IMPROVING DECISION SUPPORT IN

TYPE 1 DIABETES MANAGEMENT

By

Alejandro Z. Espinoza

A THESIS

Presented to the Department of Biomedical Engineering
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

February 2024

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF EQUATIONS

Acknowledgements

I am deeply grateful for all my mentors and peers at the Artificial Intelligence for Medical Systems (AIMS) lab at OHSU. This work would not have been possible without your guidance. Specifically, I'd like to thank:

- Dr. Peter Jacobs, for his support in the AIMS lab throughout my time in biomedical engineering, for always providing valuable insight and suggestions, and the opportunity to collaborate on projects such as this.
- Dr. Clara Mosquera-Lopez, for acting as chair of my thesis committee, for all of the feedback provided throughout this process, and for sharing extremely useful pre-processing scripts to generate several features used in this work.
- Dr. Leah Wilson, for providing invaluable feedback from the clinical point of view which helped shape this document to its current state and participating in my thesis committee.
- Dr. Jessica Castle, for providing clinical feedback, collaboration, and review of work on the DailyDose project.
- Dr. Taisa Kushner, for providing insight and initial help with understanding the DailyDose data in addition to collaborating on preprocessing codes.
- Dr. Robert Dodier in the AIMS lab, for providing technical insight on DailyDose project and patience while reasoning through complex problems in the lab.

Finally, thank you to all my friends and family, including my dog Nova. Without your support I wouldn't be who I am or where I am today.

Abstract

Type 1 diabetes (T1D) is a chronic disease that requires those who live with it to regularly administer exogenous insulin to maintain blood glucose (BG) within a safe range. Insufficient insulin dosing can cause high glucose, leading to complications such as diabetic retinopathy, neuropathy, and cardiomyopathy. Too much insulin can cause low glucose, and when untreated, lead to coma or death. T1D-related complications can be reduced with proper diet, exercise, and adherence to an insulin dosing regimen. Decision support systems (DSS) can also help optimize glucose management by utilizing artificial intelligence (AI) and physician expertise to suggest adjustments to therapy parameters and behavior. However, a lack of explainability in a DSS's suggested changes may deter participants from understanding and following AI-generated recommendations. In addition, manual calculation of meal insulin is still required by the participant. Incorrect dosing of meal insulin is common due to human error in estimating carbohydrate amounts in foods which can lead to poor postprandial glucose outcomes. Fats and proteins have also been observed to affect postprandial glucose response (PPGR) and these are typically not considered when estimating mealtime insulin.

This work begins with background on diabetes management and a discussion of the complexity of nutritional influences on PPGR, extending beyond carbohydrates to include fats and proteins. In the second chapter, an 8-week randomized controlled study of a real-world DSS is discussed. Mixed effects models revealed that time-in range was improved by an average of 6.3% (P=0.001, CI 2.5%-10.1%) when participants accepted and followed system recommendations. In the third chapter, machine learning (ML) algorithms are trained to predict three components of PPGR: net area under the curve (netAUC), minimum BG (minBG), and maximum BG (maxBG). Predictions of balanced random forest (RF) algorithms are combined with those of standard RF via regression models to reduce bias. While systemic bias remains, final predictions are highly correlated with observations for netAUC, minBG, and maxBG (R=0.62, R=0.61, R=0.7). Feature importance reveals that past BG measurements, BG trend, and insulin on board primarily influence the PPGR and macronutrient contents were less relevant. Finally, in the fourth chapter, future DSS integrations with PPGR models are discussed. Future PPGR modeling may help DSS users calculate the appropriate amount of meal insulin, provide explainability to bolus calculations, and help guide individuals towards dietary choices that optimize PPGR.

Chapter 1 - Introduction

Type 1 diabetes (T1D) is a chronic disease that affects approximately 1.6 million people in the United States alone [1]. People living with T1D are unable to produce endogenous insulin due to destruction of pancreatic beta cells caused by an autoimmune disorder, and therefore must deliver exogenous insulin to enable regulation of blood glucose [2]. Many people living with T1D are treated with multiple daily injection (MDI) therapy, a regimen in which insulin is delivered through needle syringes before meals or when glucose is too high [3]. Technological advancements have paved the way for automated insulin delivery (AID) systems, which utilize data from continuous glucose monitors (CGMs) to calculate and deliver insulin automatically [4]. Studies have indicated that people using AID systems have improved glycemic outcomes compared to people using MDI therapy. Many people choose not to use AIDs however, for a variety of reasons including cost and socioeconomic factors, physical limitations of wearing two devices on the body, and willingness or concerns by medical providers [5].

