Artificial Intelligence and Machine Learning for Improving Glycemic Control in Diabetes: Best Practices, Pitfalls, and Opportunities

Article *in* IEEE Reviews in Biomedical Engineering · November 2023

DOI: 10.1109/RBME.2023.3331297

Peter Jacobs
Oregon Health and Science University
135 PUBLICATIONS **2,613** CITATIONS

Andrea Facchinetti
University of Padova
220 PUBLICATIONS **5,276** CITATIONS

Josep Vehí
Universitat de Girona
247 PUBLICATIONS **3,707** CITATIONS

Marc D Breton
University of Virginia
225 PUBLICATIONS **8,338** CITATIONS

Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls and opportunities

Peter G. Jacobs, *Member, IEEE*, Pau Herrero, Andrea Facchinetti, Josep Vehi, Boris Kovatchev, Marc Breton, Ali Cinar, *Life Senior Member, IEEE*, Konstantina Nikita, *Fellow IEEE*, Frank Doyle Jr., III, Jorge Bondia, Tatej Battelino, Jessica R. Castle, Konstantia Zarkogianni, *Member IEEE*, Rahul Narayan, and Clara Mosquera-Lopez *Member, IEEE*

Abstract— ***Objective:*** **Artificial intelligence and machine learning are transforming many fields including medicine. In diabetes, robust biosensing technologies and automated insulin delivery therapies have created a substantial opportunity to improve health. While the number of manuscripts addressing the topic of applying machine learning to diabetes has grown in recent years, there has been a lack of consistency in the methods, metrics, and data used to train and evaluate these algorithms. This manuscript provides consensus guidelines for machine learning practitioners in the field of diabetes, including best practice recommended approaches and warnings about pitfalls to avoid. *Methods:* Algorithmic approaches are reviewed and benefits of different algorithms are discussed including importance of clinical accuracy, explainability, interpretability, and personalization. We review the most common features used in machine learning applications in diabetes glucose control and provide an open-source library of functions for calculating features, as well as a framework for specifying data sets using data sheets. A review of current data sets available for training algorithms is provided as well as an online repository of data sources. *Significance:* These consensus guidelines are designed to improve performance and translatability of new machine learning algorithms developed in the field of diabetes for engineers and data scientists.**

Index Terms— **diabetes, machine learning, artificial intelligence, deep learning, decision support, automated insulin delivery, glucose prediction, data science, feature engineering**

I. Introduction

A. Diabetes and its complications

Type 1 diabetes (T1D) is an autoimmune metabolic disorder whereby the beta cells within the pancreas are destroyed and are no longer able to produce insulin [1]. People living with T1D must therefore take exogenous insulin to enable their body to utilize glucose in the blood [2]. Without exogenous insulin, glucose levels in the blood can become dangerously high, which can be toxic and can lead to long term damage to tissue including diabetic retinopathy, neuropathy, cardiovascular disease, and limb loss [3]. Exogenous insulin delivery poses risk as well because too much insulin delivery can lead to dangerous hypoglycemia which can be fatal if extreme and untreated [4]. Type 2 diabetes (T2D) is different from T1D in that the beta cells in the pancreas can initially still produce insulin; however, the cells in the person's body have become resistant to insulin and so glucose levels can become dangerously elevated and toxic to the person when insulin secretion becomes inadequate relative to insulin resistance. In addition to T1D and T2D there is gestational diabetes, which is a condition that can happen during pregnancy whereby a woman becomes increasingly resistant to insulin and may require exogenous insulin delivery or other medications to manage their glucose [5]. New treatments in the area of diabetes care are now becoming possible because of advances in sensor technology, mobile computing, new control algorithms, data mining and also in artificial intelligence (AI) and machine learning (ML).

B. Current treatment approaches in diabetes

In T1D, the state-of-the-art therapy for managing glucose is automated insulin delivery (AID). An AID is a closed-loop system that comprises a continuous glucose monitor (CGM) that measures glucose subcutaneously about once every 1 to 5 minutes, an insulin pump that delivers insulin through a subcutaneous tube, and a control algorithm that receives the current and historical CGM data and calculates how much insulin to deliver to the person through the pump [6]. The introduction of AID into clinical care has resulted in significant improvements in glucose management such that use of AID can yield a reduction of hemoglobin A1C (HbA1c) by 0.2-0.5% compared with basal-bolus insulin therapy [7-9] in which there is no continuous feedback from glucose sensing devices and automated adjustments to insulin dosage as in closed-loop systems. Examples of open-loop therapies include basal-bolus insulin therapy and multiple daily injections (MDI). A lower HbA1c means that the person is spending less time in high glucose ranges that can cause long-term damage to health.

While AID has made a positive impact on helping people with T1D better manage their glucose levels, AID is not perfect [10]. Current commercial AIDs are so-called hybrid closed-loop systems, which means that they are not fully automated and require the person using the system to announce their carbohydrate intake to the system so that meal insulin may be dosed. People oftentimes forget to announce their meals to the system or they indicate an incorrect carbohydrate in the meal, which can cause large glucose excursions during the daytime

Manuscript submitted on 2/10/2023.

PGJ, CML, and JRC with Oregon Health & Science University, jacobsp@ohsu.edu, PH with Roche Diabetes Care, AF with Università degli Studi di Padova (Padova, Italy), JV with University of Girona and CIBERDEM, Instituto de Salud Carlos III, BK and MB with University of Virginia, AC with the Illinois Institute of Technology, JB with Universitat Politècnica de València and CIBERDEM, Instituto de Salud Carlos III, KN and KZ with the School of Electrical and Computer Engineering of the National Technical University of Athens, FD with Harvard University, TB with University Medical Center Ljubljana and University of Ljubljana funded by the Slovenian Research Agency grant # P3-0343,

when food is consumed. Exercise can be challenging because exercise (especially aerobic exercise) can cause sharp drops in glucose and dangerous hypoglycemia[11]. For these reasons, AID systems have primarily shown benefit during the overnight time, when meals and exercise do not occur [12].

Complicating the problem further is that many people still choose not to use AID systems for a variety of reasons including cost, comfort, and inconvenience of having multiple subcutaneous devices connected to their body. The majority of people with diabetes on intensive insulin therapy still use MDI therapy whereby they deliver insulin through an insulin pen. People using MDI therapy oftentimes make incorrect decisions about how much insulin to dose themselves and can therefore suffer from the complications associated with inadequate glucose management.

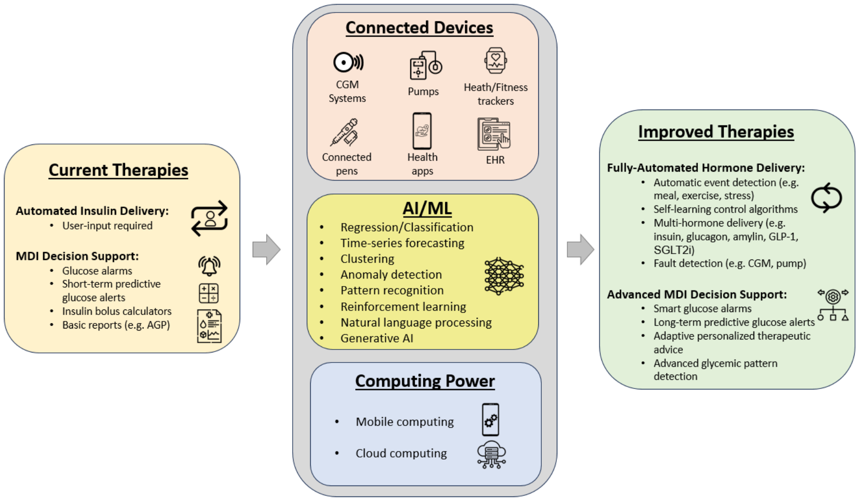

Current Therapies
Automated Insulin Delivery:
User-input required
MDI Decision Support:
Glucose alarms
Short-term predictive glucose alerts
Insulin bolus calculators
Basic reports (e.g. AGP)
Connected Devices
CGM Systems
Pumps
Heath/Fitness trackers
Connected pens
Health apps
EHR
AI/ML
Regression/Classification
Time-series forecasting
Clustering
Anomaly detection
Pattern recognition
Reinforcement learning
Natural language processing
Generative AI
Computing Power
Mobile computing
Cloud computing
Improved Therapies
Fully-Automated Hormone Delivery:
Automatic event detection (e.g. meal, exercise, stress)
Self-learning control algorithms
Multi-hormone delivery (e.g. insuin, glucagon, amylin, GLP-1, SGLT2i)
Fault detection (e.g. CGM, pump)
Advanced MDI Decision Support:
Smart glucose alarms
Long-term predictive glucose alerts
Adaptive personalized therapeutic advice
Advanced glycemic pattern detection

Figure 1: Current therapies for people living with T1D include automated insulin delivery and multiple daily injection therapy (MDI). Advances in mobile, cloud connected devices and improved computing power in combination with AI/ML are enabling new technologies in diabetes therapeutics including fully-automated hormone delivery and advanced decision support for use in MDI.

C. Improving diabetes treatment using AI and ML

AI and in particular ML, are driving discovery across the sciences in engineering, computer science, medicine and the field of diabetes treatment and therapeutics. ML has become particularly important as ubiquitous connected sensors and drug delivery devices are becoming integrated with mobile computing to generate large data sets that can be used to identify patterns that are relevant for improving health outcomes (Figure 1).

While the past 20 years have led to profound innovations in CGM and connected insulin pumps and pens, the field of ML has also had significant growth and innovation during this time. ML is a powerful tool that can be used to overcome the current challenges of current AID and MDI therapies. For example, ML can be used for identifying patterns in CGM that are useful for AID control algorithms [13]. ML can also be used to augment the automation of insulin or other hormone delivery using reinforcement learning to adapt over time to individuals' unique physiologies or to respond to disturbances such as exercise [14-18]. ML can be leveraged to develop automate recommendation systems used in decision support systems to help people living with diabetes on MDI therapy and care providers better manage insulin dosing [19-22]. There have been major successes in use of ML in applications of diabetes care. Deep learning methods have been successfully reported for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs [23, 24]. Closed-loop control algorithms for automated insulin and other hormone delivery have been augmented with ML methods for automating the detection of hypoglycemia during exercise [16, 18, 25-27], meal detection [28-37] and time series prediction models that can be incorporated into model predictive control algorithms to achieve over 70% time in glucose target range (TIR, 70-180 mg/dL) [16, 26]. Anomaly detection techniques can identify disturbances and complications in diabetes management [38, 39].

However, with the growth of new ML algorithms for use in diabetes applications, and their associated challenges in their development and implementation, there is an increasing need for best practices including guidelines on (1) how features are generated, (2) standards in metrics and how they are calculated, (3) standards on how data reconciliation, and data imputation methods are reported and performed, and (4) best practices on algorithmic approaches to enable better reproducibility and well-informed comparisons as new technologies are presented. Developing ML algorithms for diabetes applications is particularly difficult, mainly due to the scarcity and lack of structure in available datasets. Moreover, the high inter- and intra-individual variability in glucose dynamics across people living with diabetes further compounds the challenge. This variability is influenced by many factors including nutrition, lifestyle choices, medication regimens, stress, and underlying health conditions beyond diabetes [40]. In an ideal scenario, a larger volume of data would be employed to train ML algorithms on highly variable data sets. However, this is not often the case within the diabetes field. Many of the available data sets are collected either in clinical studies under highly controlled conditions that are difficult to reproduce in real-world scenarios or under free living conditions when reporting of daily activities (e.g., meals, physical activity, sleep quality, pain, etc.) and life events (e.g., those leading to high stress levels) is imperfect. Therefore, it becomes critical to establish and adhere to best practices for data processing to ensure ML models used in drug delivery and diabetes therapy are generalizable and pose minimal risk to users.

Similar to other survey manuscripts on ML in diabetes [41-43], this manuscript provides a review of prior work on ML methods in various applications in the area of diabetes with focus on T1D. Additionally, this manuscript provides a framework for how researchers can approach feature engineering, limited data set sizes, data imbalance issues, data set variability, model explainability and interpretability, personalization, and application-specific considerations for algorithm selection that can be generally applied to other applications in medicine.

We present consensus-based best practices and pitfalls to avoid when designing, training, and evaluating new models that are used in glycemic control. This guide compiles lessons learned by examining prior work in ML in diabetes. The field is dynamic, so no guide can be exhaustive. As data science evolves, some methods might need revision or could altogether be replaced by better approaches. This consensus manuscript has three primary objectives:

(i) Provide a tutorial style guide to data scientists working in the field to accelerate the development and use of ML,

(ii) Provide consensus guidelines on standards and best practices to appropriately exploit available data sets, create training/validation/test data sets, apply ML methodology, perform feature engineering, and present results,

(iii) Provide recommendations and an open-source library for standardizing calculations of common features and model evaluation metrics used in diabetes ML algorithms and an online list of data sources.

D. Methods for selecting manuscripts, reaching consensus

The PubMed, Science Direct and Google Scholar databases were considered to obtain the most relevant research works of the last thirty years using the search terms 'diabetes', 'machine learning', 'glucose prediction', 'continuous glucose monitoring', 'automated insulin delivery', 'closed-loop', 'decision support', and 'artificial intelligence'. Our search identified 189 manuscripts that were used to support the best practices and pitfalls presented. Manuscripts were reviewed based on one of the following modeling aims related to glucose control: (i) short-term continuous glucose monitoring (CGM) prediction within < 60 min, (ii) long-term CGM prediction over 60 min, (iii) CGM prediction during exercise, (iv) nocturnal CGM prediction, (v) detection and estimation of events including hypoglycemia and meals (vi) personalization and adaptation (vii) other applications of ML in diabetes management including closed-loop control and decision support. While many manuscripts have reported on T1D because of the availability of CGM in this patient cohort, most of the methods presented here may be applied to T2D and gestational diabetes as well. Certain manuscripts are included as examples of good practices or pitfalls in the field. We used a modified Delphi method [44] for reaching consensus on the guidelines. In-person meetings were arranged with each author to discuss the approach and gather initial feedback. A set of questions were distributed to authors regarding a preliminary set of guidelines. Authors provided written feedback on the questions and authors met in person and virtually at the Advanced Technologies and Treatments in Diabetes in Barcelona in April 2022. A first draft was released, and three subsequent meetings were organized to reach consensus among the authors before a final draft was completed.

E. *Related work*

There have been several survey, review, and meta-analyses manuscripts on ML in the diabetes domain [43, 45-47], presenting, mainly, summaries of various algorithms and methodologies while displaying the advantages of using ML algorithms. There are also a few manuscripts describing ML in clinical research [45], its use in education [48], and its use in clinical guidelines and recommendations [49]. A few manuscripts present ML-on-the-edge and Internet-of-things that describe methods to combine ML with smart devices [50]. Other manuscripts focus on subtopics like metrics [51], or glucose prediction [52]. This manuscript weighs the pros and cons of different approaches after forming consensus from all authors for the practitioner to make informed decisions.

II. Data

A. *Current real-world and clinical study data sets available*

The commercial availability of CGM sensors, insulin pumps, smart insulin pens, and other wearable fitness sensors that push data to cloud servers has led to rapid growth in the amount of time-matched glucose, insulin, nutrient, exercise and other sensor data. There are three types of data that are typically used in diabetes ML applications: (1) real-world data collected under free-living conditions, (2) data collected under controlled conditions within the frame of clinical trials that may take place in a hospital, in the home, or a combination of both, and (3) simulated data generated by means of executing simulation scenarios to a virtual environment (e.g. data farming). For data collected in clinical studies, participants usually adhere to strict protocols for food intake and exercise, and often the type of food and exercise is controlled as well. A controlled environment is preferred for measuring the efficacy of drugs, algorithms or interventions whereas free-living data sets are suited for the development of multiple-hormone closed-loop systems and decisions support algorithms. Simulated or synthetic data can be easily generated by physiological (compartmental) models expressed as ordinary differential equations (ODE) using simulation environments as described further in section II.C. There are methods of adding noise and variability to simulated data to make them more *real* but these *in silico* subjects still behave differently than real-world people since all factors affecting metabolism are not being modeled.

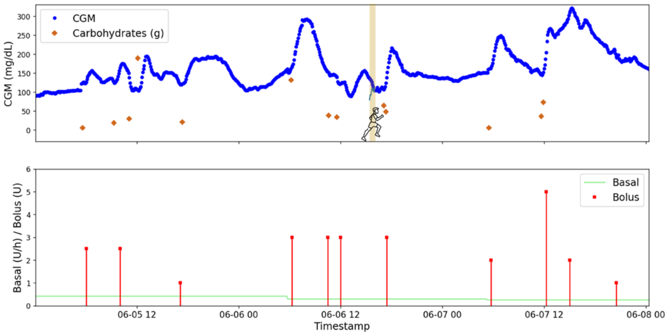


Figure 2: Top panel shows example traces of CGM (blue) and carbohydrate intake (red) in a person with T1D. Bottom panel shows examples of basal insulin (green) and bolus insulin (red lines) taken by the person.

