. This approach is informed by methodologies from previous studies, notably Berikov et al. (2022) and Sampath et al. (2016). The description of the functions used to aggregate the readings of the day are listed in Table 2

Data obtained closer to the prediction horizon holds greater significance in our predictive analysis (Metcalf et al., 2014); hence, on top of full-day trends, we extracted evening trends (7pm to 10pm). Aggregating the temporal dimension and extracting daily features is a technique to remove noise and reduce temporal dimensionality highlighting informative moments of the day.

We engineered a feature, glucose personalized (G_p), stemming that individual physiological factors—such as age, height, weight, and BMI—affect the person's glucose metabolism (Kashiwagi et al., 2023). To help the model personalise the glucose trends based on a person's physiological profile we designed the following feature,

$$G_p = \mathrm{G} \times (1 + (\mathrm{a} + \mathrm{h} + \mathrm{w} + \mathrm{b}))\,, \tag{1}$$

where G_p is the personalised glucose, G is the CGM reading, and a, h, w, b, are the age, height, weight, and BMI of the patient respectively.

Exploring the models' ability to learn on the in-house dataset, we compared performance on multiple feature sets. We analyzed seven feature sets ranging from 50 features (21 temporal, 29 static) to 16 features (6 temporal, 10 static), each chosen to isolate and highlight different aspects of the data:

All Features: Incorporates every available feature to serve as a comprehensive baseline.

Everion Daily Only: Focuses on the daily readings from the Everion device to assess the impact of daily aggregated data.

Table 2: Calculated daily features and their equations. n is the number of time steps throughout the day, G_i represents the glucose level at time step i, $\bar{G}$ is the mean glucose level, $\bar{D}$ is the mean of the first differences, σ_G is the standard deviation (SD) of glucose levels, and the "Evening" period refers to a specified time range within the day (7pm - 10pm).

Function	Equation
Coefficient of Variation	$\frac{\sigma_G}{\bar{G}}$
Liability Index	$\frac{1}{5}\sum_{i=1}^{n-1}(G_{i+1}-G_i)^2$
SD of the First Differences	$\sqrt{\frac{\sum_{i=1}^{n-1}(G_{i+1}-G_i-\bar{D})^2}{n-1}}$
Daily Minimum Value	$\min(G_1, G_2, \ldots, G_n)$
Evening Peak	$\max(G_{\text{Evening}})$
Evening Low	$\min(G_{\text{Evening}})$
Linear Regression Slope	$\frac{n\sum_{i=1}^{n} t_i G_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} G_i}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2}$

Glucose Normal: Uses standard glucose readings, providing a reference for conventional measurements.

Glucose Personalized: Adjusts glucose readings based on patient-specific factors to capture personalized insights.

Non-Aggregated Daily: Excludes features derived from daily calculations to evaluate the influence of these computed metrics.

Marx et al. (2023): Implements the feature set proposed by Marx et al. (2023) as a benchmark against established methodologies.

Reduced Selection: Employs a refined subset of features chosen based on the feature correlation aimed at optimizing model performance while reducing complexity.

Each feature set was selected with a specific rationale, allowing us to understand the contribution of various data aspects to our predictive performance.

4.3 Machine Learning Algorithms

4.3.1 Base Models

The Related Work section (Section 2) highlighted the robust performance of classical ML models, including the RFC and SVM, over the use of deep neural networks (DNN). These models consistently yield promising results, showcasing their ability to generalize well and mitigate overfitting when data points are limited.

We use three main models in this work: RFC (Do et al., 2009), Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) model, and Convolutional Neural Network (CNN) (Zhao et al., 2017). A RFC is a widely used ML model that aggregates the outputs of multiple decision trees to determine the final predicted class. It is particularly effective in scenarios involving high-dimensional inputs, where its ability to handle complex data makes it a preferred choice (Fernández-Delgado et al., 2014). LSTM and CNN models were also selected as comparative models because they are well-suited to our problem. Both models can handle two-dimensional input, which enables us to separate the time-related data from the other features in the dataset. LSTMs are popular due to their capacity to avoid vanishing gradients (Pascanu et al., 2012), capture uni-directional dependencies over long distances, and generalize effectively to unseen data (Greff et al., 2016). A convolutional neural network (CNN) abstracts the problem and excels at local pattern recognition (Alzubaidi et al., 2021), making it an ideal exploratory model for assessing the impact of temporal data.

