# SugerMate: Non-intrusive Blood Glucose Inference with Smartphone

WEIXI GU, Tsinghua University
YUXUN ZHOU, University of California,Berkeley
ZIMU ZHOU, ETH Zurich
COSTAS J. SPANOS, University of California,Berkeley
LIN ZHANG, Tsinghua University

Inferring abnormal glucose events such as hyperglycemia and hypoglycemia is crucial for the health of both diabetic patients and non-diabetic people. However, continuous or regular blood glucose monitoring can be invasive and inconvenient in everyday life. In this paper, we present SugerMate, a non-intrusive smartphone-based abnormal blood glucose monitor system. Provided with inputs of food, drug and insulin intake, it leverages smartphone sensors to automatically measure physical activities and sleep quality, and then infers the current blood glucose level at a fine-grained time resolution. Nevertheless, accurate blood glucose level inference is still challenging due to model learning difficulties brought by the imbalanced, personalized, and often limited measurements. To this end, we propose Md$^3$RNN (multi-division deep dynamic recurrent neural network), a novel learning paradigm that is able to make full use of the available measurements. Specifically, Md$^3$RNN captures complex, multi-scale blood glucose dynamics via deep networks, extracts grouped feature representations with a multi-division learning structure, and preserves user-specific characteristics using personalized output layers. Evaluations on 112 users over 7 months show that Md$^3$RNN yields an average accuracy of 82.14%, significantly outperforming previous learning methods that are either shallow, generically structured, or oblivious to grouped behaviors.

CCS Concepts: **•Human-centered computing → Ubiquitous and mobile computing systems and tools; Smartphones; Mobile phones; •Computing methodologies → Artificial intelligence; Machine learning;**

**ACM Reference format:**
Weixi Gu, Yuxun Zhou, Zimu Zhou, Costas J. Spanos, and Lin Zhang. 2017. SugerMate: Non-intrusive Blood Glucose Inference with Smartphone. *PACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 0, Article 0 (June 2017), 21 pages.
DOI: 0000001.0000001

## 1 INTRODUCTION

Blood glucose concentration plays an important role in personal health. Hyperglycemia (high blood glucose level) results in diabetes, leading to health risks such as pancreatic function failure, immunity reduce and ocular fundus diseases [23]. Meanwhile, hypoglycemia (low blood glucose level) also brings complications such as confusion, shakiness, anxiety, and if not treated in time, coma or death [19]. The International Diabetes Federation (IDF) reports that there were 415 million diabetic patients in 2015 and the number will rise to 642 million by 2040 [14]. People with diabetes need tight control of their blood glucose concentration to avoid both short-term

Authors address: Weixi Gu and Lin Zhang, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China; E-mails: guweixigavin@gmail.com, linzhang@tsinghua.edu.cn; Authors address: Yuxun Zhou and Costas J. Spanos, Electrical Engineering and Computer Sciences, University of California, Berkeley, America; E-mails: yxzhou, spanos@berkeley.edu; Authors address: Zimu Zhou, Computer Engineering and Networks Laboratory, ETH Zurich, Switzerland; E-mail: zzhou@tik.ee.ethz.ch;

DOI: 0000001.0000001

and long-term physiological complications. While non-diabetic people normally have adequate self-regulation of blood glucose concentration, they can still risk hypoglycemia when taking prolonged exercises, drinking excess amounts of alcohol, having eating disorders, taking certain medicines (*e.g.*, certain painkiller and antibiotic), or having pre-diabetes [12, 15].

While continuous or regular blood glucose monitoring is essential for blood glucose management and beneficial for hyper- and hypoglycemia warning, it can be invasive and inconvenient, especially during daily life. A standard and direct blood glucose measurement is to collect and analyze a drop of blood by finger pricking, which requires a new prick on the finger for every new observation. Alternatively, non-invasive (without penetrating the skin) continuous glucose monitoring (CGM) has attracted extensive research leveraging techniques such as thermal infrared spectroscopy, Raman spectroscopy and impedance spectroscopy [9, 46]. However, most CGM devices are expensive, cumbersome to wear for extended time, complicated in terms of operation/maintenance, and are usually limited to clinical uses, making them unattractive for both diabetic patients and non-diabetic people.

Towards more ubiquitous blood glucose monitoring when traditional CGM devices are unavailable or inconvenient to general users, researchers propose to explore the increasingly rich sensors embedded in commercial fitness wearables and smartphones as a complement. In addition to the glucose metabolism that is difficult to measure directly, blood glucose also correlates to easily measurable physiological activities such as food and drug intake, energy expenditure, sleep quality and emotional states [22]. Pioneer works [33, 40, 45] have proposed preliminary systems leveraging commodity bio-sensors (*e.g.*, ECG electrodes) and fitness wearables (*e.g.*, accelerometer and galvanic skin response sensors) to predict blood glucose concentrations and alarm abnormal blood glucose events. Nevertheless, they are validated with limited number of measurements [40, 45] or still require complex multi-sensory platforms [33, 45].

In this work, we design SugerMate, the first personalized smartphone-based non-invasive blood glucose monitoring system that detects abnormal blood glucose events by jointly tracking meal, drug and insulin intake, physical activity and sleep quality. When SugerMate detects an abnormal blood glucose event, it reminds the user to double-check by finger pricking or using CGM devices. SugerMate exploits recent advances in smartphone-based automatic human activity recognition [25] and sleep quality measurement [20] to acquire external factors as input for abnormal blood glucose event detection, making it widely applicable in daily use. In addition, SugerMate considers both *generic* and *user-specific* correlations between blood glucose levels and the measurable external factors, which are largely overlooked in previous works.

However, it is challenging to learn effective, accurate and personalized blood glucose models in practice. While there have been general blood glucose models that characterizing universal trends between blood glucose concentration and various external factors [37], they have to be adjusted based on user-specific data to account for inter-user differences [3]. Yet it is often difficult to collect sufficient data to directly build up personalized models [31]. *(i)* A disposable enzyme of glucose sensor embedded in the CGM device is only capable of a few days [29, 49], and most users are unwilling to wear CGM devices frequently due to discomfort. *(ii)* Despite their importance, measurements of hyper- and hypoglycemia events are rare compared with normal blood glucose concentrations, making it difficult to accurately detect abnormal blood glucose events.

To take full advantages of the *sparse*, *imbalanced* measurements to build *personalized* blood glucose level models, we first conduct feature extraction from both physiological and temporal data analysis viewpoints. More importantly, we propose Md$^3$RNN (multi-division deep dynamic recurrent neural network), a novel learning paradigm that efficiently extracts general blood glucose level relevant features and preserves user-specific characteristics. Our Md$^3$RNN advances previous recurrent neural network (RNN) structures in two perspectives. On the one hand, it replaces the single hidden layer with multiple deep stacked layers to describe complex, multi-scale dynamics in blood glucose metabolism. On the other hand, it leverages a group-shared input layer to extract distinctive feature representations within the same group (*i.e.*, non-diabetic, type I and type II diabetic), and adds a personalized output layer to capture individual differences. In short, Md$^3$RNN can be regarded as both a

deep extension of RNN and a combination of single-task and multi-task learning. Evaluations on the blood glucose dataset composed of 112 users lasting 7 months show that our novel Md$^3$RNN framework outperforms both generic learning (*i.e.*, ignoring inter-user differences) and personalized learning (due to lack of measurements), and also achieves notably higher inference accuracy than conventional shallow learning algorithms.

The key contributions of this work are summarized as follows.

- To the best of our knowledge, SugerMate is the first smartphone-based personalized abnormal blood glucose event detection system that works without CGM data as input. It automatically collects exercise levels and sleep quality, together with manual records of food and drug intake, and infers the current blood glucose level of users.
- We propose Md$^3$RNN, a novel multi-division deep dynamic RNN framework able to *(i)* depict complex dynamics via deep dynamic layers, *(i)* extract group-distinct feature representations via grouped input layers, and *(iii)* preserve user-specific characteristics via the personalized output layer. It tackles the typical sparsity and imbalance problems in datasets for blood glucose modeling and offers an opportunity to build personalized blood glucose models for the general public based on limited personal measurements.
- We conduct extensive evaluations on both diabetic patients and non-diabetic people. Experimental results from a dataset covering 35 non-diabetic people, 38 type I and 39 type II diabetic patients in 7 months demonstrate that SugerMate yields an average accuracy of 82.14%, and outperforms traditional general learning, group-level learning, personalized learning and shallow learning algorithms in precision and recall.

In the rest of this paper, we review related works in Sec. 2, present an overview of SugerMate in Sec. 3, and elaborate on the detailed design and evaluation in Sec. 4 and Sec. 5, respectively. Finally we conclude in Sec. 6.

## 2 RELATED WORK

Research on inferring blood glucose concentrations or abnormal events dates back to the 1960s and continues to attract extensive research interest [37]. Physiological models [8, 21] mathematically formulate the whole process of glucose metabolism and are widely used for simulations and studies involving glucose regulation. One major drawback of physiological models is the requirement for prior knowledge to adjust the physiological parameters. Alternatively, researchers propose to combine machine learning techniques with a non-specific physiological model or directly correlating blood glucose levels with insulin, food intake and other inputs without physiological parameters. For instance, Plis *et al.* [38] apply a generic physiological model of blood glucose dynamics to extract features for support vector regression to infer blood glucose levels. Reymann *et al.* [42] replace the physiological model by an online simulator and bring blood glucose tracking on mobile platforms.