An example of the type of CGM, carbohydrate, insulin, and exercise data collected from people living with T1D is shown in Figure 2. One clinical data set that is widely used in ML in diabetes applications is the Ohio T1DM Data set which was originally published in 2018 and then updated in 2020 [53]. The most recent Ohio data set includes time matched CGM and insulin data from 12 people with T1D over 8 weeks under free-living conditions. Physical activity and self-reported stress are also included in this data set. Various algorithms have been trained using this data set including Zhu et al. [17], however this data set is small compared with some other recently available data sets.

A much larger real-world data set is being collected by a company called Tidepool [54]. People with T1D donate their CGM data, insulin data, and other data types including physical activity data to the Tidepool Big Data Donation Data set. Tidepool then licenses the data set to companies and academic institutions interested in extracting knowledge and mining the data to develop new ML algorithms. The Tidepool data set has been used to train ML algorithms for predicting overnight hypoglycemia at the time when a person goes to sleep [55] and also to predict short-term glucose and hypoglycemia up to 60 minutes in the future [26]. It has also been used to develop a DNN to detect meals, exercise and their concurrent occurrences as well [25, 56].

A data set that was recently collected and released to the public in 2022 is the T1-Dexi data set [57]. The T1-Dexi data set is one of the largest data sets comprising time-matched CGM, insulin, genetics data, food intake, and physical activity data (heart rate and accelerometry). It was obtained through the execution of a 4-week study involving 497 people with T1D who performed aerobic (n=162), resistance (n=170), or interval (n=165) exercise several days per week while recording nutrition information using a custom smart phone app [58]. It is an excellent resource for designing ML algorithms, especially as related to exercise and food intake.

In addition to these data sets, the Jaeb Center for Health Research maintains a web site listing data sets available for use [59].

B. *Simulators available to generate data in diabetes research*

A simulator in diabetes comprises a set of equations that describe the dynamics of glucose metabolism as a set of compartments in the body representing subcutaneous tissue, the gut, plasma, and other non-observable compartments. Parameters of the metabolic model can be statistically sampled from a distribution of parameter values. The distribution of parameter values is typically identified using physiology tracer-study experiments [60] to generate a virtual patient population of simulated people with diabetes with different insulin absorption kinetics and dynamics, carbohydrate absorption kinetics and dynamics, other hormones (e.g. glucagon and pramlintide), and different responses to exercise (e.g. aerobic, resistance, interval). The T1D simulator has been an important tool that has helped in the design and commercialization of the first commercial AID systems [7, 9] and

This article has been accepted for publication in IEEE Reviews in Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/RBME.2023.3331297

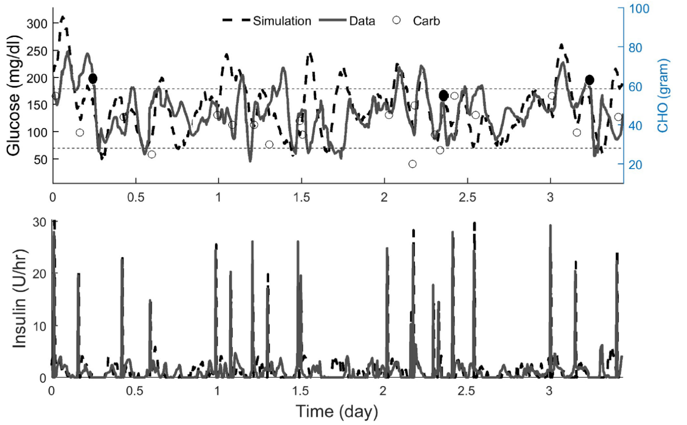


Figure 3: Example of CGM and carbohydrate traces (top panel) and insulin delivery (bottom panel) from a metabolic simulator (dashed line) vs. real-world data (solid line) [68].

also in the design of multi-hormone delivery algorithms [16, 34, 61-66]. Many ML algorithms are initially designed and tested on simulated data prior to being evaluated in human studies[20, 22]. In this way, simulators based on compartmental modeling play an important role in the preliminary design process of ML-based closed-loop control algorithms and decision support algorithms in diabetes an example of an open source simulator is shown in Figure 3.

The UVA-Padova T1D simulator [67] is the only FDA-accepted simulator that has been published as a substitute for animal trials. The simulator has been used in a variety of applications and has been instrumental in the preliminary design and evaluation of several control algorithms including the now commercially available Control IQ [9] and others described below.

The open-source OHSU T1D simulator was published in 2019 and is available for download and extensions from a Git repository [68]. This simulator has been used to evaluate several control algorithms prior to a clinical study [16, 61, 64] and also to pre-train ML algorithms prior to evaluation in human studies [20, 26, 55, 69-71].

Other simulators have also been described and can be obtained by contacting the authors. These include a statistical virtual patient population published by Haidar et al. [72], a multivariable simulator developed by Cinar and colleagues [73] that permits scheduling of exercise bouts (with intensity and duration that can be randomly modified) and also provides as outputs the values of physiological variables that are reported or predicted (energy expenditure) by wearable devices, a simulator developed by Vehi and colleagues [74], and a simulator developed by Wilinska and Hovorka [75].

Data scientists should use caution when designing new ML algorithms with simulators as the algorithms may work well on simulated data, but perform poorly in a real-world situation because simulation does not capture real-world events that influence glucose levels that are not included in the models such as medications, menstrual cycle, and stress. Data scientists should always verify their algorithm performance on real-world data and include these results in their publication (see Common Pitfall 9 and Best Practice 12).

C. Standardization of reporting on data sets used in ML algorithm development and evaluation

Currently, it is challenging to compare algorithms described in different publications because the algorithms are typically trained and evaluated on different data sets. Comparing algorithms across benchmark data sets is critical for improving reproducibility of ML algorithms [76, 77]. In the area of short-term glucose forecasting, it is especially important to clearly indicate the variability of the glucose data [78]. The use of AID or automated multi-hormone delivery, may results in less glucose variability compared to those obtained by applying sensor-augmented pump therapy or multiple daily injection therapy [79]. Consequently, an algorithm trained and evaluated on a data set with AID control may have superior performance and less error compared with an algorithm trained and evaluated on MDI data. Mosquera-Lopez et al. [26] discussed utilizing regression metrics to quantify how glucose prediction changes with variability in the glucose data set using the glucose variability impact index (GVII) and the glucose prediction consistency index (GPCI). GVII is the slope of a regression line between the RMSE error and the variance of the glucose data. If the GVII (slope) is flat, then it means that the error is not significantly impacted by the variance of the glucose data. The GPCI is the standard deviation about the regression line, indicating how consistent the RMSE is across the data set.

Common Pitfall 1: Beware of training and evaluation data that lack heterogeneity or with low glucose variance. It is easy to achieve good accuracy on a data set with low variability.

Best practice 1: Evaluate algorithms on open-source baseline data sets if available (Table 1 or [59]). Alternatively, evaluate on a data set that is included with the publication.

Best practice 2: Include performance across data sets with differing variability using metrics such as the GVII and GPCI [26] for any data set used to train or evaluate an ML algorithm.

Best practice 3: When training an algorithm that is designed to work regardless of the therapy (e.g. MDI, closed-loop, sensor augmented pump), include data balanced across all of those therapies in the training and test data sets and ensure that the data sets are well representative.

Best practice 4: Include a data sheet [80] *for the data set used in training, validation and testing.*

Data scientists should publish detailed information on the data sets used in algorithm development as a formal data sheet [80]. The data sheet (Supplemental Table 1) should include information about the data set including information regarding volume (e.g. number of participants), demographics, and treatment plan, along with missing data and interpolation done if any, variability of the data, and other relevant information to be considered when developing comparator algorithms.

Common Pitfall 2: Mixing individual data with both training and test data sets can lead to reporting of unrealistically high accuracy compared with when evaluated on individuals who were not included in the training set. In ML algorithm development, there is a training data set, a validation data set used for hyperparameter tuning, and a test data set. The test data set is not used at all in training the algorithm or in hyperparameter tuning. When training and cross-validation are done, a portion of data is held out as the test data set is used for evaluation. One of the natural consequences of limited data sizes in the field of diabetes is the temptation to increase the data available for training by including in the training data set data from a patient reserved for testing. This is problematic for several reasons. First, it implies that the accuracy reported on the algorithm would require access to a certain amount of the person's own data for use in training. A population-level ML model is not re-trained when new data is available from an individual. An adaptive or personalized model, however, can be potentially re-trained when new data is observed to improve the accuracy for a given individual.

An example of an algorithm that was designed to be personalized for specific individuals is presented in Zhu et al. [17] whereby they designed a CNN for forecasting glucose 30 minutes in the future. They included data from all six of the participants in the Ohio T1DM Data set in the training, and then forecasted on future data from one of the participants. In addition, data was augmented by extending by 50%, each of the six participants' glucose data sets with a mixture of the other five participants' glucose data. In this way, each participant's data set was doubled in size by including a mixture of data from the five other participants. This type of an algorithm would be appropriate for a personalized model, but would not be appropriate as a general population model. Population models can be personalized using transfer learning [81, 82] or meta learning [83-86] approaches.

TABLE I
COMPARISON OF DATA SET SIZES IN DIABETES (TOP) COMPARED WITH OTHER FIELDS (BOTTOM).

Name	Description	Field	# Observations
T1-Dexi [47]	Time-matched glucose, insulin, nutrition, and exercise data collected from n=497 people with T1D under free-living conditions over 28 days.	Diabetes	3,737,664
Tidepool Big Data Donation Set [44]	Time-matched continuous glucose, insulin, physical activity data from people with T1D	Diabetes	3,263,904
Ohio T1DM dataset [43]	Small open-source data set on time-matched glucose, insulin data	Diabetes	193,536
Other data sets outside the field of diabetes			
MovieLens	Moving ratings	Rating systems	20,000,263
Google SmartRep	Interactions with smart-agent	Intelligent dialog	238,000,000
Objects365	Object detection imaging dataset	Object detection	10,000,000+
Translate	Google translation data set	NLP	> 1 trillion

The heterogeneity of the test set should also be balanced with that of the training set. For example, if there are no children or adolescents in the training set, but they are present in the test set, then the algorithm may not perform as well on the adolescents during evaluation (Best Practice 5). Furthermore, the training and test sets should include approximately the same number of observations for the different target classes (classification problems) or comparable overall dynamics (regression problems). Checks should be made to ensure that training, validation, and test sets are balanced as much as possible. Care should be taken not to under-sample the majority class in case the dynamics are not preserved. Under- or over-sampling must be done in a smart way so as not to introduce bias using methods such as cluster-based centroid sampling described by Yen et al. [87].

Best practice 5: Ensure that the training, validation, and test sets are balanced and cover the same population groups.

D. Handling missing, calibration, interpolated, and synthetic data

CGM and insulin data are frequently incomplete because they are collected from devices that are wirelessly connected and they sometimes fail. Sensor misplacement or infusion site failures can cause gaps in data as well. Sensor faults can occur in closed-loop systems [88] which can be caused by pressure-induced sensor attenuations as reported by Bequette and colleagues. Insulin pumps can also fail during usage as caused by infusion set actuation problems [89]. Machine learning approaches have been applied to detecting these anomalies and alerting patients to these failures [90]. Data scientists must decide how to handle missing CGM and insulin data in the training, validation, and test sets. Some examples for handling missing data include (1) linear and nonlinear interpolation, (2) extrapolation if current data is not available using forecasting models, (3) zero-order hold, and (4) exclude missing data from the training and test sets. CGM tends to change rapidly enough that linear interpolation is a good choice if the gaps are less than about 20 minutes. After that, interpolation may not be appropriate. Data scientists should clearly describe their methods for handling missing data. Data scientists should also be careful not to report prediction accuracy on interpolated values to prevent data leakage from future values. Furthermore, data sets also may include calibration data from blood glucose meters. Data scientists should be clear to specify how calibration data or data from blood glucose meters is handled differently than CGM data.

Insulin data collected from pumps or smart pens (e.g. the Tidepool data set [54]) may not always clearly indicate if the insulin was taken for a meal or as a correction for a high glucose reading. Meal insulin is typically calculated by dividing the grams of carbohydrate consumed by a carbohydrate ratio, but this information may not be available in a pump record. If there is knowledge of the person's correction factor (CF), their target glucose ($\text{CGM}_{\text{target}}$), and the glucose at the time that insulin was dosed ($\text{CGM}_{\text{current}}$), we may presume that a portion of the insulin dosed was to get their glucose to return to their target glucose using their correction factor minus any insulin on board (IOB) that is not being used [91]. The inferred meal insulin is then just the difference of the actual insulin dosed minus the inferred correction dose as shown in Equations 1 and 2.

$$inferred\ correction\ dose = \frac{CGM_{current} - CGM_{target}}{CF} - IOB_{unused} \quad (1)$$

$$inferred\ meal\ dose = actual\ dose - \ inferred\ correction\ dose \quad (2)$$

Note that inference introduces inaccuracies and should be clearly explained as a limitation by data scientists if used.

Best practice 6: Report methods to handle calibration data and interpolation of missing data in training, validation, and test sets.

Common pitfall 3: Reporting accuracy on interpolated data in the test set can lead to invalid estimates of accuracy on actual data. Performance should be reported without using interpolated values in the test data set.

Common pitfall 4: Use caution when applying imputation or smoothing to the test data set as it can cause future data points to impact current data points if done incorrectly. CGM cannot change faster than what is physiologically plausible. In a manuscript by Clarke and Kovatchev [92], they showed that the CGM does not typically change faster than ± 4 mg/dL/minute. Therefore, smoothing outliers may help to remove noise that is not physiologically possible.

Common pitfall 5: Evaluating algorithms on synthetic data may yield invalid accuracy results. Data scientists may improve algorithms by including synthetic data using methods such as SMOTE [93] or generative adversarial neural networks to fabricate synthetic glucose data [94] to improve model accuracy or handle class imbalance during training. It is important to apply the synthetic method only on the training data and ensure that synthetic data are not in the test set.

E. Data size considerations in diabetes ML

The size of publicly available data sets in diabetes are typically a lot smaller than data sets used in other fields of ML (Table 1). For this reason, the machine learning methods employed in diabetes and medicine that rely on smaller data sets need to be different than the ones used on larger text and imaging data sets. It is important for data scientists to consider the size of the data set prior to selecting various candidate ML algorithms. As a rough rule of thumb, a model should train on at least an order of magnitude more examples than trainable parameters [95]. Simple models trained on large data sets generalize better and therefore perform better than more complex models trained on small data sets. Particular care has to be taken when exploring the use of deep neural networks (DNNs) on small data sets, since there are a large number of parameters that must be learned and this can lead to overfitting [96]. DNNs often need more data than traditional ML methods to train, and do not generalize well when the data set is small relative to the number of parameters. In computational learning theory there exists the concept of the Vapnik–Chervonenkis (VC) dimension [97], which gives a lower bound on the minimal number of training examples required to learn a model correctly. However, the VC-dimension is a theoretical concept and not often used in practice. More often, a data scientist can explore how accuracy of an algorithm changes when trained on increasing fractions of the complete development data set. The expectation is that the performance will improve with increasing amounts of training data and then plateau after a certain upper bound amount of data is reached [95].

III. Features and outcomes useful in diabetes ML

When designing an ML algorithm for predicting future glucose, the common data types that may be useful as input features for the algorithm include (1) recent CGM measurements and statistics on CGM, (2) recent nutrient intake, especially carbohydrates, (3) recent insulin doses, (4) recent other hormone doses (if applicable), (5) recent physical activity, stress, or other physiologic measures as estimated from wearable sensor data, and (6) demographics information. When wearable device data are used to determine the metabolic state of a person, additional features may be helpful including classification of physical activities and stress to improve accuracy of estimated glucose concentrations [25, 98-100]. While it is important for data scientists to freely explore and experiment with many ways of representing features as inputs to glucose forecasting algorithms, this section provides standard ways of representing common features. In addition, in online supplementary materials, we provide functions in Python to calculate each of these features to help improve standardization and repeatability.

A. Statistical representations of CGM as possible features in ML algorithms and for use as performance metrics

For most commercial glucose sensors, data is sampled every one or five minutes. The Dexcom G6/G7 and Medtronic Guardian Sensor 3 provide data every 5 minutes while the Abbott Freestyle Libre 3, Waveform (Agamatrix) and the GlucoMen (A. Menarini Diagnostics) sensors provide data every minute. Most CGM forecasting models use a history of CGM data as input features. Autoregressive (AR) and autoregressive with exogenous inputs (ARX) models are examples whereby the use of this history of CGM is explicit [101, 102].