4.3.2 DAILY VARIABLE MODELS

The in-house dataset has a mixture of static and temporal features, we maintained this temporal separation by customizing deep learning architectures. We adjusted the LSTM network and CNN architectures by passing the temporal features through the temporal layers (LSTM and CNN layers) to concatenate their non-temporal vector with our daily variables. For the hidden layers, we pass a Rectified Linear Unit (ReLU) (Agarap, 2018) as an activation function and for the binary classification a sigmoid (Narayan, 1997) activation function. Kernel regularization on the dense layers was used to mitigate overfitting.

4.3.3 TRANSFER LEARNING USING OHIOT1DM

Transfer learning involves leveraging knowledge gained from training on a larger, more general dataset (the OhioT1DM dataset) and applying it to a smaller, higher-feature dataset (the in-house dataset). In this process, a model pre-trained on a large dataset applies learned patterns to a new task. Fine-tuning on a smaller dataset allows the model to adapt quickly, needing less data and computational resources than starting from scratch. This approach makes sense for our limited, costly-to-collect dataset, enabling more efficient and effective model training.

The dataset was built from two studies where the participants used different sensor bands. This led to one of the studies not having the temporal physiological readings of the patients. To maximize the data points consistently and still demonstrate the potential of transfer learning, we chose to only extract the glucose recordings of this dataset.

The model used for transfer learning was a neural network consisting of LSTM layers pretrained on the glucose values of the OhioT1DM dataset. We then froze the trained layers and extended the architecture to encompass a set of hand-selected features from the in-house dataset based on the results of the exploratory feature sets. This allows us to leverage the OhioT1DM's cardinality.

The hand-selected features are the following: glucose, hypoglycemic flag, GSR electrode values, activity classification, blood pulse wave, core temperature, heart rate, heart rate variability, motion activity, number of steps, perfusion index, and respiration rate.

5 EXPERIMENTAL DETAILS

Implementation Our dataset's cardinality remained small throughout the studies; model training was done on CPUs. The libraries used were Numpy (Harris et al., 2020), Pandas (McKinney, 2010), Tensorflow (Abadi et al., 2015), PyTorch (Paszke et al., 2019), imblearn (Lemaître et al., 2017), and scikit-learn (Pedregosa et al., 2011). The RFC performed best with 1000 trees and a balancing of class weights. All models ran for 100 epochs. We employed early stopping with a patience of 30 epochs to accommodate the model's highly fluctuating performance during training, ensuring sufficient opportunity for stabilization and convergence. For our models' results, we used Stratified 5-fold Cross Validation.

Metrics and Evaluation Medical datasets often exhibit a significant class label imbalance (Kim, 2017; Rahman & Davis, 2013), which presents a major challenge, especially for conditions like NH. In such cases, the risk of false negatives is particularly critical because failing to predict an NH episode can severely affect patient safety (Allen & Frier, 2003; Kalra et al., 2013). Standard accuracy measurements fail to adequately address this scenario (Sun et al., 2011), necessitating careful consideration of metrics. To mitigate this issue, we utilized the binary cross entropy focal loss, (Lin et al., 2017), defined as

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{2}$$

with $FL(p_t)$ as the focal loss for a given probability p_t, with p_t being the model's estimated probability for the class with the true label t, α_t as the weight factor for class t, helping mitigate class imbalance by assigning more weight to the rare class, γ being the focusing parameter that smoothly adjusts the rate at which easy examples are down-weighted, and $\log$ denoting the natural logarithm. We opted for the AUROC score as a critical metric to assess and compare each model's performance. The AUROC score is a performance measurement for classification problems at various threshold

settings, well-defined under imbalanced datasets. It calculates the area under the receiver operating characteristic curve (Hajian-Tilaki, 2012). This curve illustrates the performance of a binary classification task along different threshold values. For completeness, we also report the F1-score

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

where Precision $= \frac{TP}{TP+FP}$, with TP as true positives and FP as false positives, and Recall $= \frac{TP}{TP+FN}$, with FN as false negatives. The F1 score gives a more comparative overview of the models' performances.

6 RESULTS

6.1 IN-HOUSE: FEATURE EXPLORATION

Our initial results, reported in Table 3, show the AUROC performance of the different models over the different feature sets. Table 7 in Appendix C details the corresponding F1 scores. Figure 2 visualizes the distribution of mean AUROC scores across models and feature sets, highlighting performance variations. This provides a more comprehensive understanding of how different feature sets impact predictive performance across models.

Table 3: This table compares different feature selection paradigms (Section 4.2) across different predictive models. Reporting the mean±standard deviation AUROC scores across three different random seeds, five-fold stratified cross-validation, and ADASYN oversampling on the in-house dataset, for a prediction horizon of 10 pm to 7 am.