The safe range for blood glucose is 70-180 mg/dL [6]. Blood glucose above this range is referred to 'high glucose', and below is referred to as 'low glucose'. The percentage of time spent within the safe range is referred to as % time in range (TIR). CGMs are commonly used to aid in assessing blood glucose levels, enabling users to monitor glucose every 1-15 minutes and inform insulin dosing based on treatment modality [6]. Long-acting insulin is delivered by MDI users once or twice daily, and releases from the subcutaneous space into the blood stream to help keep blood glucose within range. In addition to long-acting insulin, fast-acting insulin must be delivered when food is consumed to compensate for the rise in glucose caused by carbohydrate intake. The amount delivered is typically determined by establishing an insulin-to-carbohydrate ratio (ICR) with support from a physician and adjusting when necessary. The person oftentimes uses a bolus calculator that calculates the meal and correction insulin based on the person's current glucose level, the amount of correction insulin on board (IOB), and the amount of carbohydrates (CHO) in the meal that the person is going to eat [7]. This calculation is detailed in Equation 1. The goal of the bolus calculator is to recommend delivery of enough insulin to avoid excessive high glucose while avoiding low glucose. There are serious risks associated with incorrect insulin doses. Dosing too much or too little insulin can result in dangerously adverse events such as severe low glucose (<54 mg/dL) or extreme high glucose (>250 mg/dL). Chronic exposure to elevated glucose levels can damage tissue and lead to peripheral neuropathy, retinopathy, and cardiovascular complications [8]. Exposure to severe low glucose can damage the heart and if left untreated can lead to seizure, coma, and even death [9].

$$Bolus\ insulin = Meal\ insulin + Correction\ insulin - IOB$$

$$= \frac{CHO}{ICR} + \frac{Current\ BG - Target\ BG}{Correction\ Factor} - IOB$$

Equation 1 - Bolus insulin calculation [7]

Carbohydrates are considered to be the primary macronutrient which cause increases in glucose [10]. A person must therefore properly estimate the amount of carbohydrates in their meal to determine the correct amount of meal insulin to administer. While insulin dosing can be made easier through the use of AID systems, accurate estimation of carbohydrates before meals is still required for people using current commercial AID systems and MDI systems. In addition to glucose rise caused by carbohydrates, meals with high fat content are known to impact delayed gastric emptying, which can reduce the rate of carbohydrate absorption [11]. When the rate of carbohydrate absorption is reduced, blood glucose rises more slowly, which can further complicate insulin dosing and glucose regulation. Another potential interaction was found in a study that suggested high fat in meals may decrease insulin sensitivity, requiring more insulin to be dosed as a result [12]. Meals with moderate to high protein content (28g – 57g) have been shown to elevate postprandial blood glucose in individuals with T1D, which will also increase insulin dosing requirements [13]. Consuming meals with high levels of fat and protein can introduce additional complexity when determining an appropriate amount of mealtime insulin. There is insufficient information available, however, to incorporate macronutrient content into bolus calculators. Fats and proteins impact glucose responses differently across various individuals with T1D and this makes it hard to define a set of guidelines on how to incorporate these macronutrients into bolus calculators [14, 15]. Varying strategies have been suggested to adjust dosing calculations based on fats and proteins, but including macronutrient data is not always beneficial. In a study of pediatrics and adolescents, there was increased risk of low glucose when including macronutrient data [16]. Due to these complexities, bolus calculators currently only utilize carbohydrates when calculating and recommending insulin dosage. Lack of reliable methodology to incorporate contributions of protein and fats to glycemic response remains both a problem for people with T1D and an area of opportunity for improving diabetes management. This work aims to help address this problem in chapter three, where machine learning algorithms including macronutrient content predict the PPGR.