The collinearity of CGM as measured by autocorrelation tend to disappear after about 1 hour [103]. For regression-based models, collinearity can be a problem, whereas for time-series models like AR and ARX, the collinearity is positive, and histories are selected based on the autocorrelation being above a certain threshold. Choosing the history length is application specific and data scientists should explore different history lengths when designing their algorithm. Some groups have used grid search to determine what an optimal history of CGM is required to maximize performance. For example, Mosquera-Lopez and Jacobs compared CGM history of 1, 2, and 3 hours and found that 3 hours was optimal for short-term prediction of glucose using a long-short-term memory neural network [26]. However, other algorithms have reported on shorter histories, though not indicating if other history lengths were explored (e.g. Perez-Gandia et al. [104] used a 20-minute history of CGM as their input to a neural network).

The downside of choosing longer histories is that CGM sometimes drops out due to connectivity problems, and so there could be gaps in the data. Interpolating large gaps in data may negatively affect the glucose performance of a forecasting algorithm. While the history length is application and algorithm specific, it could be preferable to choose shorter histories of glucose for calculating CGM-related features to minimize the impact of CGM drop-out when used in practice.

<u>*Best practice 7: For short-term glucose forecasting tasks (e.g. 30-60 minutes), it could be preferable to choose a short history of glucose for calculating CGM-related features as inputs to ML algorithms to minimize the impact of device-related missing CGM, which happens in real-world practice.*</u> To mitigate the impact of missing data, data imputation can be used to fill in gaps in data. In addition, during training, missing data should be introduced into the data set to ensure that the algorithm can appropriately handle it.

The history length should be determined to optimize the performance of the prediction task. For example, if the data scientist is designing an algorithm for predicting overnight hypoglycemia prior to bedtime, over the course of the next 4-8 hours, summary measures of CGM that have occurred over the past one to two days could be important as well as recent CGM. When predicting glucose over an 8-hour window overnight to estimate likelihood of hypoglycemia during this window, Mosquera-Lopez et al. used statistical measures of glucose across the prior 1 h, 3 h, 6 h, 9 h, 12 h and also summary measures across the past week [55].

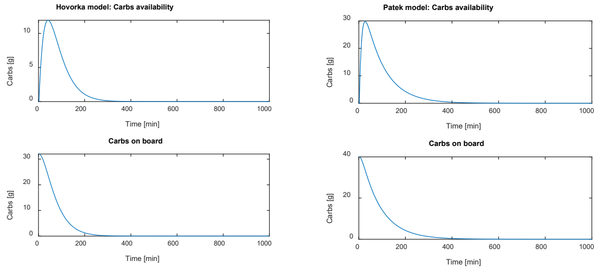


Figure 4: Carbohydrate availability and carbohydrate on board by Hovorka et al. and Patek et al. The top plot shows the plasma glucose following a 40 g carbohydrate meal. The Patek et al. model peaks earlier (27 min) compared with the Hovorka et al. model (40 minutes). Also note that the Hovorka model only presumes that 80% of the meal (32 g) is available to the body.

Glucose outcome metrics that are traditionally used as indicators of glucose management performance may also be useful as inputs to ML algorithms. Common outcome metrics that may optionally be used as inputs to glucose forecasting algorithms include statistical measures of the glucose across a historical time window such as mean, variance, skewness, and kurtosis. In addition, there are clinically relevant metrics that are also used as inputs to ML algorithms. These metrics are included in Table 2. The clinically relevant ranges have been determined by a consensus of clinical experts [105] and would be useful whenever considering clinically relevant input features to ML algorithms. Many of the glucose outcome measures are correlated with each other and using all of these features as inputs may induce collinearity in data and not be helpful in ML. For example, the % time in range is correlated with the % time below plus % time above range. To avoid the problem of correlated features, principal component analysis (PCA) can be used to do dimensionality reduction of correlated features into a smaller set of orthogonal features [100, 106, 107]. The disadvantage of using PCA-based features is that the features are less interpretable. Another option for handling correlation amongst features is to perform feature selection by eliminating less relevant features that are correlated with more relevant features.

<u>*Best practice 8: When including clinically relevant summary measures to quantify performance of the forecasting algorithm in clinically relevant ranges, it is important to use the ranges agreed upon by an international consensus group and summarized in [105]. These measures may also be useful as ML features, but the ranges and features should be selected based on the task of the algorithm and their impact on accuracy and explainability.*</u>

B. Representation of carbohydrate intake

Carbohydrate intake frequently causes a significant increase in glucose levels in people with diabetes. For this reason, carbohydrates on board or carbohydrates in plasma can be used as an input to glucose forecasting algorithms. One way to represent carbohydrates as a feature in an ML algorithm is to use a two-compartment differential equation *carbohydrate-absorption in plasma* model. In his 2004 manuscript, Hovorka et al. [108] described a second order differential equation representation of carbohydrate absorption in plasma. The equations for this model are given below whereby the carbs in grams is represented by the variable *c*. $Q_1(t)$ is the amount of glucose in the gut and $Q_2(t)$ represents available carbs in plasma. The constant t_{max} represents the time constant for meal absorption and its default value is 40 minutes but this might change depending on the type of meal.

$$\dot{Q}_1(t) = -\left(\frac{Q_1(t)}{t_{max}}\right) + 0.8c,$$

$$\dot{Q}_2(t) = \left(\frac{Q_1(t)}{t_{max}}\right) - \left(\frac{Q_2(t)}{t_{max}}\right). \tag{3}$$

Notice in that the Hovorka et al. [108] representation of carbohydrate availability assumes that only 80% of the carbohydrate consumed is utilized (e.g. 0.8c). Another model for estimating carbohydrate availability is described by Patek et al. [109]. The Patek model is also a two-compartment model. However, there are two time constants, a short-acting meal absorption t_I of 11.2 *min*, and a longer-acting absorption constant t_{abs} of 83.8 *min*.

$$\dot{Q}_1(t) = -\left(\frac{Q_1(t)}{t_1}\right) + c,$$

$$\dot{Q}_2(t) = \left(\frac{Q_1(t)}{t_1}\right) - \left(\frac{Q_2(t)}{t_{abs}}\right). \tag{4}$$

Others have described meal models including Dalla Man et al. [110]. Regardless of the method for representing carbohydrate distribution and disposal, we define carbohydrate in plasma and carbohydrates on board using Equation 5 and 6.

$$Estimated\ carbs\ in\ plasma\ (t) = Q_2(t) \tag{5}$$

$$Carbs\ on\ board\ (t) = c(z) - \int_{z=carb\_start}^{t} Q_2(z), \tag{6}$$

In Equation 6, *carb_start* is the time of carbohydrate ingestion and *c(z)* is the grams of carbohydrate ingested at time *z*.

Figure 4 shows carbohydrate availability and carbohydrates on board and the comparison of these two methods for estimating carbohydrate availability after consuming a 40 g carbohydrate meal. The Patek et al. model peaks somewhat earlier (27 minutes) compared with the Hovorka model (40 minutes). This leads to a higher peak carbs on board, which is even larger because the Patek et al. model presumes that the entire carbohydrate consumed is utilized, rather than 80% of it. The Python functions for calculating carbohydrate in plasma for these two methods is included in Supplemental Materials.

Carbohydrate in plasma is not monotonic, as it will have the identical value when it is rising as when it is falling. This is important to consider if used as a feature in ML algorithms, because a monotonic representation of carbohydrates provides more information about the future than a non-monotonic representation. Therefore, a better feature for a forecasting algorithm can be to use carbohydrates on board rather than carbohydrate availability.

Importantly, current meal models do not account for other nutrients (fat, protein and fiber) which also affect glucose response, primarily in the area of delayed gastric emptying.

C. Representing estimated plasma insulin vs. insulin on board as features

Insulin that is injected subcutaneously does not appear immediately in the plasma. There is a delay that is caused by the metabolism of insulin from a hexamer into a monomer and then movement from the subcutaneous space to plasma. The peak appearance of fast-acting insulin in plasma after injection subcutaneously is typically 40-60 minutes. Estimated plasma insulin is the amount of insulin in plasma, and as with carbohydrate availability, it can be represented by a set of differential equations or alternatively as a linear function over time. Most glucose metabolism simulators [67, 68] use multi-compartment differential equation models to represent the kinetics of insulin into plasma. For example, estimated plasma insulin can be represented using a 3-compartment model described by Hovorka et al. [108]. This model of estimated plasma insulin (I) is given by Equation 7 whereby u_I is the insulin injected subcutaneously and I is the insulin in plasma or estimated plasma insulin and t_{maxI} is the time constant for insulin absorption into plasma.

$$\dot{S_1}(t) = u_I - \left(\frac{S_1(t)}{t_{max}}\right),$$

$$\dot{S_2}(t) = \left(\frac{S_1(t)}{t_{maxI}}\right) - \left(\frac{S_2(t)}{t_{maxI}}\right),$$

$$\dot{I}(t) = \left(\frac{S_2(t)}{t_{maxI}}\right) - S_2(t)I(t). \tag{7}$$

TABLE 2

CGM FEATURES USED IN IN DIABETES RESEARCH.

Type	Name	Description
Statistical Features	Mean	Average CGM over time window
	Variance	Variance of CGM over time window
	Covariance	Covariance of CGM over time window
	Coef of variation	Coefficient of variation over time window
	Skewness	Skewness of CGM over a time window
	Kurtosis	Kurtosis of CGM over a time window
	Maximum	Maximum CGM over a time window
	Minimum	Minimum CGM over a time window
	Slope	Rate of change of glucose, typically estimated through a regression across the most recent 10-15 minutes
Clinically relevant features	HbA1c	Rate of hemoglobin glycation
	Mean glucose	Mean glucose across a time frame
	% time in range	% Time glucose is between 70-180 mg/dL
	% tight range	% Time glucose is between 70-140 mg/dL
	% low	% Time when glucose is < 70 mg/dL
	% very low	% Time when glucose is < 54 mg/dL
	% high	% Time when glucose is > 180 mg/dL
	% very high	% Time when glucose is > 250 mg/dL
	# hypo events	Number of times over a window when glucose dropped < 70 mg/dL
Demo-graphic features	Sex	Male, female, non-binary
	BMI	Body mass index
	Duration diabetes	Years of diabetes since diagnosis
	Race/ethnicity	

Insulin on board is defined according to Equation 9

$$Insulin\ availability\ (t) = I(t) \tag{8}$$

$$Insulin\ on\ board\ (t) = u_I(z) - \int_{z=ins\_dose}^{t} I(z), \tag{9}$$

where *z=ins_dose* is the time when $u_I(z)$ insulin was dosed.

Other groups have used a triangular compartmental model trained on data from Swan et al. to describe action and then use convolution with past insulin injected [111]. It is also possible to represent estimated plasma insulin and insulin on board as a simple linear decay with a 3- or 4-hour linear decay over time from the time that it is delivered. While the linear decay representation of estimated plasma insulin is simpler, it is clear that it ignores the delayed peak in estimated plasma insulin that is representative of insulin kinetics. A real-time personalized plasma insulin concentration estimation based on CGM and insulin data, and demographic information has also been developed and used in AID systems [112, 113]. Figure 5 shows a comparison of estimated plasma insulin availability and insulin on board as calculated by Equations 8 and 9, respectively.

<u>*Best practice 9: When considering food and drug intake as features in an ML model, it is important to consider the kinetics and dynamics of these compounds within the body as they metabolize. Selecting the way to represent these compounds should be considered based on application to maximize algorithm performance and to minimize the risk to the person using the algorithm.*</u> Estimated plasma carbohydrate and insulin as calculated using Equation 5 and Equation 8, respectively could be suitable features for short-term glucose forecasting. Carbohydrates on board (Equation 6) and insulin on board (Equation 9) may be more useful for longer-term predictions, because the carbs

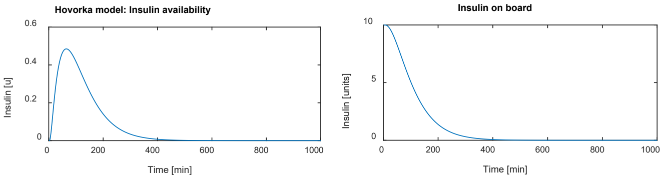


Figure 5: Insulin on board and plasma insulin availability as calculated by Hovorka et al. 3-compartment insulin kinetics model

and insulin consumed will tend to act for many hours in the future and the monotonic nature of these parameters is helpful. As new insulin formulations become available, ODE model parameters for insulin on board calculations will need to be updated. Note that deep learning approaches will not need these transformations as the insulin and carbohydrate dynamics can be learned by the network.

D. Representation of exercise

1) Exercise is challenging in diabetes

Exercise can impact glucose changes in people with diabetes in different ways depending on the type of exercise, the duration, the intensity of the exercise, and the timing of the exercise relative to meal intake and insulin dosing [57]. Exercise causes increases in insulin-mediated and non-insulin mediated glucose uptake [114]. Sharp drops in glucose can occur when people with T1D perform aerobic exercise 1-2 hours following a meal when insulin on board is high. While people with T1D are advised to reduce basal and meal bolus insulin 1-2 hours prior to an exercise event to avoid hypoglycemia during and following exercise, they oftentimes do not do this and therefore suffer from exercise-induced hypoglycemia [11, 115]. Current commercial AID systems do not automatically respond to or anticipate exercise, and this can result in hypoglycemia during and following exercise.

2) Enabling automated response to exercise using ML

Automating the detection of exercise and the automated response of an AID to exercise represents an opportunity for the use of ML in AID therapy. Some AID algorithms have features that enable the user to announce exercise in advance so that they can exercise more safely by raising the glucose target, but this is not automated. Multivariable AID control algorithms have been reported that automatically detect exercise from wearable fitness sensors and adjust insulin dosing in response to different types of exercise [18, 113, 116]. ML algorithms that attempt to predict glucose changes during exercise therefore must consider many factors during the prediction. Riddell and colleagues [57] identified the most relevant features related to glucose drops during exercise in a large free-living data set in T1D (the T1-Dexi Initiative) given in Table 3. These features are useful as inputs to ML algorithms for forecasting change in glucose during exercise.

3) Sensors used to capture and quantify physical exertion

Enabling an AID system to automatically respond to exercise requires collection of physiologic metrics representative of exercise. Exercise physiology data is typically in the form of either heart rate or accelerometry data. There has been an explosion of wearable fitness sensors that can be worn on the wrist or the chest to track heart rate and accelerometry during exercise. It is important to consider where on the body heart rate and accelerometry data are being acquired. For example, accelerometry data acquired from the wrist will look very different than when acquired from a sensor worn on the chest. Heart rate data acquired on the chest will typically be more accurate than when acquired on the wrist [117], although it is more convenient for a person to wear a wrist-based activity monitor than a chest-based monitor.

Common pitfall 6: An exercise detection algorithm trained on physical activity data collected from a chest-mounted sensor may not work as well if tested on activity data collected from the wrist, and vice-versa. When reporting results on algorithms utilizing physical activity data, it is important to consider and report where the sensor was positioned on the body. Accuracy of commercial wrist-worn devices has been assessed and there can be variability across different manufacturers and models [117]. Skin color can also affect accuracy [118].

AID algorithms that use ML to respond automatically to exercise will need to be robust so that they can handle the different types of exercise being done. There are three broad categories of physical activity, aerobic (e.g. jogging), resistance (e.g. weight lifting), and interval exercise (e.g. Crossfit, soccer, etc.) and these different types of physical activity can impact glucose in different ways. For example, Riddell and colleagues showed in a large free-living study that aerobic exercise can cause an average drop in glucose of -18±39 mg/dL while resistance and interval exercise cause drops of -14±32 and -9±36 mg/dL, respectively [57]. Exercise can also cause glucose to increase, especially when interval exercise is done in the fasted state or during competition [115]. Algorithms have been published to classify the type of exercise [119, 120]. Various groups have published methods for detecting the onset of physical activity [121, 122] and categorizing the types of physical activity [98, 100, 123]. These algorithms typically use blood volume pulse or heart rate, averaged over a window of time, and a tri-axial accelerometer magnitude, also averaged over a period of time, as features within the algorithm. Depending on the forecasting task being done, the time window across which heart rate and accelerometry data should be averaged should be carefully considered. For example, during interval exercise and resistance exercise, the heart rate and accelerometry signals tend to change very rapidly from minute-to-minute. Therefore, using a shorter time window for averaging these signals would be important for an algorithm classifying exercise type.

TABLE 3
FEATURES STATISTICALLY RELATED TO GLUCOSE CHANGE DURING EXERCISE.