While physiological models and the underlying glucose metabolism dominate the dynamics of blood glucose, the impact of seemingly "secondary" factors, such as those related with individual's lifestyle, can be quite significant. Variations of the personalized external lifestyle factors such as meals, insulin or drug intake, exercises, and sleep quality, etc., can also lead to blood glucose dynamics, which are not captured by a universal physiological model [22]. Consequently, it is crucial to monitor these external lifestyle factors as inputs to improve the performance of blood glucose concentration inference. METABO [17] is a client-server architecture based system that records dietary, physical activity, medication and medical information for hypoglycaemic and hyperglycaemic event prediction. Marling *et al.* [31] improve hypoglycemia detection by combining CGM data with heart rate, galvanic skin response and skin temperature collected from a fitness band. However, these works all require CGM data as input, making them invasive and inconvenient for both patients and non-diabetic people.

Alternatively, there has been attempt at non-invasive blood glucose monitoring with pervasive wearable and mobile devices. Nguyen *et al.* [33] observe distinct patterns in ECG signals during hypoglycemia and hyperglycemia in type 1 diabetic patients. Sobel *et al.* [45] integrate five types of sensory data from two

accelerometers, a heat-flux sensor, a thermistor, two ECG electrodes and a galvanic skin response sensor to predict blood glucose concentration. Ranvier *et al.* [40] leverage ECG signals, and energy expenditure (estimated by an accelerometer and a breathing sensor) to detect hypoglycemic events. Our work is inspired by this body of research. We propose a smartphone-based non-invasive blood glucose monitoring system that jointly considers meals, drugs and insulin intake, physical activity and sleep quality without CGM data as inputs. Practically, physical activity level and sleep quality are automatically tracked without manual input, which notably improves the useability of our system.

Moreover, personalized blood glucose models are also important. It is because those models with generic parameters may not reflect user-specific factors, such as age, weight and insulin-to-carbohydrates ratio [37]. Both the physiological parameters and the impact of life events on blood glucose need to be trained on user-specific data to account for inter-person differences [3]. However, a primary impediment to build up such model is lacking of sufficient personalized blood glucose data [31]. In this paper, we advance previous works by carefully designing a multi-division deep learning framework that shares blood glucose information among groups of users, but preserves user-specific blood glucose characteristics via personalized learning, thus making full use of the limited, sometimes incomplete user-specific data, and achieving higher prediction accuracy than both generic learning and personalized learning.

## 3 OVERVIEW

SugerMate is a smartphone-based blood glucose level tracking system that *(i)* non-intrusively collects important external impacting factors and conducts feature engineering, *(ii)* efficiently trains a personalized blood glucose level model via a novel multi-task deep learning framework, and *(iii)* timely reminds users of abnormal blood glucose levels. Fig. 1 shows the architecture of SugerMate, which consists of three modules.

The **external factor collection and feature engineering** module first records external factors that are important to infer blood glucose concentration and convenient to input via smartphones. A user inputs meta information (*e.g.*, health status, age, weight), and records daily meal, drug and insulin intake. Meanwhile, SugerMate will automatically measures physical activities and sleep quality of the user via embedded sensors (*i.e.*, accelerometer, microphone and light sensor). After collecting data from multiple users, SugerMate conducts feature engineering from physiological and temporal perspectives, and feeds them into a multi-task deep learning framework: **Md$^3$RNN**. Md$^3$RNN first learns feature representations from users in the same group (non-diabetic, type I and type II diabetic), and then adopts a deep RNN layer to learn a general blood glucose level model on the dataset of all users. Finally it outputs a personalized blood glucose level model for each individual via a personality layer. The personalized inference results of blood glucose level at fine-grained time resolution are eventually shown in the **blood glucose level tracking** module, reminding the users of the abnormal levels of blood glucose. In SugerMate, we consider 4 blood glucose levels as in Table 1.

Table 1. Normal and abnormal blood glucose levels [50]. The normal blood glucose concentration ranges from 4.4 mmol/L to 6.1 mmol/L. The blood glucose can grow to 7.8 mmol/L after eating. Blood glucose above 7.8 mmol/L for a prolonged period indicates the risk of diabetes mellitus. Blood glucose below 4.4 mmol/L is a sign of hypoglycemia.

| Blood Glucose Value (mmol/L) | Glocose Level | Explanation |
|---|---|---|
| (0, 4.4] | Level 1 | Low blood glucose (hypoglycemia) |
| (4.4, 6.1] | Level 2 | Normal level of fasting blood glucose |
| (6.1, 7.8] | Level 3 | Normal level of postprandial blood glucose |
| (7.8, $+\infty$) | Level 4 | High blood glucose (hyperglycemia) |

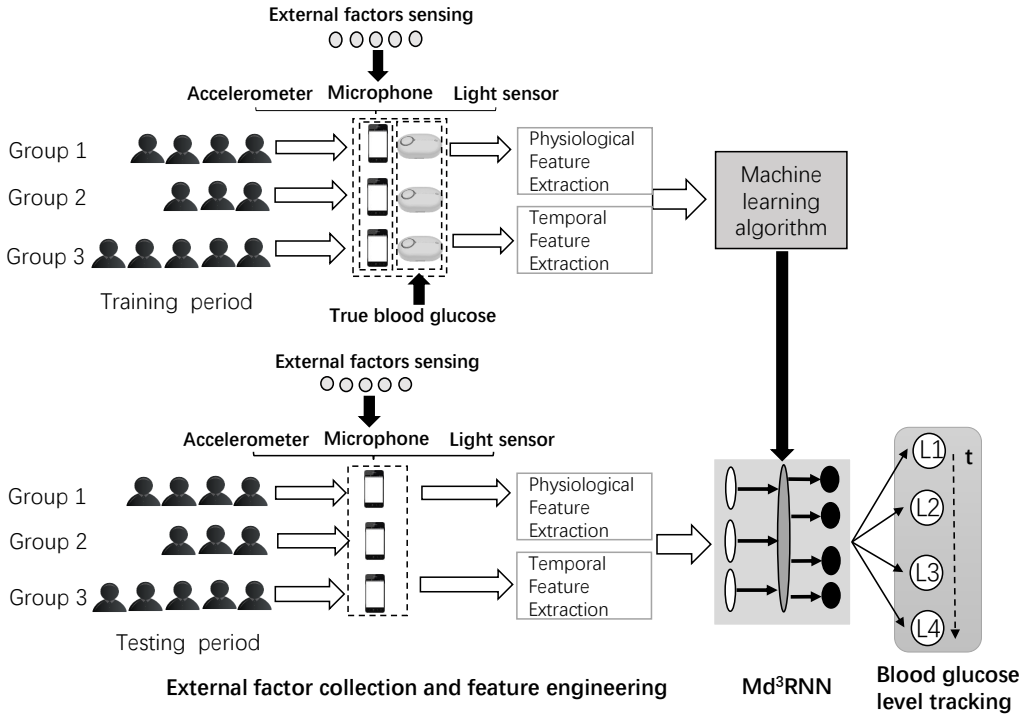


Fig. 1. Architecture of SugerMate.

## 4 DESIGN

This section presents the detailed design of SugerMate.

### 4.1 External Factor Sensing

It is neither possible nor necessary to exhaust a complete list of influential factors on blood glucose concentration. In SugerMate, five major external factors are measured, including food, drug and insulin intake, as well as physical activities and sleep quality.

**Food Intake.** As food intake is a main source of carbohydrate, SugerMate provides a food menu for users to keep track of their meals. Based on the carbohydrate food list [35], meals are categorized into five types, including grains, vegetables, milk and egg, fruits ,as well as meats. Users are asked to enter the food types and amounts of their meals, based on which SugerMate calculates $U_C$, the carbohydrate of a meal.

**Drug Intake.** Oral diabetic drugs enhance the secretion of insulin into the blood and are usually used by type II diabetic patients. In SugerMate, a drug menu of 11 common oral diabetes is presented for users to input their drug intake based on [5]. Users are required to select the drug name and record the dosage into SugerMate.

**Insulin Injection.** Inulin injection is widely used for blood glucose control for type I and type II patients. SugerMate provides an insulin type list based on [4] for users to record the usage and dosage of their insulin injection. SugerMate automatically transforms drug intake and insulin injection into the amount of acting insulin $U_I$ via bolus and basal rate information [38].

**Physical Activity.** Daily activities *e.g.*, exercises, consume the carbohydrate and affect blood glucose levels. In SugerMate, we adopt an efficient activity recognition scheme [25], which leverages the accelerometer of smartphones to automatically record six common physical activities (walking, running, going upstairs, going downstairs, sitting and standing) along with the corresponding durations. SugerMate then calculates the caloric expenditure based on the widely used calorie calculator [28, 32, 43].

$$CalorieBurn = (BMR/24) * MET * T, \tag{1}$$

where BMR (Basal Methobolic Rate) is the amount of energy required to simply sit or lie quietly, and MET (Metabolic Equivalent) is the ratio of the work metabolic rate to the resting metabolic rate. T is the activity duration time (in hours). SugerMate leverages the calorie expenditure $U_E$ as input for physiological feature extraction.

**Sleep Quality.** Sleep quality has a long-term influence on the blood glucose level [22]. SugerMate automatically measures sleep quality using smartphones as in [20]. It invokes the sensors (*i.e.*, accelerometer, microphone and light sensor) of smartphones to collect the ambient sleep-related factors, and then infersthe sleep quality by a statistics model. The output sleep quality score $U_S$ is then used for physiological feature extraction.

In summary, the five external factors are transformed into four categories of measurements including the carbohydrate of a meal ($U_C$), the amount of acting insulin ($U_I$), calorie expenditure ($U_E$) and sleep quality score ($U_S$), which are then used as input to extract important features for blood glucose level inference.