Feature set	**RFC**	**LSTM**	**CNN**	**DailyLSTM**	**DailyCNN**
All features	0.66 ± 0.25	$\mathbf{0.67 \pm 0.22}$	0.49 ± 0.23	0.52 ± 0.20	0.49 ± 0.19
Everion daily only	$\mathbf{0.64 \pm 0.22}$	0.62 ± 0.17	0.53 ± 0.17	0.58 ± 0.16	N/A
Glucose normal	$\mathbf{0.69 \pm 0.20}$	0.68 ± 0.21	0.57 ± 0.18	0.56 ± 0.21	0.54 ± 0.12
Glucose personalised	$\mathbf{0.75 \pm 0.21}$	0.70 ± 0.24	0.53 ± 0.24	0.59 ± 0.18	0.63 ± 0.19
Non-Aggregated Daily	$\mathbf{0.68 \pm 0.24}$	0.63 ± 0.19	0.49 ± 0.18	0.52 ± 0.19	0.49 ± 0.19
(Marx et al., 2023)	0.70 ± 0.21	$\mathbf{0.70 \pm 0.17}$	0.53 ± 0.21	0.63 ± 0.18	0.59 ± 0.19
Reduced selection	$\mathbf{0.69 \pm 0.20}$	0.63 ± 0.20	0.57 ± 0.19	0.57 ± 0.17	0.54 ± 0.14

Abbreviations: AUROC, Area Under the Receiver Operating Curve; RFC, Random Forest Classifier; LSTM, Long Short Term Memory; CNN, Convolutional Neural Network.

6.2 OHIOT1DM

We evaluated our models on the OhioT1DM dataset, reporting the mean AUROC score across three random seeds (see Table 4). Because the dataset already had predefined training and test splits, we did not perform stratified cross-validation. To ensure consistency with the in-house data, we selected features that were both reliably recorded and comparable. Specifically, we choose: glucose, hypoglycemia, GSR electrode, basal values, and skin temperature.

Table 4: AUROC scores and standard deviation averaged across three different random seeds for the NH prediction on the OhioT1DM dataset with a prediction horizon of 9 hours (10pm - 7am).

Metric	**RFC**	**LSTM**	**CNN**	**DailyLSTM**	**DailyCNN**
AUROC score	0.60 ± 0.03	0.60 ± 0.03	0.56 ± 0.03	0.47 ± 0.09	$\mathbf{0.67 \pm 0.04}$

Abbreviations: AUROC, Area Under the Reciever Operating Curve; RFC, Random Forest Classifier; LSTM, Long Short Term Memory; CNN, Convolutional Neural Network.

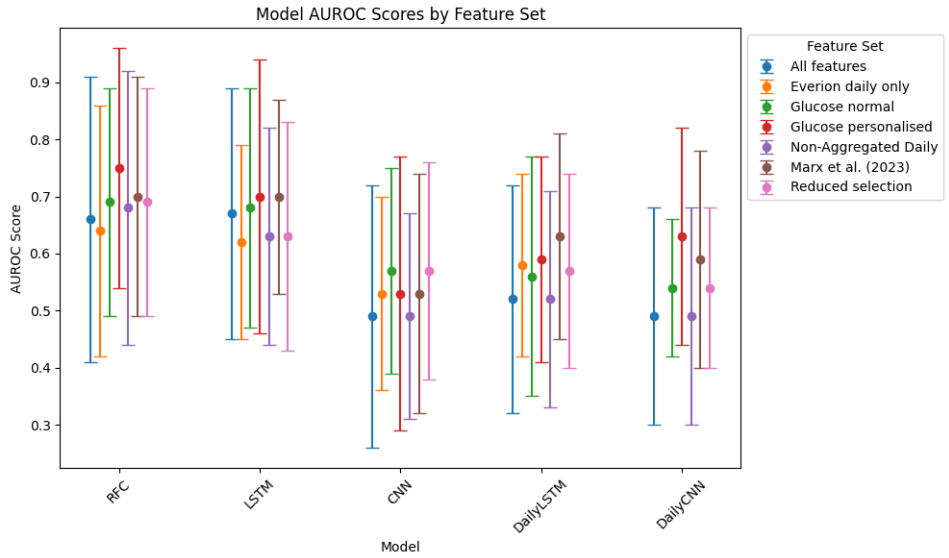

Figure 2: This figure displays the distribution of the mean AUROC cross-validation scores across three different random seeds for the different feature sets across models.