Feature
Exercise type (aerobic, interval, resistance)
Sex
HbA1c
Baseline glucose at start of exercise
Rate of change of glucose in 15-minutes before start of exercise
Percent time < 70 mg/dL in 24-hours prior to exercise
Heart rate at start of exercise
Time-of-day of exercise
Insulin on board at start of exercise

Best practice 10: Heart rate and accelerometry offer a reliable set of physiological measurements based on wearable devices to quantify exercise and impacts on glucose changes. Selection of time windows for averaging these signals and generating features should be specific to the forecasting task.

Heart rate and accelerometry offer their own advantages and disadvantages for quantifying energy expenditure during exercise. If heart rate is estimated from blood volume pulse signals from a wristband, the arm movement causes large artifacts on blood volume pulse. In this case artifacts must be eliminated before computing heart rate as described [98, 100, 129]. Accelerometry has the advantage of capturing the more rapid onset and offset of exercise, since it takes time for the heart rate signal to increase or decrease during transitions from rest to exercise and vice versa. However, heart rate may increase as a result of stress [130] instead of physical activity. In this way, a combination of heart rate and accelerometry may jointly provide a complete set of features for predicting the impact of exercise on changing glucose levels.

Common pitfall 7: Utilizing only accelerometry signals to estimate physical activity can be inaccurate, especially with wrist-worn fitness watches.

4) Quantifying exercise features using ODE-based models

Features used in an ML algorithm that quantify exercise can be derived from ODE-based compartment models that describe the impact of exercise on glucose dynamics. These ODE models could be useful for deriving features used in glucose forecasting algorithms during and following exercise. A number of ODE exercise models have been published and most use heart rate and/or accelerometry as an input to the model. For example, the OHSU T1D compartment [68] model metabolic simulator includes a model of exercise described by Hernández-Odoñez et al. [124] that uses metabolic expenditure as a function of active muscle mass and metabolic equivalent of task (METs) to impact glucose disposal. As METs increases, insulin sensitivity also increases and thereby more glucose is disposed. Hobbs et al. [125] proposed a more comprehensive model of the effects of physical activities on glucose concentration that was instrumental in the multivariable glucose-insulin-physiological variables simulator that provides estimates of various physiological variables as outputs. Dalla Man and colleagues have also incorporated exercise into a metabolic compartment model [126] following on work by Breton [127]. Ozaslan et al. [128] introduce the idea of physical activity on board (similar to insulin and carbohydrates on board), which presumes that past exercise has an additive effect on future changes in glucose.

Best practice 11: Compartment models offer physically interpretable models of metabolism and can be used to generate features for ML algorithms for representing exercise by heart rate and accelerometry.

IV. ML METHODS APPLIED TO MODELING IN DIABETES

While much of ML research has been based on algorithms trained on very large data sets oftentimes involving 2-dimensional data (e.g. images), ML efforts in the application area of forecasting and modeling in diabetes typically involve much smaller data sets primarily using multivariable time-series data. In addition to the CGM data that are sampled every 1-5 minutes, insulin data are available from pumps that deliver insulin either continuously throughout the day as basal insulin or as bolus doses for meals and correction of hyperglycemia. Meal carbohydrate estimates are usually acquired via electronic logbooks [57].

A. Short-term glucose prediction (less than 60 minutes)

Many ML algorithms have been developed for predicting glucose over short-term prediction horizons of 5-60 minutes and also over longer prediction horizons of 1-4 hours. Predicting glucose in the short-term of 5-60 minutes can be useful in automated insulin delivery systems that shut off insulin in response to predicted low glucose (Figure 6).

The prediction horizon over which the forecasting is being done can help indicate the best type of algorithm to use for the prediction. When predicting over 5-30 minutes, it is possible to estimate the glucose within a reasonable error tolerance (e.g. 14-24 mg/dl) using a regression-based algorithm (e.g. linear regression, support vector regression, long-short-term memory neural network, convolutional neural networks). One of the early algorithms for predicting glucose 30 minutes in the future was by Sparacino and colleagues who used an AR model and achieved an accuracy of about 18 mg/dL root mean squared error (RMSE) [101]. Turksoy et al. used an AR predictive model with exogenous inputs (ARX) to recommend carbohydrates if low glucose was predicted to reduce hypoglycemia, though no accuracy measures were included in the manuscript [102]. Random forests have also been used to predict glucose in the short-term with very low RMSE reported [131]. However, the RMSE reported in this manuscript (8.15 mg/dL) is lower than the typical error of a glucose sensor, which at the time of that publication was on the order of about 10-12%. The ML community including Schwartz-Ziv and Armon [132] showed that for tabular data, random forests tend to outperform deep learning methods when the number of observations is relatively small (e.g. < 1 million), which is typically the case for T1D data sets. Georga et al. published an algorithm on short-term glucose prediction using support vector regression with RMSE even lower at 6 mg/dL [133]. RMSE can be low if there is not much glucose variability in a data set. For this reason, it is important to report glucose variability along with RMSE. Neural network algorithms have also been used to predict short-term glucose [26, 104, 134-138]. Li et al. [134] utilized a convolutional recurrent neural network and achieved an RMSE of 9.38±0.71 mg/dL on simulated patient data and 21.07±2.35 mg/dL on real-world data. Results from this manuscript highlight the importance of evaluating glucose forecasting algorithms on *real-world data* compared with *simulated data* since the performance can be higher in simulation.

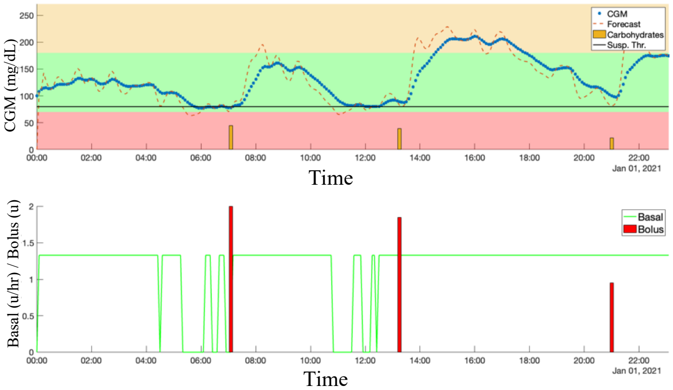


Figure 6: Top panel shows CGM over time (blue) and forecasted CGM (dashed red) where the pink region indicates hypoglycemia region (<70 mg/dL), the green region is a target glucose range (70-180 mg/dL), and beige is a hyperglycemia region (>180 mg/dL). Bottom panel shows meal insulin (red) and basal insulin in green. Notice that the basal insulin is turned off when CGM is predicted to go into the hypoglycemia region.

Common pitfall 8: RMSE can be found to be very low if the variability of the data set is low. When presenting RMSE results, it is also important to present information about the glucose variability such as standard deviation or coefficient of variation in the data set. Another option is to present accuracy results as the normalized RMSE whereby RMSE is normalized with respect to the inter-quartile range, the standard deviation, or the coefficient of variation.

Common pitfall 9: Presenting glucose forecasting accuracy results only on simulated glucose data can yield results that are overly optimistic. Introducing noise, missing data, outliers, and other disturbance artifacts can help to make simulated data sets more realistic. When presenting forecasting results on simulated data, glucose variability should also be reported along with information about how meal insulin dosing was done and how meal time were presented to the simulator, and how meal time and meal amount variability were imposed.

Common pitfall 10: When reporting algorithm accuracy results on a data set, it is important to clearly state that the algorithm was evaluated on either the entire data set or a subset of data. An explanation should be provided if only a subset was used.

Best practice 12: Always present final results on free-living real-world human data if available. If results are shown only on simulated data or on in-clinic data collected under prescribed conditions to a particular study design, this should be listed as a limitation in the results.

Best practice 13: When presenting results on a new glucose forecasting algorithm, it is important to compare it with best-in-class previously published algorithms and also include results of that algorithm on a benchmark data set for a comparison (Table 1). This is especially true if very low RMSE prediction results are found. This is possible if the prior publications have included code for implementing the algorithm. If the algorithm must be retrained, differences in performance could be expected compared with publication.

Best practice 14: When presenting results on a new algorithm, it is important to compare that algorithm's performance with the performance of a naïve algorithm including a zero-order hold predictor, a simple linear regression predictor, and a low-order autoregressive model. A zero-order hold algorithm simply assumes that glucose will not change in the future. Work on the Tidepool data set indicates that a zero-order-hold algorithm can achieve an RMSE of 25 mg/dL on 30-minute prediction for closed-loop data and 24 mg/dL on sensor-augmented pump data [26]. A simple linear extrapolation where you fit a regression line across the most recent 10 minutes can predict 30-minutes in the future with 21 and 20 mg/dL RMSE for closed-loop and sensor augmented pump in the Tidepool data sets, respectively [26]. In addition, a 3rd or 4th order autoregressive model can be used as a comparator model. Importantly, a zero-order hold algorithm will work well when there is less variability in a data set while a simple linear regression algorithm will work well when glucose is changing at a constant rate. It is important to compare algorithm performance with these naïve prediction algorithms.

Zecchin et al. [136] describe a *jump* neural network design for predicting short-term glucose. This jump neural network is a feed-forward, shallow neural network with one hidden layer of 5 neurons that have inputs connected to both the first layer but also to the output layer. Despite the small amount of data used to train the model, it achieved good performance with an RMSE of 16.6±3.1 mg/dL (mean±standard deviation) with a time gain (TG) of 18.5±3.4 minutes. Zecchin et al. [137] also demonstrated that improved performance could be obtained when including carbohydrate information as a feature. Pappada et al. [135] also proposed a shallow neural network for predicting glucose with a prediction horizon of 75 minutes. Their network had a single hidden layer with nine neurons. They used CGM, SMBG glucose, CGM trend information, insulin, carbs, and hypo/hyperglycemic symptoms, activities, and even emotional factors as inputs. Perez-Gandia [104] also used a shallow neural network to predict glucose 15, 30, and 45 minutes in the future using prior glucose data from up to 20 minutes before the prediction time. Results showed an RMSE of 18 mg/dL at 30 minutes with a prediction delay of approximately 9-15 minutes. All of these early publications on ML approaches to short-term glucose forecasting had very little data for training and evaluation. More recently, Mosquera-Lopez and Jacobs [26] showed that on a large real-world data set from the Tidepool Big Data Donation Data set with 175 people and 41,318 days of data from people on both closed-loop (CL) and sensor-augmented pump (SAP) therapy, a long short-term memory neural network could achieve an RMSE of 19.8±3.2 mg/dL (CL) and 19.6±3.8 mg/dL (SAP) for a 30-minute prediction horizon, with 99.6% of predications within the A+B zones of the Parkes Consensus grid. Because of the larger data size, the architecture was more complex with 5 hidden layers including an LSTM layer with 128 units, and 4 dense units with 64, 32, 16, 12 units respectively. The higher RMSE compared with other studies is because the model was trained and evaluated on real-world data across a large heterogeneous population of people with T1D. Simpler models were considered, but the higher complexity LSTM was best.

Common pitfall 11: Using an overly complex model that is trained on a small data set may not yield good performance. Time-series or tabular data with fewer than 1 million observations may be more accurately modeled using regression models or random forests than deep learning [132].

Best practice 15: If a data set is of limited size, select an ML algorithm that requires fewer parameters to tune such as AR/ARX models, support vector regression, random forest, or shallow neural networks. For larger data (e.g. Tidepool), more complex models can be used.

Best practice 16: When reporting on results on an algorithm, include the training time, model and number of CPUs/GPUs, and include the model and code in a repository.

Common pitfall 12: Presenting glucose forecasting results may not be accurate if the delay of the prediction algorithm is not also reported. Some algorithms impose a certain amount of delay in the prediction and this should be reported.

Best practice 17: When predicting glucose in the future, include an estimate of the time gain (TG) of the prediction model (defined below). TG is the prediction horizon over which prediction is desired, minus the delay of the prediction model. Delay in the prediction is found by performing cross-correlation of the actual glucose with the predicted signal.

$$TG = PH - Delay \tag{10}$$

B. Binary classifiers of hypoglycemic events and glucose prediction over longer-term horizons over 1 hour

Predicting glucose over 1-4 hours is more challenging within a reasonable accuracy. Kushner et al. [139] used shallow neural networks to predict glucose 1-4 hours in the future and they reported accuracy of 38±6 mg/dL for a 2-hour prediction horizon and 43±12 mg/dL at 4 hours. For longer prediction horizons, it may be optimal to instead predict binary events such as hypoglycemia. Regression algorithms can still be used for the prediction, but the algorithms are trained to only predict when a threshold is exceeded. For example, Mosquera-Lopez et al. described use of a support vector regression algorithm to predict hypoglycemia overnight up to 8 hours in advance. The algorithm had a sensitivity of 94.1% with a specificity of 72% and an area under the receiver operating characteristic curve (AUC-ROC) of 86%. The optimal threshold was selected using decision theory to optimize net benefit of acting on the classifier output whereby net benefit was defined as the negative of the sum of the low blood glucose index and the high blood glucose index (see section on metrics below). Jensen et al. [140] also designed a forecasting algorithm for predicting nocturnal hypoglycemia. They used a linear discriminant analysis classifier and achieved a sensitivity of 75% and specificity of 70%. Various ML algorithms are used for physical activity and psychological stress detection and characterization by Sevil et al. for glucose concentration estimation and in AID systems [98-100] and by Askari et al. for meal and exercise detection [25].

Common pitfall 13: When predicting glucose and the error is less than the error of the CGM (e.g., 8-10% for commercial CGM), there could be a problem with the algorithm.

Common pitfall 14: When predicting glucose or binary events, it is important to consider how unanticipated events may affect the prediction accuracy. For example, a prediction model may provide good accuracy in forecasting glucose 30 minutes in the future, but if a meal is consumed 5 minutes after the prediction is made, this will degrade the prediction accuracy since glucose may unexpectedly rise rapidly in response to the meal. When designing a forecasting algorithm, it is important to consider how unexplained events impact training and also evaluation of the algorithm.

Best practice 18: Prediction tasks and prediction metrics should be selected based on the forecast interval, and care should be taken in properly selecting features and prediction events depending on this forecast window. For long-term glucose prediction, it is challenging to achieve low error (RMSE or MARD) when predicting glucose values. When using classifiers instead, it is important to include the definition of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Outcomes will differ depending on these definitions. For example, in meal detection, a TP could be defined as a detection within 30 minutes after a meal has been ingested. For short-term hypoglycemia detection, a TP could detection of glucose dropping below 70 mg/dL for at least 15 minutes. In any case, the definition of the prediction task will must be defined. Error metrics like RMSE and MARD are more relevant prediction metrics when predicting short-term glucose (e.g., 30 minutes).

Although regression algorithms can be used to predict binary events (e.g., hypoglycemia or hyperglycemia), an algorithm used for binary classification will perform better if it is trained specifically to predict that event. For example, a glucose prediction algorithm may have a very low RMSE, but it may have a poor sensitivity for predicting low glucose readings. This is likely to occur because of the large imbalance in glucose data in the low glucose range (<70 mg/dl) compared with glucose above this range. Binary classifiers that are used to predict events like hypoglycemia need to be designed by dealing with the inherent imbalance in the data set.

It is helpful to use an outcome metric weighted by the glucose range in which the error was made. It is more dangerous to make an error in a low glucose range (<70 mg/dL) than in a normal glucose range (90-140 mg/d). Del Favero, Facchinetti and Cobelli [141] described a glucose-range-specific metric to better capture the increased risk of making errors in different ranges of glucose such that errors made in dangerous regions were weighted more heavily than those made in less dangerous regions. This group also showed in Faccioli et al. [142] that through use of this glucose-specific weighted error term, they could improve forecasting of hypoglycemia. Cameron et al. also report on a risk-based closed-loop algorithm that prioritizes mitigation of the increased risk of hypoglycemia compared with hyperglycemia [143].

Common pitfall 15: When using a regression algorithm to detect a rare event such as low glucose (e.g., < 70 mg/dL), a meal event, or an exercise event, sensitivity and specificity may be poor because the cost function for a regression algorithm is designed to minimize overall error, not detection of the hypoglycemia or meal event. These events are defined based on thresholds, such glucose dropping below a threshold of 70 mg/dL, glucose rising faster than 5 mg/dL/minute in response to a forecasted meal event. A cost function that penalizes based on glucose error alone may not be sufficient to optimize detection of the actual event of interest such as low glucose or a meal event. Furthermore, a person's glucose can vary right around the threshold defining the binary event (e.g., the threshold for low glucose is < 70 mg/dL, but glucose varies right around 70, 71, 68, etc.). When designing a binary classifier, it is important to consider points near the threshold that defines the binary event and how detection errors near this threshold are considered during training of the algorithm. Alternatively, a cost function that weights error near the threshold based on risk associated with that error may yield better performance (see Best Practice 33).