## 4.2 Feature Engineering

We extract features $X = \{X_P, X_T\}$ from external sensory data $U = \{U_C, U_I, U_E, U_S\}$ from both the physiological view ($X_P$, 10-dimension) and the temporal view ($X_T$, 51-dimension) to infer blood glucose levels.

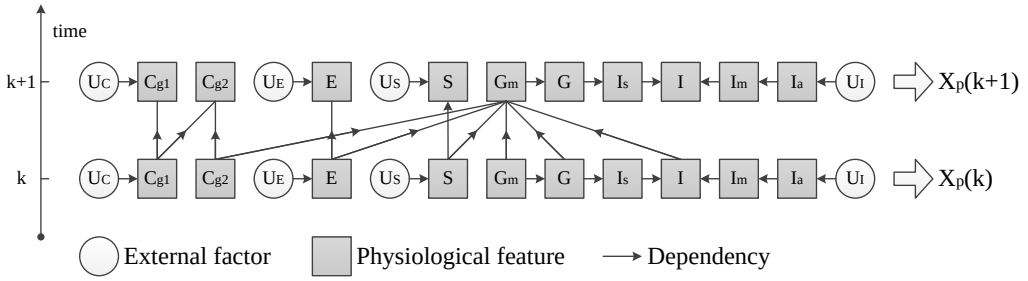


Fig. 2. An illustration of dependencies of physiological features $X_P = \{C_{g_1}, C_{g_2}, I_a, I_m, I_s, I, E, S, G_m, G\}$ and of external factors $U = \{U_C, U_I, U_E, U_S\}$.

*4.2.1 Features from Physiological View $X_P$.* Physiological features are essentially medical/biological indexes in a person's body, that describes the dynamics of glucose related variables [10, 38]. We extract physiological features based on the physiological model in [10], which characterizes carbohydrate dynamics, insulin dynamics, exercise dynamics, and glucose dynamics. We further include sleep dynamics, another important external factor that affects blood glucose levels [22]. Specifically, the following physiological features are extracted to represent blood glucose relevant dynamics.

- Features from *carbohydrate dynamics*: carbohydrate consumption $C_{g_1}$ and carbohydrate digestion $C_{g_2}$.
- Features from *insulin dynamics*: subcutaneous insulin absorption $I_a$, insulin secretion by pancreas $I_s$, insulin mass $I_m$ and active plasma insulin level $I$.
- Features from *exercise dynamics*: long-term effect of exercises on insulin $E$.

Table 2. Transition equations to calculate physiological features at time $k+1$ based on external factors measured at time $k+1$ and physiological features at time $k$ [10, 38].

| **Dynamic Process** | **Transition Equations** |
|---|---|
| Carbohydrate dynamics | $C_{g_1}(k+1) = C_{g_1}(k) - \alpha_1^c \times C_{g_1}(k) + U_C(k+1)$<br>$C_{g_2}(k+1) = C_{g_2}(k) + \alpha_1^c \times C_{g_1}(k) - \alpha_2^c \times C_{g2}(k)$ |
| Insulin dynamics | $I_a(k+1) = I_a(k) - \alpha^I_{f,r,m} \times I_a(k) + U_I(k+1)$<br>$I_s(k+1) = I_s(k)$<br>$I_m(k+1) = I_m(k) + \alpha^I_{f,r,m} \times I_a(k) + \alpha^I_a \times I_a(k) - \alpha^I_c \times I_m(k)$<br>$I(k) = I_m(k) \times S^I/(142 \times bm)$ |
| Exercise dynamics | $E(k-k_0+1) = (k-k_0) \times E(k-k_0) + U_E(k-k_0+1)$ |
| Sleep dynamics | $S(k-k_0+1) = (k-k_0) \times S(k-k_0) + U_S(k-k_0+1)$ |
| Glucose dynamics | $G_m(k+1) = G_m(k) + \delta_{abs} + \delta_{egp} - \delta_{ind} - \delta_{dep} - \delta_{clr}$<br>$G(k+1) = G_m(k+1)/(2.2 \times bm)$<br>$\delta_{abs} = \alpha_3^c \times \alpha_2^c \times C_{g_2}(k)$<br>$\delta_{egp} = \alpha_2^{egp} \times \exp(-I(k)/\alpha_3^{egp}) - \alpha_1^{egp} \times G(k)$<br>$\delta_{ind} = \alpha_1^{ind}/\sqrt{G(k)}$<br>$\delta_{dep} = \alpha_1^{dep} \times E(k) \times S(k) \times I(k)/(G(k) + \alpha_2^{dep})$<br>$\delta_{clr} = \alpha_1^{clr} \times (G(k) - \tau)$ |

**Note 1:** $S^I$ and $bm$ refer to the insulin sensitivity and body mass, respectively.
**Note 2:** All $\alpha$'s in the transition equations are user-specific parameters that need to be tuned per-person. The default values of parameters in the physiological model are set based on [10] and are further tuned for each person via 10-cross validations.
**Note 3:** We set the time duration $k - k_0$ to 24 hours in exercise dynamics and 7 days in sleep dynamics, which optimize our experimental results.

- Features from *sleep dynamics*: long-term effect of sleep quality on insulin $S$.
- Features from *glucose dynamics*: glucose mass $G_m$ and glucose concentration $G$.

The features are inter-dependent and are also related to the external factors. Fig. 2 illustrates the dependencies among variables and Table 2 summarizes the transition equations to calculate the physiological features at time $k+1$ using external factor measurements at time $k+1$ and historical physiological features at time $k$. The transition equations involve a set of parameters $\alpha$'s, e.g., the insulin sensitivity $S^I$ and the body mass $bm$. Note that it is generally difficult to determine the parameters in the transition equations. Therefore, the glucose concentration $G$ derived from the physiological model usually deviates from the actual blood glucose level. It can be viewed as a "nominal" blood glucose indicator averaged out from the overall population. Consequently, we exploit the glucose concentration $G$ derived from the physiological model as one dimension of the physiological features, which are fed into our Md$^3$RNN framework to learn the complex, personalized dynamics of blood glucose.

In summary, we extract a 10-dimension physiological feature vector $X_P$ for blood glucose level inference.

*4.2.2 Features from Temporal View $X_T$.* As blood glucose level naturally varies over time, we extract two temporal features for blood glucose level inference.

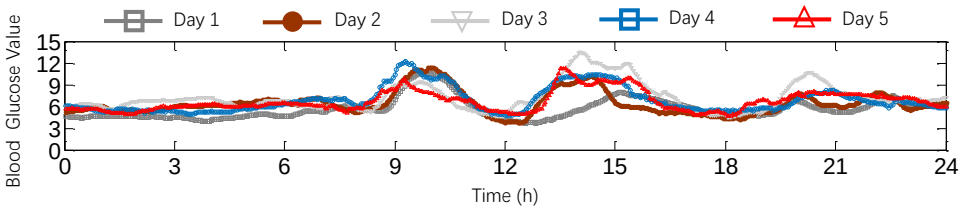


Fig. 3. Daily blood glucose traces of a volunteer.

- History blood glucose concentration $X_{T_1}$ at time $k$: Since people usually tend to lead a regular lifestyle, *e.g.*, having meals in the morning, at noon and in the evening and taking drugs at certain times, the blood glucose concentration also exhibits rough daily cycles. Fig. 3 plots the daily blood glucose traces of a volunteer for five successive days measured by a CGM device. As shown, the blood glucose trances always grow up signifcantly in the durations of 8:40 to 9:40, 12:30 to 14:00 due to the breakfast and lunch, but increase moderately from 18:00 to 20:00 because of taking the drugs before the dinner. This motivates us to adopt the historical blood glucose concentrations at time stamp $k$ (averaged over $D$ days) as one temporal feature to infer the blood glucose level at time $k$. In SugerMate, we set $D = 5$ and infer blood glucose level at a time resolution of 3min, which is in accordance with the time resolution of commercial CGM devices [29].
- Most recent physiological features $X_{T_2}$: As shown in the physiological models, the current blood glucose concentration is relevant to the recent blood glucose concentration and physiological features. However, in the physiological features $X_P$, only the physiological features at the last time stamp are considered. To account for more short-term temporal dependencies, we propose to include the $l$ most recent physiological features $X_{T_2}(k) = \{X_P(k), X_P(k-1), \ldots, X_P(k-l+1)\}$, where $l = 5$ in our implementation. That is, instead of considering the physiological features in the last 3min, we infer the current blood glucose level leveraging features in the last 15min.

In summary, we extract a 51-dimension $(1 + 10 \times 5)$ temporal feature vector $X_T$ for blood glucose level inference.

### 4.3 Blood Glucose Level Inference

Given the features extracted based on the physiological process, it seems plausible to perform any classification algorithm for blood glucose level inference. Nonetheless, this plug-and-play approach will neglect important information from (1) dynamics of the process, and (2) inter-user similarity among the same group of participants. Traditionally, various sequential methods, *e.g.*, hidden Markov model (HMM), recurrent neural networks (RNN) and time series models, are used to capture the temporal correlation of the input feature. The inter process correlations are often times incorporated with the so-called multi-task learning approaches [13], which learns processes (or tasks) in parallel to improve classification or to reduce the data sample requirement.