6.3 Transfer Learning

After optimally pretraining our model using the OhioT1DM dataset, the cross-validation results on the in-house dataset gave an average AUROC score of **0.78** $\pm$ **0.05**, using the LSTM described in Section 4.3.3.

7 Discussion

Feature Selection and Impact Among the evaluated feature sets, the personalized glucose feature set and Marx et al. (2023) feature set consistently outperformed the alternative sets. This suggests that a smaller, more informative feature subset enhances model learning. In contrast, using all available features resulted in poorer performance, likely due to noise and redundancy. Furthermore, models restricted to only global features underperformed, highlighting the importance of temporal information. Notably, the consistent drop in AUROC scores when excluding daily computed features further confirms their significance for accurate classification.

Model Performance on the In-House Dataset On the in-house dataset, the RFC model using the personalized glucose feature set achieved the highest AUROC score of 0.75 (Table 3). This indicates that reducing model complexity benefits model performance in such datasets. While both RFC and LSTM models showed competitive AUROC scores across feature sets, more complex models like DailyCNN and DailyLSTM performed worse. Suggesting that simpler architectures can generalize better given the dataset's constraints.

Comparative Dataset Overview The in-house and OhioT1DM datasets differ in both scope and study design. The in-house dataset captures children's glucose dynamics in a controlled setting, relying on wearable sensors. In contrast, OhioT1DM includes a broader age range with a greater reliance on self-reported life events. These differences are reflected in model performance, particularly in higher standard deviations for the in-house dataset due to its smaller sample size. Hence, increasing data samples improves convergence and model stability. Additionally, transfer learning with two different datasets still proves valuable in addressing model convergence, despite their dissimilarities in nature.

Cross-Dataset Model Comparison Comparing AUROC scores between datasets (Tables 3 and 4), key differences emerge. LSTM-based models (LSTM and DailyLSTM) consistently performed

better on the in-house dataset. CNN-based models (CNN and DailyCNN) exhibited higher AUROC scores on the OhioT1DM dataset. This discrepancy likely stems from CNN models leveraging larger sample sizes more effectively. Considering how prone the models are to overfitting for this particular dataset and classification task, fine-tuning remains key to achieving the best performance.

Transfer Learning Effectiveness Our transfer learning approach yielded the best results, reinforcing its robustness. Despite differences between the datasets, the method achieved an average AUROC of 0.78 (SD: 0.05). This demonstrates low variability and strong predictive performance, indicating that leveraging pre-trained knowledge improves predictive capabilities in varied contexts.

Comparison to Previous Studies Despite a smaller sample size (fewer than 314 data points), our findings remain competitive with prior state-of-the-art studies that used significantly larger CGM datasets (Mosquera-Lopez et al., 2020; Berikov et al., 2022). While our AUROC scores are slightly lower, our model maintained strong performance using a smaller and more diverse dataset over a broader prediction horizon, which solves a more clinically relevant and challenging problem. Notably, our model outperformed the 3 am - 6 am prediction horizon from Vu et al. (2020), despite having to apply transfer learning from an adult dataset to a children's dataset over an extended 9-hour horizon. This underscores the potential of informative features and transfer learning.

Limitations and Future Works The most obvious limitation of this study is the size of the dataset. A small dataset restricts the amount of information available for training and testing the models, leading to overfitting and poor generalizability (Brigato & Iocchi, 2020). This results in decreased performance when applied to new or unseen data. Despite transfer learning, manually entered logbook entries are inconsistent for both datasets, creating data gaps that need imputation. Overall, small sample sizes can introduce uncertainty and inaccuracies, resulting in significantly skewed model results.

For future research, an important focus should be on maintaining the temporal dimensions of our features when oversampling the minority class. One promising approach could be using SMOTE specifically for time-series data (Zhao et al., 2022), ensuring that the temporal dimension is accounted for during the data augmentation. An important aspect of ML in the medical field is its potential to assist healthcare professionals in making informed decisions. Future research could include confidence intervals in our predictions, which can provide a range of expected values conveying the model's uncertainty.