Best practice 19: When designing an algorithm to predict an event such as low glucose (<70 mg/dL), it may be optimal to design a binary classifier rather than a classifier based on a regression algorithm whereby there is a penalty for failing to identify the binary event. Alternatively, it could be optimal to use a glucose-range-specific penalty [141, 142].

C. Use of ML in detecting meal events

Another type of sparse event classifier in diabetes is detecting a meal event (Figure 7). Detecting meal ingestions using various features including CGM, insulin, and other features may be an important step towards enabling a fully-automated hormone delivery system. The ideal time to dose meal insulin is before the meal is consumed due to the delayed kinetics of insulin relative to carbohydrate absorption. This requires the person with T1D to remember to dose meal insulin and also to properly estimate the carbohydrates consumed in the meal. This is a burden leading some to miss mealtime insulin and for those that do take mealtime insulin, carbohydrate estimation is prone to errors [58]. However, if a meal can be detected using an ML algorithm, a portion of the meal insulin bolus can be dosed once a meal has been detected[144]. The amount of meal dosed in response to the detected meal should depend on the estimated size of the meal and the estimated time that the meal was taken. Most meal detection algorithms can detect the meal within 25-45 minutes of the consumption of the meal. Although this is not an ML method, Mahmoudi et al. [145] used a Kalman filter within a control framework to determine if a meal has been consumed. They demonstrated that when dosing for this missed meal using the UVA-Padova simulator, they could improve TIR from 53% to 83%. However, the algorithm still needs to be evaluated in a human study. Samadi and Cinar [146] reported on a qualitative trend analysis and fuzzy logic method for meal detection that achieved 87% sensitivity *in silico* and 93% sensitivity on actual human data. Their algorithm could detect a meal on average 34.8 minutes after the meal was consumed. One of the few ML-based meal detection algorithms was done by Garcia-Tirado et al. who showed that integrating a bolus priming system (BPS) to estimate the probability of a meal being consumed into their model predictive control algorithm improved time in range in an in-clinic human study [37]. Another ML approach for automated meal detection in T1D was based on an ensemble of LSTMs that received as input sequences of CGM records and classified the most recent CGM records as positive or negative for a meal onset. The *in silico* evaluation demonstrated the potential of the approach to achieve acceptable performance (mean c-statistic: >75%, mean detection time errors: 7-13 *min*) [147].

Since meal events are relatively rare (3-5 per day) compared with the total number of CGM readings in a day (~288), it is important that data scientists balance the data sets during training. A result of the imbalance in the data set is that accuracy can seem very high even for algorithms with poor sensitivity since there are so many non-meal events in a day. Therefore, reporting on balanced accuracy, area under the curve [148], and F1 are important metrics for evaluating meal detection. Lastly, specificity as a percentage is not that helpful for meal detection because of the data imbalance. Rather, the specificity of a meal detection algorithm should be reported as number of false positives per day.

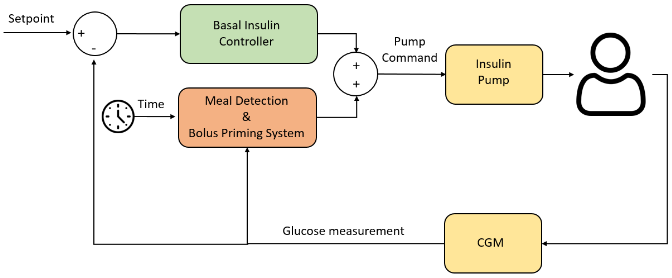


Figure 7: Block diagram of a fully-automated insulin delivery system including a basal insulin controller and a meal detection and bolus priming module.

Best practice 20: When designing an algorithm to detect a meal, it is important to report on the sensitivity, the specificity, and the area under the receiver operating characteristic curve (AUC-ROC) of the algorithm. Specificity should be reported as number of false positives per day. A false positive event for a meal detection algorithm may carry a significant risk of hypoglycemia, and the impact of acting on these false positive events should be evaluated in simulations. The additional outcome metrics to report on are covered in section III. The challenge with predicting sparse events such as a meal is that the prediction accuracy is highly dependent on the definition of the event. For example, a meal event can be defined as consuming a 10 g carbohydrate meal, or as consuming a 100 g carbohydrate meal, which yield vastly different glucose responses. It is far easier to detect a large carbohydrate meal compared with a small carbohydrate meal. Data scientists must carefully consider how the meal event is defined, how the detection window is defined, and how a detection event from the algorithm will be used when defining the definition for the meal size and when reporting accuracy.

Best practice 21: The mean time of the meal consumption relative to prediction time should be reported with all meal prediction algorithms. An optimal meal detection time is as soon as possible relative to the start of the meal consumption, though the change in glucose in response to a meal does not typically appear until 20-45 minutes after the meal. Insulin on board and macronutrient contents of the meal (protein and fat) will also affect the response to the meal [149].

Best practice 22: For automated meal detection, it may be important to both detect a meal and also to estimate the size of the meal, though this is application-specific. Certain meal forecasting algorithms may be used to automate meal insulin dosing in full closed-loop AID applications in response to detected meal events. Certain algorithms can utilize a size estimation as well to determine how much meal insulin should be dosed in response to the detected meal event [144]. Other algorithms that automate meal insulin dosing, may not require a meal size estimation to determine how to dose for meal insulin but instead use anticipation using probabilities derived from prior meal events [150]. Meal anticipation is possible as done by Garcia-Tirado and colleagues if conservative meal dosing is employed. It is important to report the sensitivity and false positives per day of the meal detection event and optionally on the accuracy of the size estimation algorithm if size estimation is important for the application.

D. Hyperparameter tuning

In all ML algorithms, there are hyperparameters that need to be tuned. These hyperparameters may include the features used, aspects of the architecture such as the number of layers or nodes in a layer of a neural network, or the selection of gamma, C, or the kernel in support vector machines. The hyperparameters can be tuned through a search across the parameter space, such as grid search. It is important for data scientists to (1) report final hyperparameters of their model and (2) clearly report how hyperparameters were tuned (e.g. using a validation set or N-fold cross validation).

Common Pitfall 16: Finalizing a model structure without hyperparameter tuning or not discussing how the hyperparameters were chosen could lead to sub-optimal model performance and lack of reproducibility.

Best Practice 23: Hyperparameters in a model should be tuned. Examples of hyperparameter tuning approaches may include grid search, random search, evolutionary algorithms, or Bayesian sampling. Data scientists should clearly state how hyperparameter tuning was done to enable reproducibility.

E. Considerations on types of sensors and devices used

Different studies are done using various types of commercial and/or developmental sensors and insulin delivery devices. This should be considered when comparing ML algorithms trained and evaluated from different studies using different devices. The accuracy of different CGM sensors under different conditions (e.g. during exercise, or with certain types of medication) should be considered.

Common Pitfall 17: Comparing algorithms built on glucose data collected using different CGM sensors may yield inconclusive results. CGM manufacturers use different algorithms, that may introduce bias on the CGM data. Data scientists should report on the devices used in the study and be aware that performance may differ if algorithms are evaluated on devices on which the algorithm was not trained.

Best Practice 24: ML algorithms designed using data from different models may need to consider manufacturer and model as inputs so that bias differences between the CGM manufacturers can be appropriately handled by the algorithm.

F. Personalized ML in diabetes forecasting and treatment

There is heterogeneity in physiology, glucose responses, insulin dosing, nutrient intake, exercise patterns, sleep patterns etc. in people living with diabetes. This large amount of heterogeneity can make it challenging to design an ML algorithm that will work well for everyone. Adapting a population model once new data from an individual becomes available may lead to significant improvements in accuracy. Romero-Ugalde et al. developed an autoregressive with exogenous inputs (ARX) model for glucose prediction during and after exercise [151]. They found that personalization could improve the prediction accuracy. Tyler and colleagues also found that using data recorded from prior exercise sessions could be used to personalize a glucose forecasting algorithm and improve accuracy during future exercise events [71]. Askari et al. used the Tidepool data set to develop personalized DNN models for meal and physical activity detection [25]. Reinforcement learning could be an important approach at personalizing glucose control algorithms in the future as discussed by Tejedor and colleagues [152] and Fox and colleagues [153]. However, it remains an open question of whether population models are better than personalized models when used in the real-world. Individuals' glucose profiles change a lot day-to-day. Herrero and colleagues showed that individuals' glucose traces could uniquely identify each person thereby generating a CGM equivalent to a 'fingerprint' [154]. If individual glucose profiles have a larger inter-day variation than the variation observed in a population model, then there could be no benefit of personalization. It is important for any personalized model to be compared with a population model.

Best practice 25: When presenting results on a personalized ML model, compare performance with a population model to show the benefit of personalization. Architecture should be identical between personalized and population models.

V. Applications of ML in Decision Support and Closed-Loop

ML has the potential to improve glucose control beyond just glucose forecasting. ML techniques can be used to optimally provide recommendations to patients about modifying their insulin dosing settings in closed-loop systems or using open-loop therapy.

A. ML algorithms used in decision support

Noaro et al. [155, 156] trained multiple ML models to estimate the optimal pre-meal bolus. The models were shown to improve glucose control *in silico* and results were also presented on retrospective real-world data.

Though not an ML approach, a common approach for selecting optimal recommendations for adjusting insulin settings is to use a run-to-run approach whereby settings are selected based on the similarities of the circumstances with prior circumstances. Various groups [157-160] have demonstrated how run-to-run methods can be used to provide decision support and improve glucose outcomes *in silico*.

Similar to run-to-run methods, Tyler et al. [20] developed a k-nearest-neighbors decision support system (KNN-DSS) to provide weekly optimal recommendations to help people with T1D on multiple daily injections (MDI) to modify correction factors and carbohydrate ratios. The algorithm was trained on 50,000 *in silico* observations derived via simulator [68] and then evaluated in a short 4-week real-world study that demonstrated reduction of time in hypoglycemia with use of the recommendations compared with a baseline period. Tyler et al. compared the recommendations provided by the KNN-DSS with recommendations provided by physicians and found that the algorithm agreed with a consensus of board-certified endocrinologists 67.9% of the time. This study emphasized several important best practices and also pitfalls important for ML algorithms designed for use in decision support in diabetes. First, while the algorithm was trained on a simulator, it was evaluated on clinical study data and results were reported on both in silico and real-world data (Best Practice #12). Second, a comparison was done between the algorithm's recommendations with those of board-certified endocrinologists (Best Practice #26).

Best Practice 26: When evaluating an ML-driven recommender engine, comparing a recommender engine with recommendations by board certified endocrinologists can provide some assurance of acceptable safety.

More recently, Castle et al. [161] reported that the Tyler et al. algorithm was evaluated in an 8-week trial in people with T1D. There was no improvement in the percent time in target glucose range (70-180 mg/dl) in the final 2 weeks of the study compared with the baseline two weeks. However, for weeks when participants followed most or all of the recommendations, they realized a 6.3% increased time in range compared with weeks when they did not follow the recommendations. Importantly, the *in silico* evaluation in the UVA Padova simulator done by Tyler et al. showed a 6.7% increase in time in range, which is close to the improvement shown in the Castle et al. study of 6.3%. However, when Tyler et al. evaluated the performance in the OHSU simulator using different virtual participants than the one on which the algorithm was trained, the performance was higher demonstrating an expected improvement in percent time in range of 20.3%. The variable performance across simulators emphasizes the need to evaluate algorithms across multiple simulators.

Best practice 27: If a decision support or control algorithm is trained on virtual participants from a given simulator, it is important to evaluate it on virtual participants in a different simulator, which should be at least comparable in terms of specifications to the first one. If a second simulator is not available, a data scientist may use a sub-sample of virtual participants from the same simulator not used to train the algorithm, and test on this sub-sample. However, results will likely be overly optimistic. It is also important to consider how closely the scenarios in a test set are represented by those in the training set.

Scenarios should be well represented in both training and test sets and should be close to real-world (e.g. utilizing real-world meal amounts, meal size misestimations, meal timing and exercise intensities, durations, and timing).

Best practice 28: When training and evaluating a clinical decision support or control algorithm, it is important to simulate lack of adherence to recommendations when determining changes in glucose outcomes. In decision support, it has been shown that people tend to not to adhere to recommendations about 25% of the time [161]. Lack of adherence in a closed-loop setting could be scenarios where a participant does not adequately follow meal bolus recommendations or does not dose for meals altogether in a hybrid closed-loop application. Lack of adherence in a clinical decision support system could be where a person does not announce a meal or uses an incorrect carbohydrate ratio or correction factor and does not change it when recommended to do so [162].

Additional work on clinical decision support systems (CDSS) was reported by Nimri et al. [163] on a clinical decision support algorithm (AI-CDSS) for use in adjusting insulin pump settings for open loop insulin therapy. This algorithm uses fuzzy logic rather than ML. Recommendations were provided to the participants once every 3 weeks for adjusting carb ratios, correction factors, and basal insulin rates. This CDSS was cleared by the FDA. Nimri et al. reported non-inferiority data comparing AI-CDSS with board certified endocrinologists in patients using MDI therapy [164].

Bisio et al. [165] showed that 80 participants with T1D using CGM + MDI therapy in combination with a decision support also did not improve their time in range compared with participants using CGM + MDI without decision support. However, as with Castle et al., they also showed that 'active users' of the app experienced a higher time in range than a group that was not defined as 'active users'. This further supports the need to model adherence and lack-of-adherence when evaluating expected benefit of use of decision support.

Best Practice 29: A decision support algorithm providing recommendations on changes to carb ratios, correction factors or basal rates should require a minimum amount of historical data prior to making a subsequent recommendation. Because of the variability of CGM and day-to-day behavior as well as the differences in behavior during the week vs. weekend, the recommendation is to require at least 1 week of historical data prior to making a recommendation. Herrero and colleagues found that a minimum of two weeks of observing CGM when people use ultra-long-acting insulin (e.g. Tresiba and Toujeo which reach steady state in 3-4 days) is sufficient for estimating glucose outcomes [166]. Therefore, multiple weeks could be required to assess glucose outcomes prior to providing a recommendation for these types of basal insulin. In addition, improvement should be assessed across multiple weeks using a statistical test as a comparison against another algorithm (e.g. standard care). The amount of time needed between recommendations is likely application-dependent and should optimally be determined through hyper-parameter tuning.

Best practice 30: When assessing a clinical decision support algorithm, it is best to compare glucose outcomes (Table 2) relative to a baseline time period. A recommended baseline window would be a minimum of 2 weeks [167]. Menstrual cycle can affect glucose levels, and for that reason, it would be beneficial to have at least 4 weeks in a baseline period.

Best practice 31: When comparing different algorithms, it is important to do a statistical test (e.g. t-test or a general linear model) to show if an improvement is significant. A test of normality should also be done if using a t-test to ensure normality of the distribution, otherwise a non-parametric test (e.g. Wilcoxon rank-sum test) should be done.

B. ML algorithms used in closed-loop control

ML algorithms can be used within systems that automate delivery of insulin and other hormones [168]. A model predictive control algorithm requires a forecasting model that is used to predict glucose across a future time horizon [169]. This forecasting model can be a physical model comprising ordinary differential equations as was done in Hovorka et al. [108] and by others [66, 170], or the model may also include a data-driven model as was done by Zarkogianni and colleagues [171] where they combined a compartment model with a recurrent neural network. ML algorithms can also be used within a closed-loop framework to predict a low glucose event as was done using a long-short-term memory neural network [26] within a dual-hormone closed loop system to shut off insulin in the event that a low glucose event is predicted [16]. ML algorithms can also be used to predict when a meal has occurred such that a percentage of meal insulin can be delivered when the meal is detected or to detect an exercise event so that insulin can be reduced during or following exercise [144, 150]. The application in which an algorithm is intended to be used is critical to understand when designing it because the resulting actions of its use can be catastrophic if the algorithm is incorrect. For example, depending on the dosing strategy, when designing a meal detection algorithm designed to dose insulin in the event of a non-reported meal, it is optimal to minimize or eliminate false positives at the expense of reduced sensitivity, since dosing insulin in response to a false positive meal detection could potentially result in severe hypoglycemia, which can be life-threatening. Whereas, a hypoglycemia prediction algorithm should be optimized for higher sensitivity because a hypoglycemic event is a dangerous event; if the algorithm misses a prediction, and insulin is not reduced by the closed-loop system, then it again could be harmful to the user.

Best practice 32: When designing ML algorithms used within a closed-loop system, the outcome metrics [172] should be carefully considered to minimize risk to the user.