In this paper, a novel machine learning paradigm, namely Multi-division deep-dynamic RNN (Md$^3$RNN), is proposed. To include the the aforementioned information sources in an unified framework, we develop two key ideas that extend the classical RNN. Firstly, the single hidden layer in RNN is replaced with several deep stacked layers. The deep structure in the new model is able to describe complex, multi-scale dynamics that would otherwise be ignored (or averaged out) by prior "shallow" models. Secondly, the correlations among users, being quite significant within user groups (divisions), are encoded by group-shared input layer and common hidden layers, whereas the distinct characteristics of individual users are modeled with different output layers for personalized prediction. Within a larger scope of machine learning, the proposed Md$^3$RNN aims to leverage recent advancement of deep learning and multi-task learning, to model group-interacted time series data having

complex temporal dynamics. It can be viewed as both a deep extension of RNN, and an intermediate between single-task learning and multi-task learning, hence the name. Although we develop Md$^3$RNN for the specific use case of SugerMate, it is worth pointing out that it can be readily applied to many other applications dealing with grouped dynamic data. The overall configuration of the proposed model is summarized in Fig. 4. Detailed construction of each component is given in the sequel.

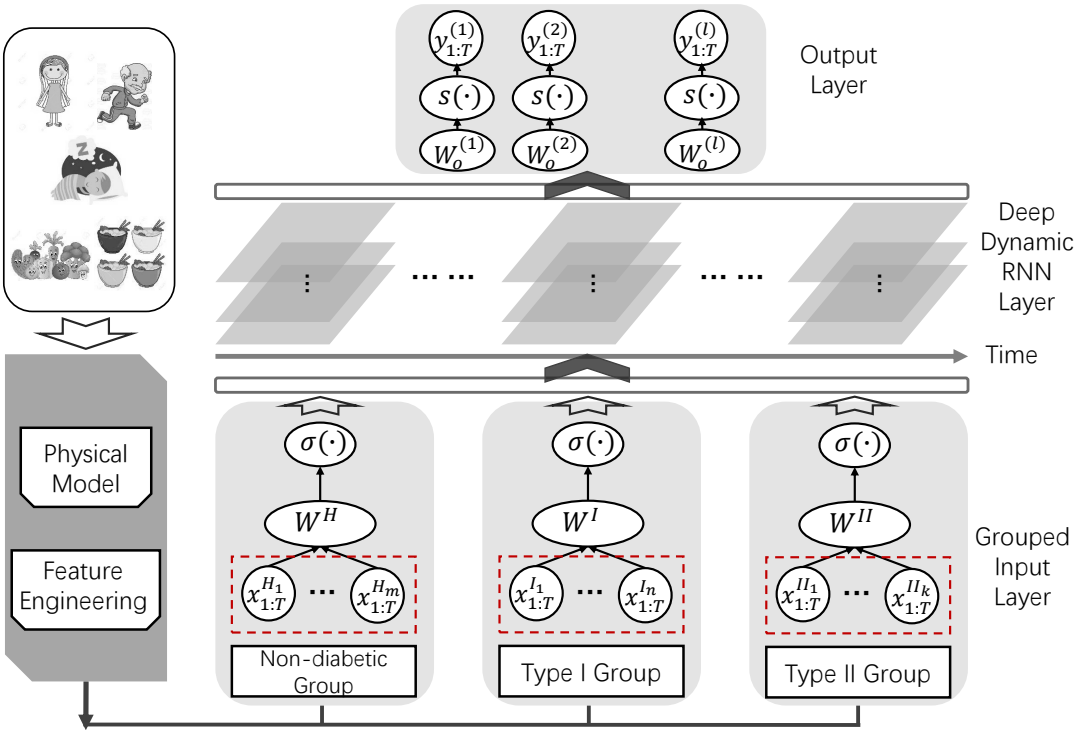


Fig. 4. Illustration of Md$^3$RNN structure.

*4.3.1 Model construction by layers.* The inputs of the Md$^3$RNN model are the features extracted following the discussion in the previous section. The labeled data sequences for user number $j$ at time $i$ are denoted by $(x_i^j, y_i^j)$. We also adopt an index set convention, that $(x_A^B, y_A^B)$ represents the data set $\left\{(x_i^j, y_i^j) | i \in A, j \in B\right\}$ given index sets $A$ and $B$.

**Grouped Input Layer.** In the context of blood glucose prediction, available inputs are naturally divided into three groups according to the health condition of the participant from whom the data was generated. Notation-wise, we utilize $H$, $I$ and $II$ to indicate the the group of non-diabetic user (healthy), user with type I diabetes and those having type II diabetes, respectively. Since the extracted features are essentially physiological indexes and temporally correlated variables, they must go though different transformations to represent useful information of three distinct groups. This consideration motivate the design of the input layer (bottom of Fig. 4) - it is divided into three units that performs different linear and non-linear transformation according to user groups. For instance, a

data sample $x_t^{I_j}$, generated at time $t$ from the $j^{th}$ user of type I, undergoes the following processing:

$$\tilde{x}_t^{I_j} = \sigma\left(W^I x_t^{I_j}\right) \tag{2}$$

where $W^I$ is the coefficients of the affine transformation [1], $\sigma$ is the sigmoid activation function, and $\tilde{x}_t^{I_j}$ is the output of the input layer for that data sample. Similar operations are conducted for data samples from group $H$ and $II$, but with different transformation coefficients. Intuitively, the shared transformation within groups would improve the learning of parameters (vs. single task learning), as information from all data in a homogeneous group is used. Also, the transformation can be stacked into several (say $P$) layers, for better information representation.

**Deep Dynamic Layer.** A common hidden layer is designated to capture the dynamics of the blood glucose evolution process. The underlying assumption is that, the physiological reactions governing blood glucose variation are similar for all people, despite of grouped behaviors in the representation of physiological indexes (input layer), or individual characteristics in exhibited glucose level. This assumption could be justified by a series of medical related research [8, 10, 21, 26]. Moreover, since all users share the same hidden layer, all collected data samples are eventually helping the estimation of its parameters. The availability of rich information for the hidden layer makes the learning of a deep structure possible. In SugerMate, a number of Long Short Term Memory (LSTM) networks are stacked together (middle of Fig. 4), to increase the overall model flexibility. It has been justified in both theory and practice that stacked LSTMs are able to capture dynamics occurring at different time scales [11], which in the current application would enable the modeling of both slow and rapid biological/chemical reactions. Although a wide variety of LSTM configuration exist in literature, in this work we adopt the one recently proposed by [24], which combines the forget/input gate and merges cell/hidden state for simplicity and better generalization performance. Mathematically, given the output from the grouped input layer, the deep dynamic layer performs

$$\begin{aligned}
z_t^d &= \sigma^d\left(W_z^d[h_{t-1}^d, h_t^{d-1}]\right) \\
r_t^d &= \sigma^d\left(W_r^d[h_{t-1}^d, h_t^{d-1}]\right) \\
\tilde{h}_t^d &= tanh\left(W_h^d[r_t^d * h_{t-1}^d, h_t^{d-1}]\right) \\
h_t^d &= (1 - z_t^d) * h_{t-1}^d + z_t^d * \tilde{h}_t^d
\end{aligned} \tag{3}$$

for hidden layer numbered $d = 1, 2, \cdots, D$. At the first dynamic layer with $d = 1$, the input $h_t^{d-1}$ is set to be the output from the grouped input layer, and the output of the last dynamic layer, $h_t^D$, will be used as the input of the last component of Md$^3$RNN.

**Personalized Output Layer.** Finally, each user is assigned a personalized output layer, parameterized by $W_o^j, j = 1, \cdots, l$, which performs a single linear and softmax transformation on the results of the deep dynamic layer. The particular configuration of the output layer compensates for the individual characteristics in the exhibited blood glucose (*i.e.*, measured blood glucose level). Because only data generated by a specific participant $j$ will have an effect on the its parameters $W_o^j$, the personalized output layer is set to have a "shallow structure", *i.e.*, it only performs the transformation once. More specifically, given $h_t^D$ from user $j$, it computes

$$\hat{y}_t^j = \text{softmax}\left(W_o^j h_t^D\right) \tag{4}$$

*4.3.2 Cost Sensitive Learning and Hyperparamter (Model) Selection.* Similar to other deep neural network learning, Md$^3$RNN is trained by minimizing the sum of losses over all the time steps. The definition of the loss function has much bearing on the generalization performance of the method. In particular for the current

[1] We assume that the interception is included in $W$. This can be done by simply adding a constant feature vector of 1.

application, simply minimizing a general error rate seems inappropriate, because the costs of different types of misclassifiction errors can differ a lot. For example, missing the detection of high blood glucose (type I or II) is more costly than misclassifying normal condition to an alarm for high glucose. Moreover, in the collected data set from real people, the training data is inherently imbalanced - the available samples labeled Level 1 and Level 4 are much fewer (only 30.6% in training dateset) compared to samples in the other categories (Level 2 and Level 3).

The above concerns motivate the cost sensitive learning of Md$^3$RNN. Instead of directly minimizing a surrogate of error rate, we propose to optimize over a weighted version of classification losses. More specifically, the following total loss function is considered:

$$L = \sum_t \sum_{y_t \in \mathcal{Y}} l(y_t, \hat{y}_t) C_{y_t} + \lambda^T \phi(W^H, W^I, W^{II}, W_z, W_r, W_h, W_o) \tag{5}$$

where $\mathcal{Y}$ is the label set, and $\hat{y}_t$s are prediction output from Md$^3$RNN. Our implementation uses cross entropy for $l(y_t, \hat{y}_t)$, but generally the "base" lost function $l(y_t, \hat{y}_t)$ can be any surrogate of the error function. The additional coefficient $C_{y_t}$ weights the misclassification error for category $y_t$. In the current application, $\mathcal{Y} = \{1, 2, 3, 4\}$, associated with four coefficients $C_1$ to $C_4$. Those cost weighting coefficients are treated as hyperparameters of the proposed model, but in other applications of Md$^3$RNN, they can also be determined with prior knowledge about the misclassification cost and the class imbalance. The second term in Eq.5 is a regularization for model parameters, which in the current implementation is substantiated with a frobenius norm for each matrix. The penalty coefficient $\lambda$s are left as model hyper-parameters, whose selection procedure will be discussed later.