8 Conclusion

In conclusion, our work shows successful results in a relevant and critical issue in pediatrics: improving diabetic management for children with T1D, where even small advances can have a major effect. Particularly, we focus on a long prediction horizon, which has a broader impact on the clinics. Using a challenging dataset, due to its size and nature, we experimented with a variety of machine learning techniques, extensive feature engineering, and multiple models to address data complexity. Among these, the best performance on our in-house data was an AUROC score of 0.75 using the personalized glucose feature set and a Random Forest Classifier. Moreover, integrating adult data from a different distribution through transfer learning boosts the average AUROC scores from 0.75 to 0.78, and reduces the standard deviation from 0.21 to 0.05 (a 76% decrease). This way, we show both positive results in NH prediction and the benefits of incorporating diverse data sources to enhance model robustness and predictive accuracy. These findings pave the way for new research in predictive NH that moves beyond glucose-only methods by incorporating broader physiological data in long prediction horizons. This approach opens the door to more effective and less invasive clinical decision-making tools in pediatrics.

Acknowledgments

SL is supported by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number MB22.00047.

REFERENCES

M. Abadi, A. Agarwal, and P. et al. Barham. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

A. F. Agarap. Deep learning using rectified linear units (relu). *ArXiv*, 2018.

K. Alex, S. Ilya, and E. H. Geoffrey. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

K. V. Allen and B. M. Frier. Nocturnal hypoglycemia: clinical manifestations and therapeutic strategies toward prevention. *Endocr Pract*, 9(6):530–43, 2003. doi: 10.4158/EP.9.6.530.

L. Alzubaidi, J. Zhang, and A. J. et al. Humaidi. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021. doi: 10.1186/s40537-021-00444-8.

American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 33(Suppl 1):S62, 2009. doi: 10.2337/dc10-S062.

F. H. Araújo, M. X. Ribeiro, M. S. Marcolino, A. L. P. Ribeiro, V. Nobre, and W. Meira Jr. Using machine learning to support healthcare professionals in making preauthorisation decisions. *International Journal of Medical Informatics*, 94:1–7, 2016. doi: 10.1016/j.ijmedinf.2016.06.007.

Otto F. Barak, Djordje G. Jakovljevic, Jelena Z. Popadic Gacesa, Zoran B. Ovcin, David A. Brodie, and Nikola G. Grujic. Heart rate variability before and after cycle exercise in relation to different body positions. 9(2):176–182, 2010. ISSN 1303-2968.

V. B. Berikov, O.A. Kutnenko, J. F. Semenova, and Klimontov V. V. Machine learning models for nocturnal hypoglycemia prediction in hospitalized patients with type 1 diabetes. *J. Pers. Med.*, 2022.

A Bertachi, C. Viñals, L. Biagi, I. Contreras, J. Vehí, I Conget, and M Giménez. Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor. *Sensors (Basel)*, 20(6):1705, March 2020. doi: 10.3390/s20061705.

L Brigato and L Iocchi. A close look at deep learning with small data. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2490–2497, 2020.

M. Brod, T. Christensen, and D. M. Bushnell. Impact of nocturnal hypoglycemic events on diabetes management, sleep quality, and next-day function: results from a four-country survey. *Journal of Medical Economics*, 15(1):77–86, 2012. doi: 10.3111/13696998.2011.624144.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

T. Do, P Lenca, S. Lallich, and N. Pham. Classifying very-high-dimensional data with random forests of oblique decision trees. *Advances in knowledge discovery and management*, 01 2009. doi: 10.1007/978-3-642-00580-0_3.

S. V. Edelman and J. S. Blose. The impact of nocturnal hypoglycemia on clinical and cost-related issues in patients with type 1 and type 2 diabetes. *The Diabetes educator*, 2014.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90): 3133–3181, 2014.

A. Franzese et al. Management of diabetes in childhood: Are children small adults? *Clinical Nutrition*, 23(3):293–305, 2004. doi: 10.1016/j.clnu.2003.07.007.

T. I. G. Robinson G. A. Gregory, S. E Linklater, et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. *Lancet Diabetes Endocrinol*, 10(10):741–760, 2022. doi: 10.1016/S2213-8587(22)00218-2. [published correction appears in Lancet Diabetes Endocrinol. 2022 Nov;10(11):e11. doi: 10.1016/S2213-8587(22)00280-7].

K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2016. doi: 10.1109/TNNLS.2016.2582924.

H. Haibo, B. Yang, E. A. Garcia, and L. Shutao. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969.

K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4(2):627–635, 2012.

C. R. Harris, K. J. Millman, and S. J. et al. van der Walt. Array programming with NumPy. *Nature*, 585(7825):357–362, Sep 2020. doi: 10.1038/s41586-020-2649-2.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *ArXiv*, 2015.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

M. Javaid, A. Haleem, R.P. Singh, R. Suman, and S. Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3: 58–73, 2021. doi: 10.1016/j.ijin.2022.05.002.