VI. Metrics used to evaluate algorithms

A. Glucose forecasting metrics

Kovatchev provides a review of metrics that can be used to assess outcomes of predictive algorithms [51]. When predicting glucose, the most common metrics used to assess performance are root mean squared error (RMSE), mean absolute error (MAE), mean absolute relative difference (MARE), mean error (ME), mean relative error (MRE), and time gain (TG). In the equations below, $\hat{y}(k+PH)$ is the predicted glucose at time k, for a given prediction horizon (PH) while y(k+PH) is the measured glucose at that time.

$$RMSE = \sqrt{\frac{1}{N}\left(\sum_1^N(\hat{y}(k+PH) - y(k+PH))^2\right)} \qquad (11)$$

$$MAE = \frac{1}{N}\sum_1^N|\hat{y}(k+PH) - y(k+PH)| \qquad (12)$$

$$MARE = \frac{1}{N}\sum_1^N\left|\frac{\hat{y}(k+PH)-y(k+PH)}{y(k+PH)}\right| \qquad (13)$$

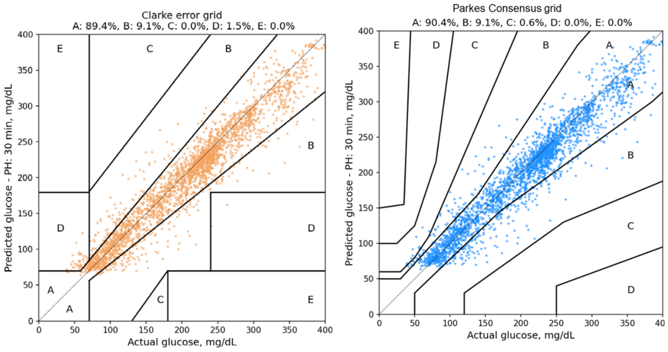
$$ME = \frac{1}{N}\sum_1^N \hat{y}(k+PH) - y(k+PH) \qquad (14)$$

$$MRE = \frac{1}{N}\sum_1^N\frac{\hat{y}(k+PH)-y(k+PH)}{y(k+PH)} \qquad (15)$$

Notice that RMSE and MAE are very similar measures. However, RMSE includes the square root of the sum of the square of errors, which yields an error metric that is more sensitive to large deviations. A large glucose prediction error can be life-threatening to a person with T1D using that prediction to dose insulin. RMSE should be the primary metric used to present results on glucose forecasting.

Common pitfall 18: When reporting on absolute error for glucose forecasting, MAE may not provide a complete picture as it does not weight large mistakes in prediction like RMSE.

Best practice 33: When reporting on an absolute error for glucose forecasting performance, include RMSE as the primary reporting measure. It may also be helpful to report glucose-specific RMSE



(a) (b)

Figure 8: Clarke error grid (a) and Parkes Consensus grid (b).

(gRMSE) which weights the risk associated with errors in different glucose ranges [141].

When used in glucose prediction, RMSE and MAE are typically in units of either *mg/dL* or *mmol/*L. MARE is helpful in that it is given as a percentage, which is an important way to display the data since a 30 *mg/dL* error in glucose estimation is far more concerning when glucose is low compared with when glucose is high, for example. ME and MRE provide information about the bias of a prediction estimate, indicating if the prediction tends to be lower or higher than the actual value on average. While designers may prefer to have an algorithm that has no bias, it is important to consider that a negative bias could be preferable to a positive bias. Hypoglycemia can be life-threatening while some amount of hyperglycemia can be tolerated, and so an algorithm that errs on the side of estimating glucose as too low, is actually safer than an algorithm that estimates glucose as higher than the actual value. It is important to limit false positives as this can cause alarm fatigue.

Common pitfall 19: Presenting prediction results only as an absolute value (e.g. RMSE, MARE, or MAE) is not sufficient because there is no information about bias in the prediction.

Common pitfall 20: When calculating an error metric such as RMSE for a given prediction horizon (e.g. 30 minutes), averaging the RMSE across multiple horizons yields an error that is incorrectly small. For example, if a data scientist was to report an RMSE for a 30-minute horizon as an average of the RMSE at 5-, 10-, 15-, 20-, 25-, and 30-min horizons, the RMSE will be unrealistically lower than if reported solely at 30 min.

Best practice 34: In addition to including absolute outcome metrics of RMSE, MARE, and MAE, also include relative error measures including MRE and ME. One of the disadvantages of including mathematical metrics of accuracy such as RMSE, MARE, MAE, MRE, and ME is that there is no penalization for errors that are made in regions that are dangerous when predictions are used by an insulin dosing algorithm or a decision support algorithm. Del Favero et al. [141] introduced a new metric called the glucose-specific mean squared error (gMSE) that penalizes mean squared error more significantly in low and high regions of glucose where the impact of an error is more clinically significant. If the glucose is less than a lower threshold of 70 mg/dL (TL), then a penalty is applied.

The Clarke error grid [173], the Parkes consensus grid [174], the continuous error grid [175], and the surveillance error grid [176] are other important tools that can be used to demonstrate how bias in a prediction is more or less clinically relevant. The Clarke error grid shown in Figure 8a [173] is a plot of predicted vs. reference glucose values over the range of 0-400 mg/dL (0-22.2 mmol/l). Regions of the plot are indicated by the letters A, B, C, D, and E whereby regions A and B are considered safe and regions C, D, and E are considered progressively more dangerous and even life-threatening if a person or device was to act on these inaccurate predictions. The Clarke error grid was originally designed to help people with diabetes gauge awareness their glucose levels. The Parkes Consensus grid [174] in Figure 8b and the surveillance error grid also have the regions A-E, but the boundaries are smooth and were determined by a consensus among a group of endocrinologists who agreed on the regions that were most dangerous to a patient if a mistake in forecasting is made. Importantly, the risk attributed to the different regions of these grids was designed for decision making based on real-time blood glucose meter measurements, not predictions made 30-60 minutes in the future. Because risk based on forecasting errors is likely different than risk based on real-time blood glucose measurements, there is an opportunity in the future to re-think how risk should be quantified for glucose forecasting.

Best Practice 35: Present glucose prediction results in the form of a Parkes consensus grid or surveillance error grid figure while also summarizing results in a table showing the percent of predictions in the A, B, C, D, and E regions.

B. Metrics used for sparse event prediction (e.g. hypoglycemia, hyperglycemia, meal events, etc.)

When assessing accuracy for a sparse event / binary classifier, it is important to account for the potential imbalance in the data set. For example, if there are 4% CGM readings below 70 mg/dL, an algorithm which always predicts that the glucose is not hypoglycemic will therefore have an accuracy of 96%, even if it misses every hypoglycemic event. For this reason, it is important to report both sensitivity and specificity as well as area under the receiver operating characteristic curve along with accuracy measures for sparse event classifiers (see Supplementary Equations 3.1-3.5).

In addition, rather than reporting accuracy, it is important to report prediction accuracy using a metric that is robust to this imbalance. Balanced accuracy is the arithmetic mean of the sensitivity and specificity, and so it is a more robust metric to imbalance than simply reporting accuracy. Another metric robust to imbalance for binary classifiers is Matthews Correlation Coefficient (MCC) [177] (Supplementary Equation 3.5).

Supplementary Equation 3.5 shows that when the classifier is perfect (FP = FN = 0) the value of MCC is 1, indicating perfect accuracy. Conversely, when the classifier always misclassifies (TP = TN = 0), we get a value of -1, representing perfect negative correlation (in this case, you can simply reverse the classifier's outcome to get the ideal classifier). MCC value is always between -1 and 1, with 0 meaning that the classifier is no better than a random flip of a fair coin. MCC is also symmetric, so no class is more important than the other.

Common pitfall 21: If accuracy is reported on a sparse event classifier algorithm in a highly unbalanced data set with an excessive number of negative observations, it is possible to have a very high accuracy, but a very low or even zero value for sensitivity of detecting the positive event. The definitions of the thresholds used to define the sparse event is critical; it will impact the sensitivity and specificity of the algorithm.

Best practice 36: When data sets are being used for a sparse event predictor, and if there is a large class imbalance between positive and negative observations, report sensitivity, specificity (FP/day), area under the ROC curve [148], and balanced accuracy such as MCC.

C. Metrics for evaluating postprandial glucose prediction

When predicting the glucose response after a meal (i.e. postprandial glucose response), there are various metrics that are of interest and can be used to determine if the amount of insulin dosed for the meal was appropriate. The area under the curve (AUC) of the CGM trace is calculated by using the trapezoidal rule and summing the area under this curve from the start of a meal to 3-4 hours after the meal. The AUC is not appropriate to assess postprandial glucose response because it inherently includes the starting glucose in the calculation, which is unrelated to the meal response. A better metric is the incremental area under the curve (iAUC) of the glucose trace, which sums the area under the CGM curve relative to the starting glucose value, only including in the sum the CGM values that are greater than the starting glucose value (Supplementary Figure 1). An additional metric is the netAUC which is the same as the iAUC however it also sums the negative areas under the curve relative to the starting CGM. Brouns et al. provide an overview of AUC, iAUC, and netAUC [178]. If the insulin dosing for the meal is optimal, then there will be no low glucose following the meal (<70 mg/dL) and the iAUC and netAUC will be minimal.

In addition to iAUC and netAUC, the maximum postprandial glucose, minimum postprandial glucose, and the delta between the peak and starting glucose are also useful.

Common pitfall 22: AUC as a postprandial glucose metric can be deceptive because it is correlated with starting glucose.

Best practice 37: When assessing prediction of postprandial glucose responses, appropriate metrics are the iAUC, netAUC, maximum delta glucose from the glucose at the start of the meal. All account for glucose at the start of the meal.

D. Metrics for assessing explainability of algorithms

Regulatory bodies like the Food and Drug Administration (FDA) in the U.S. will typically require that ML algorithms used within life-critical operations such as drug delivery maintain a certain level of explainability and interpretability. Interpretability implies that it is possible to understand how an algorithm arrived at giving a certain prediction or recommendation. Explainability describes how certain aspects, parameters, or features in a model influence the output of that model. Simple classes of algorithms (e.g. logistic regression, decision tree) are inherently interpretable. However, other algorithms such as random forest and deep learning algorithms are not interpretable, but explainability can be incorporated into them. Some of the popular methods of incorporating explainability into complex black-box ML algorithms include SHAP [179], LIME [180], DeepLIFT [181], MACE [182], GAN based methods [183]. These tools help illuminate ML models making predictions more comprehensible.

SHAP stands for SHapley Additive exPlanations and is more widely used and more similar to human explanations. The core idea behind Shapley value-based explanations of ML models is to use fair allocation results from cooperative game theory to allocate credit for a model's output among its input features.

Best practice 38: Data scientists should strive to develop algorithms that are explainable and interpretable. When using algorithms that are not inherently explainable, methods like SHAP, LIME, etc. should be used to provide a certain measure of explainability into the algorithm.

Common pitfall 23: When features are correlated, if the algorithm is used to identify input variables that are significantly related to the outcome, these correlations may lead to conclusions that are incorrect. For example, people with T1D take insulin at the same time as consuming a meal. In this way, both insulin and meal intake are correlated with rises in glucose. However, insulin does not cause the rise in glucose. The meal causes the rise. When exploring explainability of a model, it is important to be aware of correlated features to avoid invalid conclusions about which features are impacting prediction.

VII. Concluding remarks

In this manuscript, we have presented an overview of current best practices and common pitfalls for data scientists interested in working on the development of AI and ML algorithms for diabetes and glucose management. Future guidelines may include best practices and pitfalls as they relate to newer technologies such as adaptive therapies [184, 185], adjunctive therapies such as SGLT-2 inhibitors [186], multi-hormone closed-loop systems (insulin, glucagon [16, 187], pramlintide [34]) and use of cloud-based computing approaches vs. computing on the edge [188, 189]. Our aim is to present current consensus-based guidelines and recommendations that will aid in the advancement of the field to ultimately improve glucose outcomes and overall health of people living with diabetes.

VIII. References

[1] A. Katsarou, *et al.*, "Type 1 diabetes mellitus," *Nat Rev Dis Primers,* vol. 3, p. 17016, Mar 30 2017.

[2] M. A. Atkinson, *et al.*, "Type 1 diabetes," *Lancet,* vol. 383, pp. 69-82, Jan 4 2014.

[3] K. K. Dhatariya, *et al.*, "Diabetic ketoacidosis," *Nat Rev Dis Primers,* vol. 6, p. 40, May 14 2020.

[4] J. L. Chiang, *et al.*, "Type 1 diabetes through the life span: a position statement of the American Diabetes Association," *Diabetes Care,* vol. 37, pp. 2034-54, Jul 2014.

[5] "Gestational diabetes mellitus," *Diabetes Care,* vol. 27 Suppl 1, pp. S88-90, Jan 2004.

[6] R. Hovorka, "Closed-loop insulin delivery: from bench to clinical practice," *Nature Reviews Endocrinology,* vol. 7, pp. 385-395, 2011/07/01 2011.

[7] R. M. Bergenstal, *et al.*, "Safety of a Hybrid Closed-Loop Insulin Delivery System in Patients With Type 1 Diabetes," *JAMA,* vol. 316, pp. 1407-1408, Oct 4 2016.

[8] S. A. Brown, *et al.*, "Multicenter Trial of a Tubeless, On-Body Automated Insulin Delivery System With Customizable Glycemic Targets in Pediatric and Adult Participants With Type 1 Diabetes," *Diabetes Care,* vol. 44, pp. 1630-1640, 2021.

[9] S. A. Brown, *et al.*, "Six-Month Randomized, Multicenter Trial of Closed-Loop Control in Type 1 Diabetes," *N Engl J Med,* vol. 381, pp. 1707-1717, Oct 31 2019.

[10] L. M. Wilson, *et al.*, "Opportunities and challenges in closed-loop systems in type 1 diabetes," *The Lancet Diabetes & Endocrinology,* vol. 10, pp. 6-8, 2022/01/01/ 2022.

[11] O. Moser, *et al.*, "Glucose management for exercise using continuous glucose monitoring (CGM) and intermittently scanned CGM (isCGM) systems in type 1 diabetes: position statement of the European Association for the Study of Diabetes (EASD) and of the International Society for Pediatric and Adolescent Diabetes (ISPAD) endorsed by JDRF and supported by the American Diabetes Association (ADA)," *Diabetologia,* Oct 13 2020.

[12] B. P. Kovatchev, *et al.*, "Evening and overnight closed-loop control versus 24/7 continuous closed-loop control for type 1 diabetes: a randomised crossover trial," *Lancet Digit Health,* vol. 2, pp. e64-e73, Feb 2020.

[13] B. Lobo, *et al.*, "A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series," *IEEE Transactions on Biomedical Engineering,* vol. 69, pp. 654-665, 2022.

[14] M. K. Bothe, *et al.*, "The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas," *Expert Review of Medical Devices,* vol. 10, pp. 661-673, 2013/09/01 2013.

[15] T. Zhu, *et al.*, "Basal Glucose Control in Type 1 Diabetes Using Deep Reinforcement Learning: An In Silico Validation," *IEEE Journal of Biomedical and Health Informatics,* vol. 25, pp. 1223-1232, 2021.

[16] L. M. Wilson, *et al.*, "Dual-Hormone Closed-Loop System Using a Liquid Stable Glucagon Formulation Versus Insulin-Only Closed-Loop System Compared With a Predictive Low Glucose Suspend System: An Open-Label, Outpatient, Single-Center, Crossover, Randomized Controlled Trial," *Diabetes Care,* vol. 43, pp. 2721-2729, Sep 9 2020.

[17] T. Zhu, *et al.*, "A Deep Learning Algorithm for Personalized Blood Glucose Prediction," in *KHD@IJCAI*, 2018.

[18] K. Turksoy, *et al.*, "Multivariable Artificial Pancreas for Various Exercise Types and Intensities," *Diabetes Technol Ther,* vol. 20, pp. 662-671, Oct 2018.

[19] N. S. Tyler and P. G. Jacobs, "Artificial Intelligence in Decision Support Systems for Type 1 Diabetes," *Sensors,* vol. 20, p. 3214, 2020.

[20] N. S. Tyler, *et al.*, "An artificial intelligence decision support system for the management of type 1 diabetes " *Nature Metabolism,* pp. 612-619, 2020.

[21] P. Pesl, *et al.*, "Case-Based Reasoning for Insulin Bolus Advice," *J Diabetes Sci Technol,* vol. 11, pp. 37-42, Jan 2017.

[22] M. Reddy, *et al.*, "Clinical Safety and Feasibility of the Advanced Bolus Calculator for Type 1 Diabetes Based on Case-Based Reasoning: A 6-Week Nonrandomized Single-Arm Pilot Study," *Diabetes Technol Ther,* vol. 18, pp. 487-93, Aug 2016.