With the technique of back-propagating, computing the gradient of Md$^3$RNN is not so different from the gradient calculation of classical RNN. In this work, we accomplish those computation using Tensorflow [1], and proceed to learn Md$^3$RNN model by stochastic gradient descent for overall loss minimization.

Last but not least, the construction of the Md$^3$RNN model involves choosing 15 hyperparameters, *e.g.*, cost coefficients $C_y$, depth $D$ of the stacked dynamic layer, learning rate, number of hidden unit in the input layer, etc. Direct application of cross validation (CV) for hyperparameter tuning, even with the help of parallel computing, seems intractable as the number of required CVs scales exponentially to the number of hyperparameters. In this regards, we adopt Bayesian optimization (BO), a recent tool developed for blackbox function optimization with limited evaluations. The decision variables of BO are those hyperparameters, and the objective is the F-score of the precision and recall on some testing data set. Note that BO has been used recently for the hyperparameter (model) selection of many deep learning paradigms [44][30].

## 5 EVALUATION

### 5.1 Experimental Settings

**Datasets.** We validate SugerMate on a dataset of 112 participants (35 non-diabetes, 38 type I diabetic patients and 39 type II diabetic patients) collected during July 2016 to January 2017. Each participant is equipped with (1) a WAVEGUIDER *U-Tang* CGM device [29] to record blood glucose concentration every 3 minutes and (2) a smartphone with SugerMate installed to collect external factors either automatically (activities and sleep quality) or manually (food, drug, and insulin intake). All participants agree to take measurements (*i.e.*, wear the CGM device and use SugerMate to record external factors) for at least 6 days, which is a normal disposable usage duration of the enzyme in the sensor of the CGM. Fig. 5 illustrates an example of data collection from a user. In total we obtain 762639 samples of blood glucose concentration and the corresponding external factors covering around 38132 hours. In brief, we collect the following categories of data:

- **Meta information.** We record basic personal data including gender, age, weight and health status to cover a wide range of users. Table 3 summarizes the basic information of the participants.


Fig. 5. An illustration of the equipments for data collection. Each participant wears a CGM device to record blood glucose concentration and uses a smartphone to collect external factors.

Table 3. Summary of participant information.

(a)

| Age (year) | # User |
|---|---|
| 15-24 | 8 |
| 25-34 | 17 |
| 35-44 | 24 |
| 45-54 | 29 |
| 55-64 | 34 |

(b)

| Weight (kg) | BMI ($kg/m^2$)[36] | # User |
|---|---|---|
| Underweight | (0, 18.5) | 18 |
| Normal weight | [18.5, 25) | 31 |
| Overweight | [25, 30) | 41 |
| Obese | [30, $+\infty$) | 22 |

(c)

| Gender | # User |
|---|---|
| Male | 57 |
| Female | 55 |

- **Blood glucose measurements.** We collect blood glucose measurements using commercial CGM devices for 6 to 30 days as labeled data. Table 4 summarizes the blood glucose measurements in our evaluation.
- **External factor measurements.** During measurements of blood glucose concentration, each participant manually inputs the times of their daily meal, drug and insulin intake. SugerMate automatically records activity levels and sleep quality as in Sec. 4.1. Fig. 5 shows the user interfaces to record external factors.

**Ground Truth.** We use the blood glucose concentrations collected by the CGM device as ground truth [2].
**Metrics.** We mainly adopt precision, recall and accuracy [18] to quantify the performance of SugerMate.

## 5.2 Inference Accuracy

*5.2.1 Overall Inference Accuracy.* Since all participants collected both measurements of CGM and external factors for at least 6 days, we use measurements during the former 5 days for training and the rest for testing.

[2]While clinical studies report that the precision and accuracy of commercial CGM devices still need improving [9, 46], they are sufficient as ground truth for the four normal and abnormal blood glucose levels.

Table 4. Summary of blood glucose measurements.

(a)

| Duration (days) | # User |
|---|---|
| 6-10 | 48 |
| 11-15 | 24 |
| 16-20 | 20 |
| 21-25 | 13 |
| 26-30 | 7 |

(b)

| Blood Glucose | # Sample |
|---|---|
| Level 1 | 75369 |
| Level 2 | 293530 |
| Level 3 | 235686 |
| Level 4 | 158054 |
| Total | 762639 |

Table 5 shows the overall performance of SugerMate. All results are averaged over the testing data. As shown, the recalls and the precisions for all the 4 blood glucose levels are above 79% and 73%, respectively. In particular, the recalls for Level 1 (low blood glucose) and Level 4 (high blood glucose) are 83.13% and 85.23%, even though the training data for Level 1 and Level 4 only take up 9.88% and 20.72% of the entire training set. This result shows that SugerMate can accurately infer low/high blood levels even with an imbalanced training dataset. Overall, SugerMate yields an accuracy of 82.14%, showing a promising performance to track blood glucose levels.

Table 5. Confusion matrix of SugerMate.

| Ground Truth | Inference | | | | | |
|---|---|---|---|---|---|---|
| | Level 1 | Level 2 | Level 3 | Level 4 | | |
| Level 1 | 62657 | 5521 | 3672 | 3519 | 83.13% | Recall |
| Level 2 | 16346 | 240584 | 27563 | 9037 | 81.96% | |
| Level 3 | 2660 | 30905 | 188472 | 13649 | 79.97% | |
| Level 4 | 3443 | 5620 | 14278 | 134713 | 85.23% | |
| | 73.62% | 85.12% | 80.55% | 83.72% | Accuracy: 82.14% | |
| | Precision | | | | | |

*5.2.2 Inference Result Analysis.* To understand the inference accuracy and the risks of different types of errors in the context of blood glucose management, we classify the inference results based on the principles of Clarke Error Grid Analysis (CEGA) [7]. The analysis classifies the inference results into correct event (Type A) and different types of errors (Type B to Type E) with increasing levels of severity. For instance, Type B errors are those that will not lead to inappropriate treatments, while Type E errors can lead to wrong treatment. Table 6 summarizes the percentages of each type of results. As shown, SugerMate will not cause inappropriate treatment (Type A and B) in almost 90% of the cases. It may lead to unnecessary worries or treatment (Type C) in 5.47% of the cases. In fewer than 5% of the cases, SugerMate will miss an abnormal blood glucose event (Type D) or confuse treatment (Type E). Therefore, SugerMate is suitable as an temporal alternative for CGM devices. However, we do not recommend SugerMate for extended duration of usage for patients serious diabetics, who need regular blood glucose management.

*5.2.3 Temporal View of Inference Results.* Fig. 6 plots the example inference results of SugerMate of three participants (one non-diabetic, one Type I diabetic patient, and one Type II diabetic patient) throughout a day. The errors are depicted at the bottom of each figure. As shown, the true blood glucose levels vary during the day after important daily activities such as food intake (5:50, 11:20 and 19:00 for the non-diabetic user; 6:00 and 16:50

Table 6. Inference result analysis.

| Type of Result | Explanation of Result | Percentage |
|---|---|---|
| Type A | The inference value is consistent with the true value.<br>(*i.e.*, the inference blood glucose level is correct.) | 82.14% |
| Type B | The inference result would not lead to inappropriate treatment.<br>(*i.e.*, Level 2 is predicted as Level 3, or vice-versa.) | 7.67% |
| Type C | The inference result will lead to unnecessary treatment.<br>(*i.e.*, Level 2 is predicted as Level 1/4,or Level 3 is predicted as Level 1/4.) | 5.47% |
| Type D | Fail to detect hypoglycemia or hyperglycemia.<br>(*i.e.*, Level 1/4 are predicted as Level 2/3.) | 3.81% |
| Type E | The predicted results that would confuse treatment by mistaking hypoglycemia for hyperglycemia or vice-versa.<br>(*i.e.*, Level 1 is predicted as Level 4, and vice-versa.) | 0.91% |

for the type I diabetic user; 6:00, 12:50 and 17:45 for the type II user), insulin injection (7:40 for the type I diabetic user), drug intake (15:10 for the type II user) and exercises (15:30 for the type II user), indicating the importance of external factors. The blood glucose levels inferred by SugerMate also match the true blood glucose levels most of the time, which validates the effectiveness of SugerMate during various daily activities.

Most errors mistake blood glucose levels 2 and level 3, and usually occur during the transition of two blood glucose levels (*e.g.*, from 6:10 to 6:30 for the type II user), or in case of sudden blood glucose concentration change (*e.g.*, at 2:30 for the non-diabetic user and at 0:30 for the type I user). Errors during blood glucose transitions are mainly brought by the temporary delays to measure the external factors. In this case, the inference result will follow up the true value in a short time. Errors of sudden blood glucose fluctuation are often resulted by the abrupt or tiny changes of the external factors, which may not be immediately detected by SugerMate. In this case, the durations of errors are too short to cause risky emergencies. Moreover, most errors belong to the results of Type B in 5.2.2, which would not lead to inappropriate treatment.

## 5.3 Model Comparison

*5.3.1 Effectiveness of Md$^3$RNN Framework.* To demonstrate the effectiveness of the multi-division framework in making full use of the training dataset, we evaluation Md$^3$RNN from two perspectives.

**Layer contribution analysis.** To evaluate the effect of different layers, we conduct blood glucose level inference with three combinations of layers.