S Kalra, J. J. Mukherjee, S Venkataraman, et al. Hypoglycemia: The neglected complication. *Indian J Endocrinol Metab.*, 2013.

K. Kashiwagi et al. Assessment of glycemic variability and lifestyle behaviors in healthy nondiabetic individuals according to the categories of body mass index. *PLOS ONE*, 18(10), 2023.

M. S. Kim. An effective under-sampling method for class. imbalance data problem. *International Symposium on Advance intelligent System*, 2017.

G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

L. Lu, J. Zhang, Y. Xie, F. Gao, S. Xu, X. Wu, Z. Ye, et al. Wearable health devices in health care: narrative systematic review. *JMIR mHealth and uHealth*, 8(11):e18907, 2020.

C. Marling and R. Bunescu. The ohiot1dm dataset for blood glucose level prediction: Update 2020. *CEUR Workshop Proceedings*, 2675:71, 2020.

A. Marx, F. Di Stefano, H. Leutheuser, K. Chin-Cheong, M. Pfister, M. A. Burckhardt, and J. E. Bachmann, S. Vogt. Blood glucose forecasting from temporal and static information in children with t1d. *Frontiers in pediatrics*, 2023.

P. Mathew and D. Thoppil. *Hypoglycemia. [Updated 2022 Dec 26]. In*. Treasure Island (FL): StatPearls Publishing, 2024.

W. McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

K. M. Metcalf, A. Singhvi, E. Tsalikian, M. J. Tansey, M. B. Zimmerman, D. W. Esliger, and K. F. Janz. Effects of moderate-to-vigorous intensity physical activity on overnight and next-day hypoglycemia in active adolescents with type 1 diabetes. *Diabetes care*, 37(5):1272–1278, 2014.

C. Mosquera-Lopez, R. Dodier, N. Tyler, S., L. M. Wilson, J. El Youssef, J. R. Castle, and P. G. Jacobs. Predicting and preventing nocturnal hypoglycemia in type 1 diabetes using big data analytics and decision theoretic analysis. *Diabetes technology therapeutics vol. 22*, 2020.

S. Nadella, J. A. Indyk, and M. K. Kamboj. Management of diabetes mellitus in children and adolescents: engaging in physical activity. *Translational Pediatrics*, 6(3):215–224, July 2017. doi: 10.21037/tp.2017.05.01.

S. Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information Sciences*, 99(1):69–82, 1997. ISSN 0020-0255. doi: https://doi.org/10.1016/S0020-0255(96)00200-9.

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2012.

A. Paszke, S. Gross, and F. et al. Massa. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

L. Piwek, D. A. Ellis, S. Andrews, and A. Joinson. The rise of consumer health wearables: promises and barriers. *PLoS medicine*, 13(2):e1001953, 2016.

M. M. Rahman and D. N. Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

S. Sampath, P. Tkachenko, E. Renard, and S. V. Pereverzev. Glycemic control indices and their aggregation in the prediction of nocturnal hypoglycemia from intermittent blood glucose measurements. *Journal of Diabetes Science and Technology*, 10(6):1245–1250, 2016. doi: 10.1177/1932296816670400.

Y. Sun, A. Wong, and M. S. Kamel. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23, 11 2011. doi: 10.1142/S0218001409007326.

T. Suresh, Z. Brijet, and T. D. Subha. Imbalanced medical disease dataset classification using enhanced generative adversarial network. *Comput Methods Biomech Biomed Engin*, 26(14):1702–1718, Oct-Dec 2023. doi: 10.1080/10255842.2022.2134729. Epub 2022 Nov 2.

L. Vu, S. Kefayati, T. Idé, et al. Predicting nocturnal hypoglycemia from continuous glucose monitoring data with extended prediction horizon. *AMIA Annu Symp Proc.*, 2020.

B. Zhao, H. Lu, S Chen, Liu J., and Wu D. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017. doi: 10.21629/JSEE.2017.01.18.

P. Zhao, C. Luo, B. Qiao, L. Wang, S. Rajmohan, Q. Lin, and D. Zhang. T-smote: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*. International Joint Conferences on Artificial Intelligence Organization, 2022.

X Zhu, L Zhao, J Chen, C Lin, F Lv, S Hu, X Cai, L Zhang, and L. Ji. The effect of physical activity on glycemic variability in patients with diabetes: A systematic review and meta-analysis of randomized controlled trials. *Front Endocrinol (Lausanne).*, 2021.

A FEATURE SETS

The following subsections contain the features used for the respective feature sets that were compared throughout this study.