[23] V. Gulshan, *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA,* vol. 316, pp. 2402-2410, 2016.

[24] F. Arcadu, *et al.*, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," *npj Digital Medicine,* vol. 2, p. 92, 2019/09/20 2019.

[25] M. R. Askari, *et al.*, "Detection of Meals and Physical Activity Events From Free-Living Data of People With Diabetes," *Journal of Diabetes Science and Technology,* p. 19322968221102183, 2022.

[26] C. Mosquera-Lopez and P. G. Jacobs, "Incorporating Glucose Variability into Glucose Forecasting Accuracy Assessment Using the New Glucose Variability Impact Index and the Prediction Consistency Index: An LSTM Case Example," *Journal of Diabetes Science and Technology,* vol. 0, p. 19322968211042621, 2022.

[27] R. Reddy, *et al.*, "Prediction of Hypoglycemia During Aerobic Exercise in Adults With Type 1 Diabetes," *J Diabetes Sci Technol,* vol. 13, pp. 919-927, Sep 2019.

[28] J. Daniels, *et al.*, "A Deep Learning Framework for Automatic Meal Detection and Estimation in Artificial Pancreas Systems," *Sensors,* vol. 22, p. 466, 2022.

[29] K. Turksoy, *et al.*, "Multivariable adaptive closed-loop control of an artificial pancreas without meal and activity announcement," *Diabetes Technol Ther,* vol. 15, pp. 386-400, May 2013.

[30] K. Turksoy, *et al.*, "Use of Wearable Sensors and Biometric Variables in an Artificial Pancreas System," *Sensors (Basel),* vol. 17, Mar 7 2017.

[31] F. Cameron, *et al.*, "Probabilistic evolving meal detection and estimation of meal total glucose appearance," *J Diabetes Sci Technol,* vol. 3, pp. 1022-30, Sep 2009.

[32] F. M. Cameron, *et al.*, "Closed-Loop Control Without Meal Announcement in Type 1 Diabetes," *Diabetes Technol Ther,* vol. 19, pp. 527-532, Sep 2017.

[33] E. Dassau, *et al.*, "Detection of a Meal Using Continuous Glucose Monitoring," *Implications for an artificial β-cell,* vol. 31, pp. 295-300, 2008.

[34] A. Haidar, *et al.*, "A Novel Dual-Hormone Insulin-and-Pramlintide Artificial Pancreas for Type 1 Diabetes: A Randomized Controlled Crossover Trial," *Diabetes Care,* p. dc191922, 2020.

[35] E. Palisaitis, *et al.*, "A Meal Detection Algorithm for the Artificial Pancreas: A Randomized Controlled Clinical Trial in Adolescents With Type 1 Diabetes," *Diabetes Care,* vol. 44, pp. 604-606, Feb 2021.

[36] M. Zheng, *et al.*, "Automated meal detection from continuous glucose monitor data through simulation and explanation," *Journal of the American Medical Informatics Association,* vol. 26, pp. 1592-1599, 2019.

[37] J. Garcia-Tirado, *et al.*, "Advanced Closed-Loop Control System Improves Postprandial Glycemic Control Compared With a Hybrid Closed-Loop System Following Unannounced Meal," *Diabetes Care,* vol. 44, pp. 2379-2387, 2021.

[38] A. Z. Woldaregay, *et al.*, "Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes," *J Med Internet Res,* vol. 21, p. e11030, May 1 2019.

[39] M. R. Askari, *et al.*, "Adaptive-learning model predictive control for complex physiological systems: Automated insulin delivery in diabetes," *Annual Reviews in Control,* vol. 50, pp. 1-12, 2020/01/01/ 2020.

[40] "Good to Know: Factors Affecting Blood Glucose," *Clinical Diabetes,* vol. 36, pp. 202-202, 2018.

[41] I. Kavakiotis, *et al.*, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal,* vol. 15, pp. 104-116, 2017/01/01/ 2017.

[42] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Vis Comput Ind Biomed Art,* vol. 4, p. 30, Dec 3 2021.

[43] T. Zhu, *et al.*, "Deep Learning for Diabetes: A Systematic Review," *IEEE Journal of Biomedical and Health Informatics,* vol. 25, pp. 2744-2757, 2021.

[44] M. Niederberger and J. Spranger, "Delphi Technique in Health Sciences: A Map," *Front Public Health,* vol. 8, p. 457, 2020.

[45] S. Basu, *et al.*, "Use of Machine Learning Approaches in Clinical Epidemiological Research of Diabetes," *Curr Diab Rep,* vol. 20, p. 80, Dec 3 2020.

[46] I. Contreras and J. Vehi, "Artificial Intelligence for Diabetes Management and Decision Support: Literature Review," *J Med Internet Res,* vol. 20, p. e10775, May 30 2018.

[47] S. Larabi-Marie-Sainte, *et al.*, "Current Techniques for Diabetes Prediction: Review and Case Study," *Applied Sciences,* vol. 9, p. 4604, 2019.

[48] J. Li, *et al.*, "Application of Artificial Intelligence in Diabetes Education and Management: Present Status and Promising Prospect," *Front Public Health,* vol. 8, p. 173, 2020.

[49] M. Phillip, *et al.*, "The Digital/Virtual Diabetes Clinic: The Future Is Now-Recommendations from an International Panel on Diabetes Digital Technologies Introduction," *Diabetes Technol Ther,* vol. 23, pp. 146-154, Feb 2021.

[50] M. A. Makroum, *et al.*, "Machine Learning and Smart Devices for Diabetes Management: Systematic Review," *Sensors (Basel),* vol. 22, Feb 25 2022.

[51] B. P. Kovatchev, "Metrics for glycaemic control — from HbA1c to continuous glucose monitoring," *Nature Reviews Endocrinology,* vol. 13, pp. 425-436, 2017/07/01 2017.

[52] A. Z. Woldaregay, *et al.*, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artif Intell Med,* vol. 98, pp. 109-134, Jul 2019.

[53] C. Marling and R. Bunescu, "The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020," *CEUR Workshop Proc,* vol. 2675, pp. 71-74, Sep 2020.

[54] Tidepool. (1/13/2022). *Tidepool Big Data Donation Data Set*. Available: https://www.tidepool.org/bigdata

[55] C. Mosquera-Lopez, *et al.*, "Predicting and Preventing Nocturnal Hypoglycemia in Type 1 Diabetes Using Big Data Analytics and Decision Theoretic Analysis," *Diabetes Technol Ther,* May 14 2020.

[56] M. R. Askari, *et al.*, "Meal and Physical Activity Detection from Free-Living Data for Discovering Disturbance Patterns of Glucose Levels in People with Diabetes," *BioMedInformatics,* vol. 2, pp. 297-317, 2022.

[57] M. Riddell, *et al.*, "The Type 1 Diabetes EXercise Initiative (T1DEXI): Examining the acute glycemic effects of different types of structured exercise sessions in type 1 diabetes in a real-world setting," *Diabetes Care,* vol. In press, 2023.

[58] M. Gillingham, *et al.*, "Assessing Mealtime Macronutrient Content: Patient Perceptions versus Expert Analyses via a Novel Phone App," *Diabetes Technol Ther,* Aug 24 2020.

[59] J. C. f. H. Research. (1/13/2022). *Listing of available T1D data sets*. Available: https://public.jaeb.org/datasets/diabetes

[60] C. Cobelli and E. Carson, *Introduction to modeling in physiology and medicine* vol. 2: Elsevier, 2019.

[61] J. Castle, *et al.*, "Randomized outpatient trial of single and dual-hormone closed-loop systems that adapt to exercise using wearable sensors," *Diabetes Care,* vol. 41, pp. 1471-1477, 2018.

[62] A. Haidar, *et al.*, "Glucose-responsive insulin and glucagon delivery (dual-hormone artificial pancreas) in adults with type 1 diabetes: a randomized crossover controlled trial," *CMAJ,* vol. 185, pp. 297-305, Mar 5 2013.

[63] A. Haidar, *et al.*, "Comparison of dual-hormone artificial pancreas, single-hormone artificial pancreas, and conventional insulin pump therapy for glycaemic control in patients with type 1 diabetes: an open-label randomised controlled crossover trial," *The Lancet Diabetes & Endocrinology,* vol. 3, pp. 17-26, 1// 2015.

[64] P. G. Jacobs, *et al.*, "Randomized trial of a dual-hormone artificial pancreas with dosing adjustment during exercise compared with no adjustment and sensor-augmented pump therapy," *Diabetes Obes Metab,* vol. 18, pp. 1110-1119, Nov 2016.

[65] P. G. Jacobs, *et al.*, "Incorporating an Exercise Detection, Grading, and Hormone Dosing Algorithm Into the Artificial Pancreas Using Accelerometry and Heart Rate " *J Diabetes Sci Technol,* vol. 9, pp. 1175-1184, 2015.

[66] N. Resalat, *et al.*, "Design of a dual-hormone model predictive control for artificial pancreas with exercise model," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2016, pp. 2270-2273, Aug 2016.

[67] C. D. Man, *et al.*, "The UVA/PADOVA Type 1 Diabetes Simulator: New Features," *J Diabetes Sci Technol,* vol. 8, pp. 26-34, Jan 1 2014.

[68] N. Resalat, *et al.*, "A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model," *PLoS One,* vol. 14, pp. 1-17, 2019.

[69] N. Resalat, *et al.*, "Adaptive Control of an Artificial Pancreas using Model Identification, Adaptive Postprandial Insulin Delivery and Exercise," *Journal of Diabetes Science and Technology,* vol. 13, pp. 1044-1053, 2019.

[70] N. Resalat, *et al.*, "Adaptive tuning of basal and bolus insulin to reduce postprandial hypoglycemia in a hybrid artificial pancreas," *Journal of Process Control,* vol. 80, pp. 247-254, 2019/08/01/ 2019.

[71] N. Tyler, *et al.*, "Quantifying the impact of physical activity on future glucose trends using machine learning," *iScience, from Cell Press,* vol. 25, pp. 1-19, 2022.

[72] A. Haidar, *et al.*, "Stochastic Virtual Population of Subjects With Type 1 Diabetes for the Assessment of Closed-Loop Glucose Controllers," *IEEE Trans Biomed Eng,* vol. 60, pp. 3524-33, Dec 2013.

[73] M. Rashid, *et al.*, "Simulation software for assessment of nonlinear and adaptive multivariable control algorithms: Glucose–insulin dynamics in Type 1 diabetes," *Computers & Chemical Engineering,* vol. 130, p. 106565, 2019/11/02/ 2019.

[74] E. Estremera, *et al.*, "A simulator with realistic and challenging scenarios for virtual T1D patients undergoing CSII and MDI therapy," *Journal of Biomedical Informatics,* vol. 132, p. 104141, 2022/08/01/ 2022.

[75] M. E. Wilinska, *et al.*, "Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes," *J Diabetes Sci Technol,* vol. 4, pp. 132-44, Jan 2010.

[76] B. J. Heil, *et al.*, "Reproducibility standards for machine learning in the life sciences," *Nature Methods,* vol. 18, pp. 1132-1135, 2021/10/01 2021.

[77] J. Pineau, *et al.*, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," *ArXiv,* vol. abs/2003.12206, 2021.

[78] D. Rodbard, "Glucose Variability: A Review of Clinical Applications and Research Developments," *Diabetes Technol Ther,* vol. 20, pp. S25-s215, Jun 2018.

[79] P. I. Beato-Víbora, *et al.*, "Real-world outcomes with different technology modalities in type 1 diabetes," *Nutrition, Metabolism and Cardiovascular Diseases,* vol. 31, pp. 1845-1850, 2021/06/07/ 2021.

[80] T. Gebru, *et al.*, "Datasheets for Datasets," *Communications of the ACM,* vol. 64, pp. 86-92, 2021.

[81] Y. Deng, *et al.*, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *npj Digital Medicine,* vol. 4, p. 109, 2021/07/14 2021.

[82] K. Weiss, *et al.*, "A survey of transfer learning," *Journal of Big Data,* vol. 3, p. 9, 2016/05/28 2016.

[83] R. Vilalta and Y. Drissi, "A Perspective View and Survey of Meta-Learning," *Artificial Intelligence Review,* vol. 18, pp. 77-95, 2002/06/01 2002.

[84] V. Naumova, *et al.*, "A meta-learning approach to the regularized learning—Case study: Blood glucose prediction," *Neural Networks,* vol. 33, pp. 181-193, 2012/09/01/ 2012.

[85] T. Zhu, *et al.*, "Dilated Recurrent Neural Networks for Glucose Forecasting in Type 1 Diabetes," *Journal of Healthcare Informatics Research,* vol. 4, pp. 308-324, 2020/09/01 2020.

[86] T. Zhu, *et al.*, "Personalized Blood Glucose Prediction for Type 1 Diabetes Using Evidential Deep Learning and Meta-Learning," *IEEE Trans Biomed Eng,* vol. 70, pp. 193-204, Jan 2023.

[87] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications,* vol. 36, pp. 5718-5727, 2009/04/01/ 2009.

[88] B. W. Bequette, "Fault Detection and Safety in Closed-Loop Artificial Pancreas Systems," *Journal of Diabetes Science and Technology,* vol. 8, pp. 1204-1214, 2014.

[89] D. P. Howsmon, *et al.*, "Real-Time Detection of Infusion Site Failures in a Closed-Loop Artificial Pancreas," *Journal of Diabetes Science and Technology,* vol. 12, pp. 599-607, 2018/05/01 2018.

[90] L. Meneghetti, *et al.*, "Machine Learning-Based Anomaly Detection Algorithms to Alert Patients Using Sensor Augmented Pump of Infusion Site Failures," *J Diabetes Sci Technol,* vol. 16, pp. 641-648, May 2022.

[91] H. Zisser, *et al.*, "Bolus calculator: a review of four "smart" insulin pumps," *Diabetes Technol Ther,* vol. 10, pp. 441-4, Dec 2008.

[92] W. L. Clarke and B. Kovatchev, "Continuous Glucose Sensors: Continuing Questions about Clinical Accuracy," *J Diabetes Sci Technol,* vol. 1, pp. 669-75, Sep 2007.

[93] N. V. Chawla, *et al.*, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Int. Res.,* vol. 16, pp. 321–357, 2002.

[94] O. Mujahid, *et al.*, "Conditional Synthesis of Blood Glucose Profiles for T1D Patients Using Deep Generative Models," *Mathematics,* vol. 10, p. 3741, 2022.

[95] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*: Springer-Verlag, 2006.

[96] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends® in Machine Learning,* vol. 2, pp. 1-127, 2009.

[97] V. Vapnik, *The Nature of Statistical Learning Theory* vol. 2. New York, NY: Springer, 2000.

[98] M. Sevil, *et al.*, "Detection and Characterization of Physical Activity and Psychological Stress from Wristband Data," *Signals,* vol. 1, pp. 188-208, 2020.

[99] M. Sevil, *et al.*, "Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management," *IEEE Trans Biomed Eng,* vol. 68, pp. 2251-2260, Jul 2021.

[100] M. Sevil, *et al.*, "Determining Physical Activity Characteristics from Wristband Data for Use in Automated Insulin Delivery Systems," *IEEE Sens J,* vol. 20, pp. 12859-12870, Nov 2020.

[101] G. Sparacino, *et al.*, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *IEEE Trans Biomed Eng,* vol. 54, pp. 931-7, May 2007.

[102] K. Turksoy, *et al.*, "Hypoglycemia Detection and Carbohydrate Suggestion in an Artificial Pancreas," *J Diabetes Sci Technol,* vol. 10, pp. 1236-1244, Nov 2016.

[103] B. Kovatchev and W. Clarke, "Peculiarities of the continuous glucose monitoring data stream and their impact on developing closed-loop control technology," *J Diabetes Sci Technol,* vol. 2, pp. 158-63, Jan 2008.

[104] C. Pérez-Gandía, *et al.*, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes Technol Ther,* vol. 12, pp. 81-8, Jan 2010.

[105] T. Battelino, *et al.*, "Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range," *Diabetes Care,* p. dci190028, 2019.

[106] E. Montaser, *et al.*, "Essential Continuous Glucose Monitoring Metrics: The Principal Dimensions of Glycemic Control in Diabetes," *Diabetes Technology & Therapeutics,* 2022.

[107] M. R. Askari, *et al.*, "Detection and Classification of Unannounced Physical Activities and Acute Psychological Stress Events for Interventions in Diabetes Treatment," *Algorithms,* vol. 15, p. 352, 2022.

[108] R. Hovorka, *et al.*, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiol Meas,* vol. 25, pp. 905-20, Aug 2004.