- *Deep dynamic layer.* Training without considering differences in groups and persons, and only output a general model.
- *Deep dynamic layer + Grouped input layer.* Learn group-specific feature representations but ignore per-person characteristics in the output.
- *Deep dynamic layer + Grouped input layer + personalized output layer (Md$^3$RNN).* Efficiently learn features from different groups and output personalized inference results.

Fig. 7 plots the comparison results of the three combinations. As shown, both the precisions and recalls increase with more layers, with an improvement of 21.13% in average precision and 18.57% in average recall, respectively. Moreover, the standard deviations drop remarkably from 17.25% to 10.25% of average precision, and from 20.75% to 10.75% of average recall. The results demonstrate the effectiveness of Md$^3$RNN, which learns representative features from the same groups and considers individual differences in blood glucose level inference.

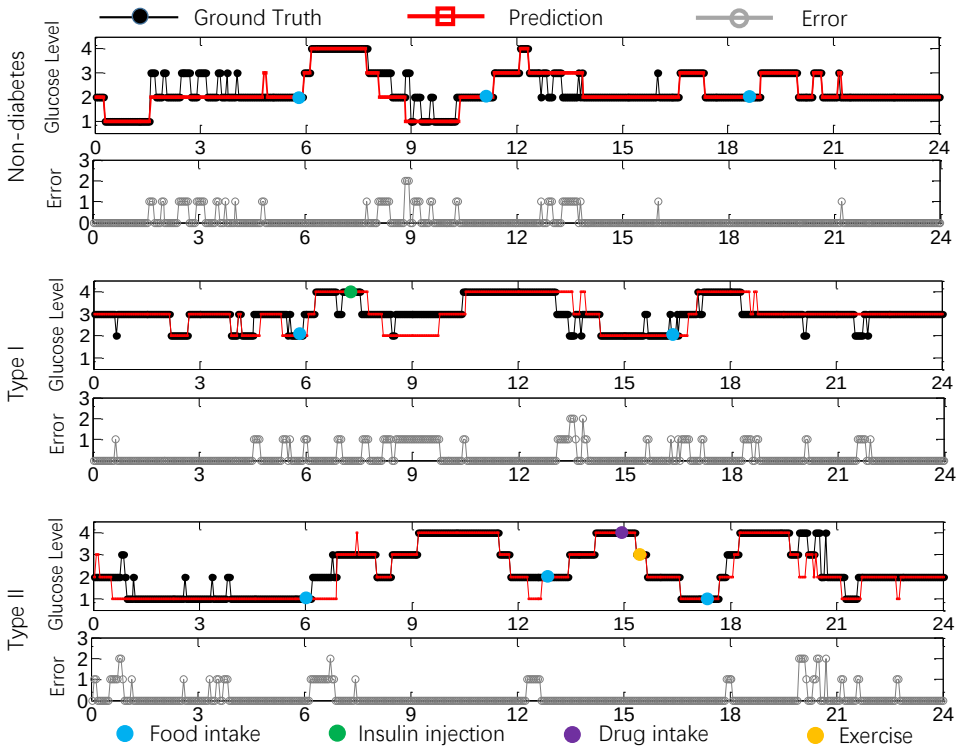


Fig. 6. Traces of blood glucose level inference results throughout a day.

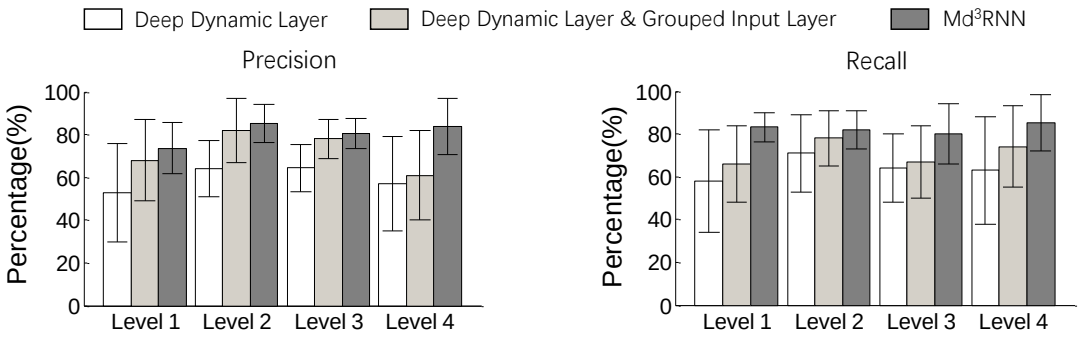


Fig. 7. Performance of layer combinations.

**Comparison of data sharing schemes.** To demonstrate the benefits of sharing data and knowledge among groups and users, we compare Md$^3$RNN with other learning frameworks with different data sharing schemes.

- *General Learning.* All the training data are directly fed into the model (*i.e.*, deep RNN) for training indifferently. General learning results in a *generic* model that assumes universal correlations between all inputs and the blood glucose levels.

- *Group Learning.* The data of users belonging to a same group are fed into a model (*i.e.*, deep RNN) for training. Three separate models are obtained for three groups (*i.e.*, non-diabetic, type I and type II diabetic). The group learning results in a *group* model that shares the general characteristics of users within the same group but without data sharing among users in different groups.
- *Single Learning.* We train a different model (*i.e.*, deep RNN) for each individual participant by feeding his/her own measurements into the model. Single learning results in a *personalized* model without sharing data and learning knowledge from measurements of other participants.

Fig. 8 shows the overall precisions and recalls of our Md$^3$RNN as well as *General learning*, *Group learning* and *Single learning*. As shown, our multi-divisional learning framework (Md$^3$RNN) performs best among the four learning approaches with an average precision of 80.75% and an average recall of 82.57%. It also yields the lowest standard deviations (17.18% of average precision and 17% of average recall). The results show that Md$^3$RNN is both effective and stable in blood glucose level inference.

General learning treats each sample of training data equally, and ignores the individual differences, so it performs poorly in most cases. Conversely, single learning approach encodes the individual characteristics but suffers from lacking of user-specific training dataset. Even though group learning learns the similarities of users within the same group, it ignores inter-person physiological differences. Md$^3$RNN combines the advantages of these three learning approaches, which makes better use of the limited individual training data by sharing measurements among users and preserves user-specific characteristics via the personal learning layer.

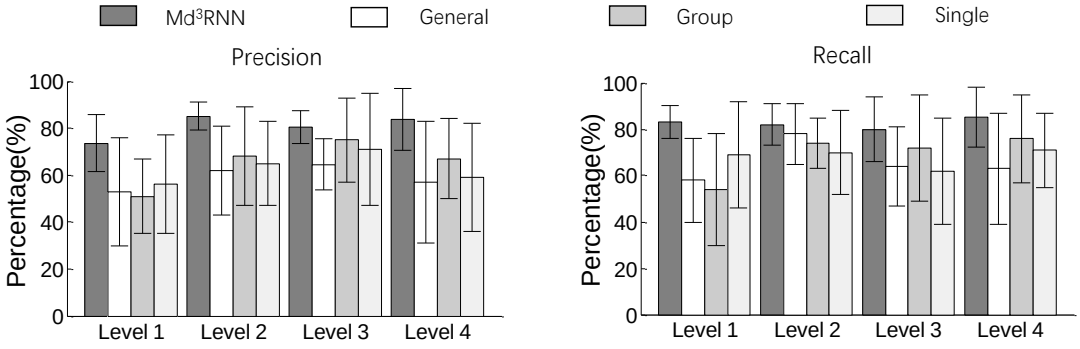


Fig. 8. Performance comparison of different data sharing schemes.

*5.3.2 Effectiveness of Md$^3$RNN Learning Algorithm.* To demonstrate the effectiveness of adopting Md$^3$RNN learning algorithms over the shallow, deep or multi-task learning algorithms, we compare it with the following advanced algorithms.

- **Random Forest (RF) [27].** As another ensemble method, RF combines many simple decision trees together and output the mode of classes for prediction. To avoid correlation among base trees, random set of features are selected in the splitting process when constructing each decision tree. For implementation, we adopt the conditional inference tree algorithm in the Party package.
- **Artificial neural network (ANN) [48].** We also included the classical ANN as a baseline, simply to justify the benefit of "structure engineering" from Md$^3$RNN. The ANN under comparison contains a single input layer, three hidden layers, and an output layer. The training of ANN is done by using the stochastic gradient descend algorithm implemented in Tensorflow [1].
- **Multi-task Support Vector Machine (mSVM).** SVM [47] bases on the idea of optimal separating hyperplane that maximizes the separation margin of two data groups (classes). Due to this construction,

it usually generalizes well, and its dual form is a quadratic programing that can be easily incorporated with kernels. We use a multi-task version of the SVM classifier proposed in [13], which incorporate the relation between tasks through a task-coupled kernel function. To eliminate scale/location discrepancies among input variables, all features are normalized before being used in the training phase.

- **Gradient Boosting (GB) [16].** GB generates a prediction model by combining many weak classifiers into a stronger classification committee. We use the implementation of the fastAdaboost package to combine basic tree classifiers for ensemble learning.
- **Multi-task Gaussian Processes (mGP).** Instead of directly parameterizing a latent function for classification, GP [41] models it with a generic Gaussian process. The posterior of the process is updated with training data set, and is "squashed" through a logistic function for classification. The multi-task version of GP can be achieved by redefining the kernel matrix to include task similarities. We adopt the multi-task GP proposed in [2], and also implement several approximation algorithms for acceleration [6].
- **Nested Dirichlet Process infinite Hidden Markov model (nDP-iHMM).** Being a classical example of dynamic Bayesian networks, HMM assumes that the observed process is driven by a hidden (unobserved) Markovian process [39]. Simple as it is, HMM is widely used in signal processing and time series analysis due to its interpretability and tractability. To further improve the model flexibility, the authors of [34] proposed using the non-parametric method for possibly undetermined state space, and imposing a nested Dirichlet process prior to share information among tasks.