A.1 ALL FEATURES

Glucose-related:
Glucose, hypoglycemia event, glucose linear regression slope, glucose evening low and peak, daily glucose minimum, standard deviation of glucose differences, coefficient of glucose variation.

Heart-related:
Heart rate, heart rate variability, heart rate evening low and peak, heart rate variability evening low and peak, heart rate variability minimum.

Physical Activity:
Motion activity, activity classification, number of steps, perfusion index, respiration rate, energy, activity score, wellness index, evening low and peak.

Temperature & Pressure:
Core temperature, temperature local, temperature object, barometer pressure.

Additional Biometrics:
Blood pulse wave, GSR electrode, gender, age, weight, height, BMI, basal percentage, basal total.

Insulin & Diabetes Data:
Glycated hemoglobin (HbA1c) reading, total daily insulin dose (TDD), max daily insulin fast, max daily insulin slow, total daily fast insulin, total daily slow insulin.

Others:
Health score, training effect score, richness score.

A.2 EVERION DAILY ONLY

Glucose-related:
Glucose, hypoglycemia flag.

Insulin & Diabetes Data:
Max insulin fast, max insulin slow, total insulin fast, total insulin slow, glycated hemoglobin (HbA1c) reading, total daily insulin dose.

Demographics:
Gender, age, weight, height, BMI, basal percentage, basal total.

Heart-related:
Heart rate variability evening low and peak, heart rate variability minimum.

A.3 GLUCOSE NORMAL

Core Features:
Glucose, hypoglycemia flag, heart rate, heart rate variability, number of steps.

Insulin & Diabetes Data:
Max insulin fast, max insulin slow, total insulin fast, total insulin slow, glycated hemoglobin (HbA1c) reading, total daily insulin dose.

Demographics:
Gender, age, weight, height, BMI, basal percentage, basal total.

Glucose Metrics:
glucose linear regression slope, glucose evening low, and peak, daily glucose minimum, standard deviation of the glucose differences, coefficient of glucose variation.

A.4 PERSONALIZED GLUCOSE

Glucose Metrics:
Glucose personalised (G_p), hypoglycemia events, glucose linear regression slope, glucose evening low and peak, daily glucose minimum, standard deviation of glucose differences, coefficient of glucose variation.

Heart-related:
Heart rate, heart rate variability.

Insulin-related:
Max insulin fast, max insulin slow, total insulin fast, total insulin slow.

A.5 NON-AGGREGATED DAILY

Core Features:
Glucose, hypoglycemia flag, heart rate, perfusion index, motion activity, activity classification, heart rate variability, respiration rate, energy, core temperature, temperature local, barometer pressure, GSR electrode, health score, training effect score, activity score, richness score, blood pulse wave, temperature object, temperature barometer.

Insulin-related:
Max insulin fast, max insulin slow, total insulin fast, total insulin slow.

A.6 (MARX ET AL., 2023)

Core Features:
Activity classification, blood pulse wave, core temperature, GSR electrode, heart rate, heart rate variability, motion activity, number of steps, perfusion index, respiration rate.

Demographics:
Gender, age, weight, height, BMI.

Insulin & Diabetes Data:
Basal percentage, basal total, glycated hemoglobin (HbA1c) reading, total daily insulin dose, max daily insulin fast, max daily insulin slow, total daily fast insulin, total daily slow insulin.

Glucose Metrics:
Glucose linear regression slope, glucose evening low and peak, daily glucose minimum, standard deviation of glucose differences, coefficient of glucose variation.

A.7 REDUCED SELECTION

Core Features:
Glucose, hypoglycemia flag, activity classification, blood pulse wave, core temperature, GSR electrode, heart rate, heart rate variability, motion activity, number of steps, perfusion index, respiration rate.

Insulin & Demographics:
Max daily insulin fast, max daily insulin slow, total daily fast insulin, total daily slow insulin, gender, age, weight, height, BMI, basal percentage, basal total, glycated hemoglobin (HbA1c) reading, total daily insulin dose.

Glucose Metrics:
Glucose linear regression slope, glucose evening low and peak, daily glucose minimum, standard deviation of glucose differences, coefficient of glucose variation.

B IN-HOUSE RECORDED FEATURES

In this appendix section, we list and describe in Tables 5 and 6 all the features recorded from the devices used during the In-house study.

Table 5: List and description of the Everion features, the bold features are the features used in this research.