[109] S. D. Patek, *et al.*, "Empirical Representation of Blood Glucose Variability in a Compartmental Model," in *Introduction to modeling in physiology and medicine*. vol. 2, E. Carson and C. Cobelli, Eds., ed: Elsevier, 2019, pp. 133-156.

[110] C. Dalla Man, *et al.*, "Meal simulation model of the glucose-insulin system," *IEEE Trans Biomed Eng,* vol. 54, pp. 1740-9, Oct 2007.

[111] K. L. Swan, *et al.*, "Effect of Puberty on the Pharmacodynamic and Pharmacokinetic Properties of Insulin Pump Therapy in Youth With Type 1 Diabetes," *Diabetes Care,* vol. 31, pp. 44-46, 2008.

[112] I. Hajizadeh, *et al.*, "Plasma-insulin-cognizant adaptive model predictive control for artificial pancreas systems," *Journal of Process Control,* vol. 77, pp. 97-113, 2019/05/01/ 2019.

[113] I. Hajizadeh, *et al.*, "Adaptive and Personalized Plasma Insulin Concentration Estimation for Artificial Pancreas Systems," *J Diabetes Sci Technol,* vol. 12, pp. 639-649, May 2018.

[114] T.-T. P. Nguyen, *et al.*, "Separating Insulin-Mediated and Non-Insulin-Mediated Glucose Uptake during Aerobic Exercise in People with Type 1 Diabetes," *American Journal of Physiology-Endocrinology and Metabolism,* 2020.

[115] M. C. Riddell, *et al.*, "Exercise management in type 1 diabetes: a consensus statement," *Lancet Diabetes Endocrinol,* vol. 5, pp. 377-390, May 2017.

[116] P. G. Jacobs, *et al.*, "Integrating metabolic expenditure information from wearable fitness sensors into an AI-augmented automated insulin delivery system: a randomized clinical trial," *Lancet Digit Health,* p. in press, 2023.

[117] R. K. Reddy, *et al.*, "Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study," *JMIR Mhealth Uhealth,* vol. 6, p. e10338, 2018.

[118] B. Sañudo, *et al.*, "Pilot Study Assessing the Influence of Skin Type on the Heart Rate Measurements Obtained by Photoplethysmography with the Apple Watch," *J Med Syst,* vol. 43, p. 195, May 22 2019.

[119] C. Manohar, *et al.*, "Comparison of physical activity sensors and heart rate monitoring for real-time activity detection in type 1 diabetes and control subjects," *Diabetes Technol Ther,* vol. 15, pp. 751-7, Sep 2013.

[120] K. Turksoy, *et al.*, "Classification of Physical Activity: Information to Artificial Pancreas Control Systems in Real Time," *J Diabetes Sci Technol,* vol. 9, pp. 1200-7, Oct 6 2015.

[121] I. S. Dasanayake, *et al.*, "Early Detection of Physical Activity for People With Type 1 Diabetes Mellitus," *J Diabetes Sci Technol,* Jun 30 2015.

[122] J. Garcia-Tirado, *et al.*, "Anticipation of Historical Exercise Patterns by a Novel Artificial Pancreas System Reduces Hypoglycemia During and After Moderate-Intensity Physical Activity in People with Type 1 Diabetes," *Diabetes Technol Ther,* vol. 23, pp. 277-285, Apr 2021.

[123] M. Cescon, *et al.*, "Activity detection and classification from wristband accelerometer data collected on people with type 1 diabetes in free-living conditions," *Comput Biol Med,* vol. 135, p. 104633, Aug 2021.

[124] M. Hernández-Ordoñez and D. U. Campos-Delgado, "An extension to the compartmental model of type 1 diabetic patients to reproduce exercise periods with glycogen depletion and replenishment," *J Biomech,* vol. 41, pp. 744-52, 2008.

[125] N. Hobbs, *et al.*, "A physical activity-intensity driven glycemic model for type 1 diabetes," *Computer Methods and Programs in Biomedicine,* vol. 226, p. 107153, 2022/11/01/ 2022.

[126] C. Dalla Man, *et al.*, "Physical Activity into the Meal Glucose—Insulin Model of Type 1 Diabetes: In Silico Studies," *Journal of Diabetes Science and Technology,* vol. 3, pp. 56-67, 2009/01/01 2009.

[127] M. D. Breton, "Physical activity-the major unaccounted impediment to closed loop control," *J Diabetes Sci Technol,* vol. 2, pp. 169-74, Jan 2008.

[128] B. Ozaslan, *et al.*, "Automatically accounting for physical activity in insulin dosing for type 1 diabetes," *Comput Methods Programs Biomed,* vol. 197, p. 105757, Dec 2020.

[129] M. R. Askari, *et al.*, "Artifact Removal from Data Generated by Nonlinear Systems: Heart Rate Estimation from Blood Volume Pulse Signal," *Industrial & Engineering Chemistry Research,* vol. 59, pp. 2318-2327, 2020/02/12 2020.

[130] H. G. Kim, *et al.*, "Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature," *Psychiatry Investig,* vol. 15, pp. 235-245, Mar 2018.

[131] E. I. Georga, *et al.*, "A predictive model of subcutaneous glucose concentration in type 1 diabetes based on Random Forests," *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2012, pp. 2889-92, 2012.

[132] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion,* vol. 81, pp. 84-90, 2022/05/01/ 2022.

[133] E. I. Georga, *et al.*, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE J Biomed Health Inform,* vol. 17, pp. 71-81, Jan 2013.

[134] K. Li, *et al.*, "Convolutional Recurrent Neural Networks for Glucose Prediction," *IEEE Journal of Biomedical and Health Informatics,* vol. 24, pp. 603-613, 2020.

[135] S. M. Pappada, *et al.*, "Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes," *Diabetes Technol Ther,* vol. 13, pp. 135-41, Feb 2011.

[136] C. Zecchin, *et al.*, "Jump neural network for real-time prediction of glucose concentration," *Methods Mol Biol,* vol. 1260, pp. 245-59, 2015.

[137] C. Zecchin, *et al.*, "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration," *IEEE Trans Biomed Eng,* vol. 59, pp. 1550-60, Jun 2012.

[138] T. Zhu, *et al.*, "Enhancing self-management in type 1 diabetes with wearables and deep learning," *npj Digital Medicine,* vol. 5, p. 78, 2022/06/27 2022.

[139] T. Kushner, *et al.*, "Multi-Hour Blood Glucose Prediction in Type 1 Diabetes: A Patient-Specific Approach Using Shallow Neural Network Models," *Diabetes Technol Ther,* vol. 22, pp. 883-891, Dec 2020.

[140] M. H. Jensen, *et al.*, "Prediction of Nocturnal Hypoglycemia From Continuous Glucose Monitoring Data in People With Type 1 Diabetes: A Proof-of-Concept Study," *J Diabetes Sci Technol,* vol. 14, pp. 250-256, Mar 2020.

[141] S. Del Favero, *et al.*, "A glucose-specific metric to assess predictors and identify models," *IEEE Trans Biomed Eng,* vol. 59, pp. 1281-90, May 2012.

[142] S. Faccioli, *et al.*, "Combined Use of Glucose-Specific Model Identification and Alarm Strategy Based on Prediction-Funnel to Improve Online Forecasting of Hypoglycemic Events," *Journal of Diabetes Science and Technology,* vol. 0, p. 19322968221093665.

[143] F. Cameron, *et al.*, "A Closed-Loop Artificial Pancreas Based on Risk Management," *Journal of Diabetes Science and Technology,* vol. 5, pp. 368-379, 2011/03/01 2011.

[144] C. Mosquera-Lopez, *et al.*, "Automated Meal Detection and Meal Size Estimation Using Machine Learning: Towards Artificial-intelligence-enabled Fully Closed-loop Insulin Delivery Systems," *Nature NPJ Digital,* vol. In press, pp. 1-28, 2023.

[145] Z. Mahmoudi, *et al.*, "An automated meal detector and bolus calculator in combination with closed-loop blood glucose control," *IFAC-PapersOnLine,* vol. 51, pp. 168-173, 2018/01/01/ 2018.

[146] S. Samadi, *et al.*, "Automatic Detection and Estimation of Unannounced Meals for Multivariable Artificial Pancreas System," *Diabetes Technol Ther,* vol. 20, pp. 235-246, Mar 2018.

[147] M. Athanasiou, *et al.*, "An LSTM-based Approach Towards Automated Meal Detection from Continuous Glucose Monitoring in Type 1 Diabetes Mellitus," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, 2021, pp. 1-5.

[148] J. N. Mandrekar, "Receiver Operating Characteristic Curve in Diagnostic Test Assessment," *Journal of Thoracic Oncology,* vol. 5, pp. 1315-1316, 2010/09/01/ 2010.

[149] H. Meng, *et al.*, "Effect of macronutrients and fiber on postprandial glycemic responses and meal glycemic index and glycemic load value determinations," *Am J Clin Nutr,* vol. 105, pp. 842-853, Apr 2017.

[150] J. P. Corbett, *et al.*, "Using an Online Disturbance Rejection and Anticipation System to Reduce Hyperglycemia in a Fully Closed-Loop Artificial Pancreas System," *Journal of Diabetes Science and Technology,* vol. 16, pp. 52-60, 2022.

[151] H. M. Romero-Ugalde, *et al.*, "ARX model for interstitial glucose prediction during and after physical activities," *Control Engineering Practice,* vol. 90, pp. 321-330, 2019/09/01/ 2019.

[152] M. Tejedor, *et al.*, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artificial Intelligence in Medicine,* vol. 104, p. 101836, 2020/04/01/ 2020.

[153] I. Fox, *et al.*, "Deep Reinforcement Learning for Closed-Loop Blood Glucose Control," presented at the Proceedings of the 5th Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research, 2020.

[154] P. Herrero, *et al.*, "Identifying Continuous Glucose Monitoring Data Using Machine Learning," *Diabetes Technol Ther,* vol. 24, pp. 403-408, Jun 2022.

[155] G. Noaro, *et al.*, "An Ensemble Learning Algorithm Based on Dynamic Voting for Targeting the Optimal Insulin Dosage in Type 1 Diabetes Management," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1828-1831.

[156] G. Noaro, *et al.*, "Machine-Learning Based Model to Improve Insulin Bolus Calculation in Type 1 Diabetes Therapy," *IEEE Transactions on Biomedical Engineering,* vol. 68, pp. 247-255, 2021.

[157] P. Herrero, *et al.*, "Automatic Adaptation of Basal Insulin Using Sensor-Augmented Pump Therapy," *J Diabetes Sci Technol,* vol. 12, pp. 282-294, Mar 2018.

[158] C. C. Palerm, *et al.*, "A Run-to-Run Control Strategy to Adjust Basal Insulin Infusion Rates in Type 1 Diabetes," *J Process Control,* vol. 18, pp. 258-265, 2008.

[159] C. Toffanin, *et al.*, "Automatic adaptation of basal therapy for Type 1 diabetic patients: a Run-to-Run approach," *IFAC Proceedings Volumes,* vol. 47, pp. 2070-2075, 2014/01/01/ 2014.

[160] H. Zisser, *et al.*, "Clinical update on optimal prandial insulin dosing using a refined run-to-run control algorithm," *J Diabetes Sci Technol,* vol. 3, pp. 487-91, May 1 2009.

[161] J. R. Castle, *et al.*, "Assessment of a Decision Support System for Adults with Type 1 Diabetes on Multiple Daily Insulin Injections," *Diabetes Technol Ther,* Aug 3 2022.

[162] M. Vettoretti, *et al.*, "Patient decision-making of CGM sensor driven insulin therapies in type 1 diabetes: In silico assessment," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 2363-2366.

[163] R. Nimri, *et al.*, "Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes," *Nat Med,* vol. 26, pp. 1380-1384, Sep 2020.

[164] R. Nimri, *et al.*, "Comparison of Insulin Dose Adjustments Made by Artificial Intelligence-Based Decision Support Systems and by Physicians in People with Type 1 Diabetes Using Multiple Daily Injections Therapy," *Diabetes Technol Ther,* vol. 24, pp. 564-572, Aug 2022.

[165] A. Bisio, *et al.*, "Impact of a Novel Diabetes Support System on a Cohort of Individuals With Type 1 Diabetes Treated With Multiple Daily Injections: A Multicenter Randomized Study," *Diabetes Care,* vol. 45, pp. 186-193, 2021.

[166] P. Herrero, *et al.*, "Robust Determination of the Optimal Continuous Glucose Monitoring Length of Intervention to Evaluate Long-Term Glycemic Control," *Diabetes Technol Ther,* vol. 23, pp. 314-319, Apr 2021.

[167] T. Battelino, *et al.*, "Continuous glucose monitoring and metrics for clinical trials: an international consensus statement," *Lancet Diabetes Endocrinol,* vol. 11, pp. 42-57, Jan 2023.

[168] K. Zarkogianni, *et al.*, "A Review of Emerging Technologies for the Management of Diabetes Mellitus," *IEEE Trans Biomed Eng,* vol. 62, pp. 2735-49, Dec 2015.

[169] B. W. Bequette, "Algorithms for a closed-loop artificial pancreas: the case for model predictive control," *J Diabetes Sci Technol,* vol. 7, pp. 1632-43, Nov 1 2013.

[170] N. Resalat, *et al.*, "Evaluation of model complexity in model predictive control within an exercise-enabled artificial pancreas," *IFAC-PapersOnLine,* vol. 50, pp. 7756-7761, 2017/07/01/ 2017.

[171] K. Zarkogianni, *et al.*, "An Insulin Infusion Advisory System Based on Autotuning Nonlinear Model-Predictive Control," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 2467-2477, 2011.

[172] M. Phillip*, et al.*, "Consensus Recommendations for the Use of Automated Insulin Delivery (AID) Technologies in Clinical Practice," *Endocr Rev,* Sep 6 2022.

[173] W. L. Clarke*, et al.*, "Evaluating clinical accuracy of systems for self-monitoring of blood glucose," *Diabetes Care,* vol. 10, pp. 622-8, Sep-Oct 1987.

[174] J. L. Parkes*, et al.*, "A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose," *Diabetes Care,* vol. 23, pp. 1143-1148, 2000.

[175] W. L. Clarke*, et al.*, "Evaluating clinical accuracy of continuous glucose monitoring systems: Continuous Glucose-Error Grid Analysis (CG-EGA)," *Curr Diabetes Rev,* vol. 4, pp. 193-9, Aug 2008.

[176] D. C. Klonoff*, et al.*, "The surveillance error grid," *J Diabetes Sci Technol,* vol. 8, pp. 658-72, Jul 2014.

[177] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics,* vol. 21, p. 6, 2020/01/02 2020.

[178] F. Brouns*, et al.*, "Glycaemic index methodology," *Nutr Res Rev,* vol. 18, pp. 145-71, Jun 2005.

[179] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.

[180] M. T. Ribeiro*, et al.*, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.

[181] A. Shrikumar*, et al.*, "Learning important features through propagating activation differences," presented at the Proceedings of the 34th International Conference on Machine Learning - Volume 70, Sydney, NSW, Australia, 2017.

[182] A. Karimi*, et al.*, "Model-Agnostic Counterfactual Explanations for Consequential Decisions," presented at the Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), Virtual, 2020.

[183] S. Liu*, et al.*, "Generative Counterfactual Introspection for Explainable Deep Learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019, pp. 1-5.

[184] J. Daniels*, et al.*, "A Multitask Learning Approach to Personalized Blood Glucose Prediction," *IEEE Journal of Biomedical and Health Informatics,* vol. 26, pp. 436-445, 2022.

[185] C. Toffanin*, et al.*, "Adaptive and Individualized Artificial Pancreas for Precision Management of Type 1 Diabetes," in *Precision Medicine in Diabetes: A Multidisciplinary Approach to an Emerging Paradigm*, R. Basu, Ed., ed Cham: Springer International Publishing, 2022, pp. 305-313.

[186] S. I. Taylor*, et al.*, "SGLT2 inhibitors as adjunctive therapy for type 1 diabetes: balancing benefits and risks," *Lancet Diabetes Endocrinol,* vol. 7, pp. 949-958, Dec 2019.

[187] T. Zhu*, et al.*, "Personalized Dual-Hormone Control for Type 1 Diabetes Using Deep Reinforcement Learning," in *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, Eds., ed Cham: Springer International Publishing, 2021, pp. 45-53.

[188] K. Cao*, et al.*, "An Overview on Edge Computing Research," *IEEE Access,* vol. 8, pp. 85714-85728, 2020.

[189] T. Zhu*, et al.*, "IoMT-Enabled Real-Time Blood Glucose Prediction With Deep Learning and Edge Computing," *IEEE Internet of Things Journal,* vol. 10, pp. 3706-3719, 2023.