Fig. 9 illustrates the results. Apparently, Md$^3$RNN achieves best performance on both precisions and recalls. More specifically, it outperforms the runner-up by at least 20% in terms of average precision, and yields much better recalls for the categories of interest, *i.e.*, level 1 and level 4. Among those baselines, it appears that no method could dominate the others, except that nDP-iHMM performs slightly better in terms of recall score. This is mainly because nDP-iHMM is the only method among baselines that allows both temporal correlation and information sharing among tasks. However, compared to Md$^3$RNN, which is able to describe multi-scale dynamics, nDP-iHMM is still worse in general. The dominating performance of Md$^3$RNN is somewhat expected, as those baselines either ignore the multi-scale dynamics of the observed data, or can not allow information sharing among available data from users. The above observation further justifies the efforts of designing the new machine learning paradigm for SugerMate, which efficiently transfers valuable knowledge between individuals.

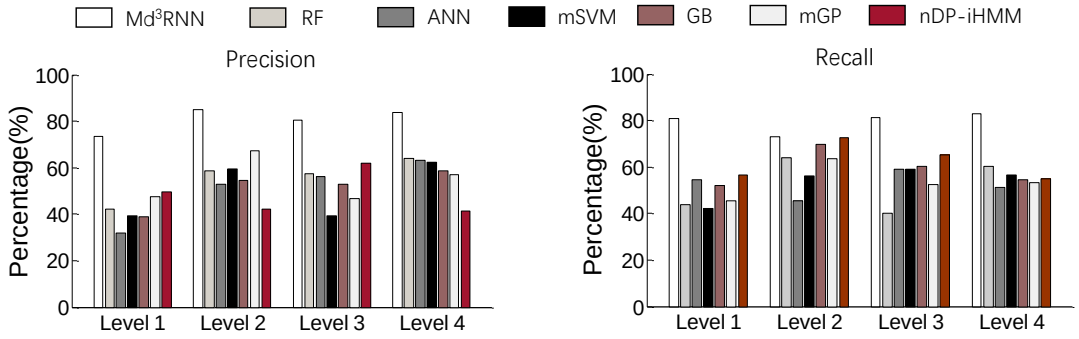


Fig. 9. Performance comparison with shallow learning algorithms.

## 5.4 Micro-benchmarks

*5.4.1 Effectiveness of Features.* Table 7 shows the average precisions and recalls for all the 4 blood glucose levels with different combinations of features. By combining physiological features ($X_P$) with temporal features ($X_T$), the

average precision and recall of the 4 blood glucose levels improve by at least 30% and 40% respectively. Specifically, all of the three features bring in notable improvements in detecting the blood glucose events. This is because $X_P$ from the physiological model can well reflect the impact of external factors (*e.g.*, food and exercise), and $X_{T1}$ from the historical trend enables to track the individual temporal dynamics of average blood glucose concentrations in recent time. As for $X_{T2}$, it brings in more short-term temporal dependencies between physiological features and the blood glucose levels. In general, the precisions and recalls of 4 levels are increasing smoothly with adding each feature, demonstrating the effectiveness of the selected features.

Table 7. Effectiveness of features.

| | Level 1 | | Level 2 | | Level 3 | | Level 4 | |
|---|---|---|---|---|---|---|---|---|
| **Features** | **Precision** | **Recall** | **Precision** | **Recall** | **Precision** | **Recall** | **Precision** | **Recall** |
| $X_P$ | 43.37% | 32.82% | 46.03% | 39.10% | 51.79% | 48.95% | 56.30% | 43.49% |
| $X_P$+$X_{T_1}$ | 58.29% | 63.60% | 72.15% | 60.17% | 68.84% | 61.49% | 73.23% | 66.74% |
| $X_P$+$X_{T_1}$+$X_{T_2}$ | 73.62% | 83.13% | 85.12% | 81.96% | 80.55% | 79.97% | 83.72% | 85.23% |

*5.4.2 Impact of amount of training samples.* In this experiment, we evaluate the performance of SugerMate with increasing numbers of training samples. Since the duration of measurements for each participant varies from 6 to 30 days, we use measurements of 5 to 25 days for training, and the rest for testing. Note that we keep the measurements for training but exclude them for testing if the duration of certain user's measurements is insufficient. For example, if the user's measurements last for 7 days, we use his measurement to evaluate the performance of using 5 days of training data, and test on the measurements of the remaining 2 days. However, when evaluating the performance with 10 days of training data, we only use his 7 days of measurements for training, but not for testing.

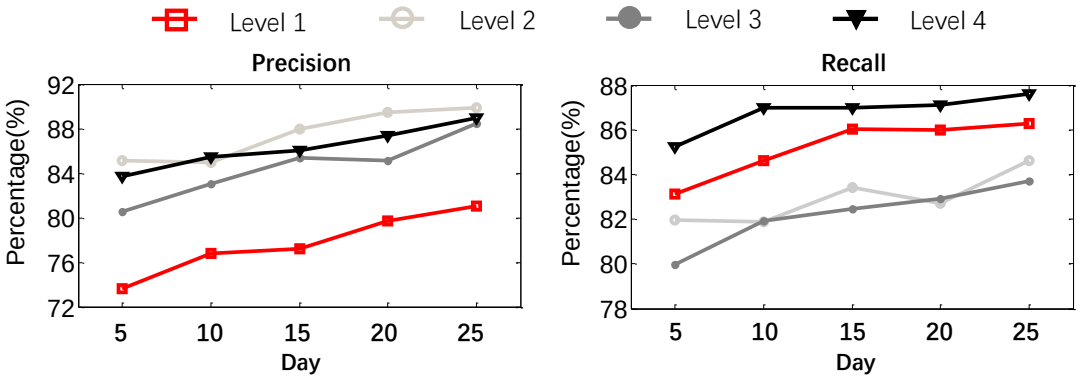


Fig. 10. Impact of increasing amount of training samples.

Fig. 10 illustrates the results for all the 4 blood glucose levels. The results are averaged over all testing samples as in previous evaluations. As expected, the precisions and recalls for all the 4 blood glucose levels improve smoothly with the increase of training samples. The results verify that the challenge (and our motivation to adopt a multi-task learning framework) is the lack of training data. Note that SugerMate is not a replacement of the current CGM devices, but rather, a complement when CGM devices are uncomfortable or inconvenient to wear. Therefore we envision the training dataset will grow gradually after wearing the CGM device multiple times (at least for diabetes patients), and the overall accuracy will also improve over time as a result.

*5.4.3 Impact of Temporal Gaps.* The blood glucose concentration is correlated with the previous blood glucose levels because of the control loop of the glucose metabolism [8, 21, 37]. Since SugerMate does not rely on the previous blood glucose level as an input, it is natural that the accuracy of SugerMate will degrade if there is a long gap between the training and the testing datasets (*i.e.*, the training dataset can be outdated).

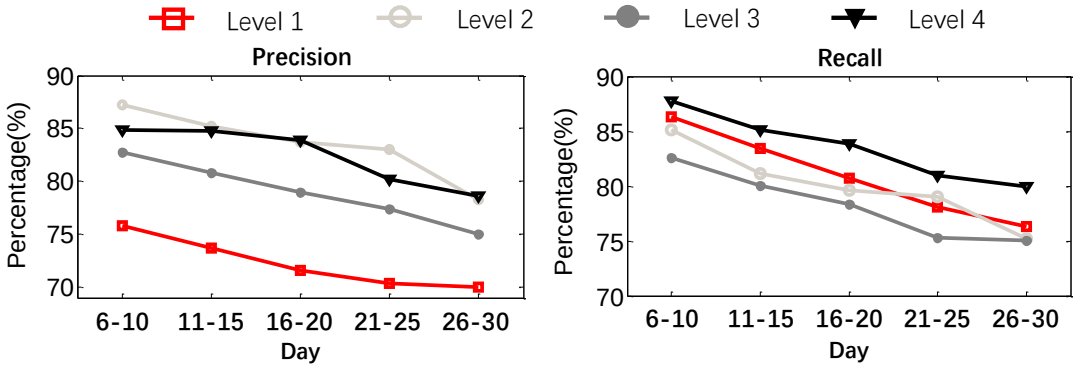


Fig. 11. Impact of temporal gaps between the training and testing datasets.

Fig. 11 plots the overall performance by training using the same 5 days of measurements, and testing on measurements collected on the 6-10th, 11-15th, 16-20th, 21-25th, and 26-30th days, respectively. As expected, both the precisions and recalls drop moderately with the increase of temporal gaps between the training and the testing datasets, with a maximum decrease of 6.73% and 7.02% in average precision and recall after 21-25 days. Note that SugerMate is not designed as a replacement of the commercial CGM devices, but rather a ubiquitous temporary alternative when CGM devices are uncomfortable or inconvenient to wear. From the results, we recommend SugerMate users to put on the CGM device to monitor the blood glucose at least every three weeks. The data sampled by the CGM will automatically feed into SugerMate for a model retraining.