Everion Feature	Description
Heart rate	Measures the number of heartbeats per minute.
Oxygen saturation	Assesses the percentage of oxygen-saturated hemoglobin.
Perfusion index	Indicates the pulse strength at the sensor site.
Motion activity	Tracks the movement activity of the wearer.
Activity classification	Categorizes the type of physical activity performed.
Heart rate variability	Monitors variations in the time interval between heartbeats.
Respiration rate	Number of breaths taken.
Energy	Energy Expenditure.
Core temperature	Monitors the internal body temperature.
Temperature local	Temperature at the device's location on the body.
Barometer pressure	The atmospheric pressure.
GSR electrode	Skin's electrical conductance.
Health score	Score of the wearer's health status.
Relax stress intensity score	Score on the intensity of stress and relaxation levels.
Sleep quality	index score Score on the quality of sleep.
Training effect score	Score on the impact of exercise on fitness levels.
Activity score	Score based on physical activity intensity and duration.
Richness score	Score based on physical activities undertaken.
Heart rate quality	Quality of heart rate measurements.
Oxygen saturation quality	Quality of oxygen saturation measurements.
Blood pulse wave	Metric of the blood pulse wave.
Number of steps	Number of steps taken.
Activity classification quality	Quality of activity classification.
Energy quality	Quality of energy expenditure estimations.
Heart rate variability quality	Quality of HRV measurements.
Respiration rate quality	Quality of respiration rate measurements.
Core temperature quality	Quality of core temperature measurements.
Temperature object	Temperature of an object in proximity to the device.
Temperature barometer	Temperature readings from the device's built-in barometer.
Timestamp UTC	Date and time in Coordinated Universal Time.
Timestamp offset	Local time offset from UTC at the time of measurement.

Table 6: List and description of metadata features.

Metadata Feature	Description
Gender	The biological sex of the individual (male or female).
Age	The age of the individual is typically measured in years.
Weight	The body weight of the individual, usually measured in kilograms (kg) or pounds (lbs).
Height	The stature of the individual, typically measured in centimetres (cm) or inches (in).
BMI (Body Mass Index)	A measure of body fat based on height and weight. It is calculated as weight in kilograms divided by the square of height in meters (kg/m²).
Basal Percentage	The percentage of total daily insulin that is basal (background insulin) to manage glucose levels over time.
Basal Total	The total daily amount of basal (long-acting) insulin, typically measured in units.
HbA1c (Glycated Hemoglobin)	A measure of average blood glucose levels over the past 2-3 months. It is expressed as a percentage and used to monitor long-term glucose control in people with diabetes.
TDD (Total Daily Insulin Dose)	The total amount of insulin taken in a day, including both basal (long-acting) and bolus (fast-acting) insulin, typically measured in units.

C Additional Results

We show in Table 7 the spread of the F1 scores for the different proposed feature sets and models applied to the In-house dataset. This table is used as an extension of Section 6.2 that contains the corresponding AUROC mean averages.

Table 7: Mean±standard deviation F1 scores across three different random seeds, five-fold stratified cross-validation, and ADASYN oversampling on the in-house dataset, for a prediction horizon of 10 pm to 7 am. This table compares different feature selection paradigms (Section 4.2) across different predictive models.

Feature set	RFC	LSTM	CNN	DailyLSTM	DailyCNN
All features	**0.43 ± 0.30**	0.43 ± 0.18	0.34 ± 0.18	0.25 ± 0.22	0.26 ± 0.14
Everion daily only	**0.47 ± 0.28**	0.45 ± 0.18	0.20 ± 0.21	0.30 ± 0.22	N/A
Glucose normal	**0.49 ± 0.28**	0.41 ± 0.21	0.32 ± 0.18	0.20 ± 0.18	0.29 ± 0.16
Glucose personalise	**0.54 ± 0.26**	0.45 ± 0.22	0.25 ± 0.25	0.16 ± 0.18	0.28 ± 0.24
Non-Aggregated Daily	**0.48 ± 0.32**	0.31 ± 0.18	0.29 ± 0.18	0.24 ± 0.18	0.08 ± 0.18
(Marx et al., 2023)	**0.52 ± 0.32**	0.41 ± 0.21	0.28 ± 0.19	0.33 ± 0.17	0.28 ± 0.25
Reduced selection	**0.54 ± 0.31**	0.42 ± 0.17	0.28 ± 0.18	0.26 ± 0.24	0.18 ± 0.18

Abbreviations: RFC, Random Forest Classifier; LSTM, Long Short Term Memory; GV, Daily Variable; REG, Regularised; CNN, Convolutional Neural Network.