## 6 CONCLUSION

Inferring blood glucose levels is important to avoid health risks incurred by hyperglycemia and hypoglycemia. Commercial continuous blood glucose monitoring devices can be invasive and inconvenient to wear, which degrades the quality of life for diabetic patients and makes them inaccessible to non-diabetic people. We present SugerMate, a ubiquitous blood glucose level inference system using commodity smartphones. It measures important external factors that affect blood glucose concentration and adopts machine learning to infer blood glucose levels at a fine-grained time resolution. The core of SugerMate is a novel learning paradigm, Md$^3$RNN, which (1) depicts complex glucose dynamics via a deep model, (2) extracts generic feature representations with a grouped multi-task framework, and (3) preserves individual differences using personalized outputs. It tackles the sparsity and imbalance problem, which is the main hurdle of high-accurate personalized blood glucose level tracking. We deploy SugerMate to 112 users and collect measurements for over 7 months. Evaluations show that Md$^3$RNN outperforms the state-of-the-arts methods on the blood glucose level inference. With fully automatic recording of external factors in the future, we envision SugerMate as a user-friendly and reliable complement for continuous blood glucose monitoring in daily life.

## REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Edwin V Bonilla, Kian Ming Adam Chai, and Christopher KI Williams. 2007. Multi-task Gaussian process prediction.. In *NIPs*, Vol. 20. 153–160.

[3] Razvan Bunescu, Nigel Struble, Cindy Marling, Jay Shubrook, and Frank Schwartz. 2013. Blood glucose level prediction using physiological models and support vector regression. In *Proceedings of IEEE International Conference on Machine Learning and Applications*, Vol. 1. 135–140.

[4] Joslin Diabetes Center. 2017. Insulin A to Z: A Guide on Different Types of Insulin. (2017). http://www.joslin.org/info/insulin_a_to_z_a_guide_on_different_types_of_insulin.html.

[5] Joslin Diabetes Center. 2017. Oral Diabetes Medications Summary Chart. (2017). http://www.joslin.org/info/oral_diabetes_medications_summary_chart.html.

[6] Krzysztof Chalupka, Christopher KI Williams, and Iain Murray. 2013. A framework for evaluating approximation methods for Gaussian process regression. *Journal of Machine Learning Research* 14, Feb (2013), 333–350.

[7] William L Clarke. 2005. The original Clarke error grid analysis (EGA). *Diabetes Technology & Therapeutics* 7, 5 (2005), 776–779.

[8] Chiara Dalla Man, Robert A Rizza, and Claudio Cobelli. 2007. Meal simulation model of the glucose-insulin system. *Transactions on Biomedical Engineering* 54, 10 (2007), 1740–1749.

[9] Carlos Eduardo Ferrante do Amaral and Benhard Wolf. 2008. Current development in non-invasive glucose monitoring. *Medical engineering & physics* 30, 5 (2008), 541–549.

[10] David L Duke. 2010. *Intelligent diabetes assistant: A telemedicine system for modeling and managing blood glucose.* PhD thesis. Carnegie Mellon University.

[11] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075* (2015).

[12] Margaret Eckert-Norton and Susan Kirk. 2013. Non-diabetic hypoglycemia. *The Journal of Clinical Endocrinology & Metabolism* 98, 10 (2013), 39A–40A.

[13] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 109–117.

[14] International Diabetes Federation. 2015. IDF Diabetes Atlas. 7th ed. (2015). http://www.diabetesatlas.org/.

[15] Philip Felig, Ali Cherif, Akira Minagawa, and John Wahren. 1982. Hypoglycemia during prolonged exercise in normal men. *New England Journal of Medicine* 306, 15 (1982), 895–900.

[16] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (2002), 367–378.

[17] Eleni Georga, Vasilios Protopappas, Alejandra Guillen, Giuseppe Fico, Diego Ardigo, Maria Teresa Arredondo, Themis P Exarchos, Demosthenes Polyzos, and Dimitrios I Fotiadis. 2009. Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes modeling and management system. In *Proceedings of IEEE Annual International Conference of the Engineering in Medicine and Biology Society*. 5633–5636.

[18] Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Advances in information retrieval*. Springer, 345–359.

[19] DCCT Research Group and others. 1991. Epidemiology of severe hypoglycemia in the Diabetes Control and Complications Trial. *The American Journal of Medicine* 90, 4 (1991), 450–459.

[20] Weixi Gu, Zheng Yang, Longfei Shangguan, Wei Sun, Kun Jin, and Yunhao Liu. 2014. Intelligent sleep stage mining service with smartphones. In *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 649–660.

[21] Roman Hovorka, Valentina Canonico, Ludovic J Chassin, Ulrich Haueter, Massimo Massi-Benedetti, Marco Orsini Federici, Thomas R Pieber, Helga C Schaller, Lukas Schaupp, Thomas Vering, and others. 2004. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement* 25, 4 (2004), 905.

[22] Shingo Iwasaki, Junji Kozawa, Kenji Fukui, Hiromi Iwahashi, Akihisa Imagawa, and Iichiro Shimomura. 2015. Coefficient of variation of R-R interval closely correlates with glycemic variability assessed by continuous glucose monitoring in insulin-depleted patients with type 1 diabetes. *Diabetes Research and Clinical Practice* 109, 2 (2015), 397–403.

[23] Ronald Klein. 1995. Hyperglycemia and microvascular and macrovascular disease in diabetes. *Diabetes Care* 18, 2 (1995), 258–268.

[24] Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber. 2014. A clockwork rnn. *arXiv preprint arXiv:1402.3511* (2014).

[25] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity Recognition Using Cell Phone Accelerometers. *SIGKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[26] ED Lehmann and T Deutsch. 1992. A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *Journal of biomedical engineering* 14, 3 (1992), 235–242.

[27] Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[28] HealthStatus LLC. 2017. Calories Burned Calculator. (2017). https://www.healthstatus.com/calculate/cbc.

[29] Shenzhen Waveguider Optical Telecom Technology lnc. 2015. Continuous Glucose Monitor of WAVEGUIDER. (2015). http://www.waveguidertech.com/html/071190.html.

[30] Gustavo Malkomes, Charles Schaff, and Roman Garnett. 2016. Bayesian optimization for automated model selection. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). 2892–2900.
[31] Cindy Marling, Lijie Xia, Razvan Bunescu, and Frank Schwartz. 2016. Machine Learning Experiments with Noninvasive Sensors for Hypoglycemia Detection. In *Proceedings of IJCAI Workshop on Knowledge Discovery in Healthcare Data*.
[32] Inc. MyFitnessPal. 2017. Calorie Counter - MyFitnessPal. (2017). https://play.google.com/store/apps/details?id=com.myfitnesspal.android&hl=en.
[33] Linh Lan Nguyen, Steven Su, and Hung T Nguyen. 2012. Identification of hypoglycemia and hyperglycemia in type 1 diabetic patients using ECG parameters. In *Proceedings of IEEE Annual International Conference of the Engineering in Medicine and Biology Society*. 2716–2719.
[34] Kai Ni, Lawrence Carin, and David Dunson. 2007. Multi-task learning for sequential data via iHMMs and the nested Dirichlet process. In *Proceedings of the 24th international conference on Machine learning*. ACM, 689–696.
[35] The Regents of the University of Michigan. 2012. Diabetes: Carbohydrate Food List. (2012). http://www.med.umich.edu/1libr/MEND/CarbList.pdf.
[36] World Health Organization and others. 2013. BMI classification 2013. *WHO [cited 2014 June 30]. Available from: apps. who. int/bmi/index. jsp* (2013).
[37] Silvia Oviedo, Josep Vehí, Remei Calm, and Joaquim Armengol. 2016. A review of personalized blood glucose prediction strategies for T1DM patients. *International Journal for Numerical Methods in Biomedical Engineering* (2016).
[38] Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, and Frank Schwartz. 2014. A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management. In *Proceedings of AAAI Workshop on Modern Artificial Intelligence for Health Analytics*.
[39] Lawrence Rabiner and B Juang. 1986. An introduction to hidden Markov models. *ieee assp magazine* 3, 1 (1986), 4–16.
[40] Jean-Eudes Ranvier, Fabien Dubosson, Jean-Paul Calbimonte, and Karl Aberer. 2016. Detection of hypoglycemic events through wearable sensors. In *Proceedings of the International Workshop on Semantic Web Technologies for Mobile and PErvasive Environments*.
[41] Carl Edward Rasmussen. 2006. Gaussian processes for machine learning. (2006).
[42] Maximilian P Reymann, Eva Dorschky, Benjamin H Groh, Christine Martindale, Peter Blank, and Bjoern M Eskofier. 2016. Blood glucose level prediction based on support vector regression using mobile platforms. In *Proceedings of IEEE Annual International Conference of the Engineering in Medicine and Biology Society*. 2990–2993.
[43] Inc. ShapeSense. 2016. Activity Based Calorie Burn Calculator. (2016). http://www.shapesense.com/fitness-exercise/calculators/activity-based-calorie-burn-calculator.aspx.
[44] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.
[45] Sandra I Sobel, Peter J Chomentowski, Nisarg Vyas, David Andre, and Frederico GS Toledo. 2014. Accuracy of a novel noninvasive multisensor technology to estimate glucose in diabetic subjects during dynamic conditions. *Journal of Diabetes Science and Technology* 8, 1 (2014), 54–63.
[46] Santhisagar Vaddiraju, Diane J Burgess, Ioannis Tomazos, Faquir C Jain, and Fotios Papadimitrakopoulos. 2010. Technologies for continuous glucose monitoring: current problems and future promises. *Journal of Diabetes Science and Technology* 4, 6 (2010), 1540–1562.
[47] Lipo Wang. 2005. *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media.
[48] Sun-Chong Wang. 2003. Artificial neural network. In *Interdisciplinary computing in java programming*. Springer, 81–100.
[49] Wikipedia. 2016. Blood glucose monitoring. (2016). https://en.wikipedia.org/wiki/Blood_glucose_monitoring.
[50] Wikipedia. 2017. Blood sugar. (2017). https://en.wikipedia.org/wiki/Blood_sugar.

Received February 2017