In addition to insufficient methodology to include more complete macronutrient information in bolus calculators, another hurdle to clear in appropriate meal insulin estimation lies in the accuracy of macronutrient estimation by individuals. In a 2016 study, 61 adults were given a carbohydrate counting test based on commonly consumed foods [17]. A total of 82% of participants overestimated carbohydrate content by an average of 40%. In another study by Gillingham, et al. [18], it was shown that while tracking food in logging apps and verified using nutritionists and food photography, only 22% of high protein meals and 17% of high fat meals were estimated accurately. Furthermore, Gillingham et al. found that people consistently overestimated smaller meals and underestimated larger meals. These findings demonstrate the challenges that bolus calculators face when used by people with T1D.

Given the overwhelming risk and complexity associated with living with T1D, much research has been invested in developing models to help simplify decision making for individuals through smartphone applications known as decision support systems (DSS). DSSs can provide recommendations through automated methods such as machine learning models and rule-based systems, as discussed in the review by Tyler and Jacobs [19]. DSSs can integrate multiple types of heterogenous data such as CGM, insulin dosed, food logs, and exercise logs [20]. These collected data are used to generate personalized recommendations. Example recommendations may include adjustments to long-acting insulin, adjustment of insulin-to-carbohydrate ratio, or adjusting correction factors at specific times of day [21]. Many DSSs have not been evaluated for clinical efficacy, and there is conflicting evidence for those which have been evaluated. While recently studies have suggested that DSS use may not result in significant improvement in %TIR for MDI users [22], there is also evidence that when people using a DSS follow recommendations provided by the DSS, there are significant improvements in glycemic outcomes [23].

The CGM measurements in the period of time following a meal is referred to as the postprandial glucose response (PPGR). A typical measure of PPGR includes calculating the area under the curve (AUC) of the CGM curve, which can be accomplished by numerical integration with the trapezoidal rule [24]. The AUC is informative because it is an indicator of whether or not the insulin dosed for the meal was sufficient. Too little insulin may result in a larger AUC due to rise in glucose caused by insufficient meal insulin, whereas too much insulin can lead to a negative AUC relative to starting CGM. Ideally, the glucose response should be minimal when insulin is appropriately dosed resulting in a near-zero AUC. The AUC may be calculated in multiple ways. The standard AUC is calculated including the CGM at the start of the meal and calculating the area under the entire CGM including the starting CGM. Including the starting CGM leads to misleading results so it is typically excluded. There are several other more relevant ways of calculating AUC: the incremental AUC (iAUC) and net AUC (netAUC). NetAUC is a quantity representing the total change in blood glucose relative to a starting value by summing both positive and negative CGM values under the curve relative to the stating glucose. This differs from a related measure, iAUC, which only considers positive CGM values that are above the starting glucose value while ignoring glucose values that are below the starting glucose value. In this work, netAUC is calculated by subtracting the median CGM over the past 30 minutes at the start of the meal from each postprandial CGM value recorded. Using the median of CGM values over the past 30 minutes is helpful to reduce the impact of CGM noise [25]. The trapezoidal rule is then utilized to approximate integration of both positive and negative area under the curve. In addition to AUC metrics, it can also be insightful to observe the minimum blood glucose (minBG) and maximum blood glucose (maxBG) of the PPGR, which can be indicators of whether or not the patient experienced an adverse glycemic event. Multiple factors impact the PPGR including the person's sensitivity to insulin, the amount of insulin the person has in their body (insulin-on-board or IOB), the accuracy of the carbohydrate estimation, the timing of the meal insulin delivery relative to the consumption of the meal, and the macronutrient content of the meal that they are consuming [12, 13, 26, 27].

Models that can predict the PPGR may be helpful for use in both DSSs and AID algorithms. For example, a person using MDI therapy typically uses an insulin bolus calculator to determine how much insulin to inject into their body prior to consuming a meal. A model that can accurately predict the person's PPGR could potentially be used to calculate the exact amount of meal bolus insulin that should be administered to optimize the person's TIR and minimize the time in high and low glucose. Likewise, an AID would make use of an accurate model that can predict PPGR by determining how much meal insulin should be delivered by a hybrid AID that requires the person to announce meals to the system which will then deliver meal insulin.

In the Chapter 2, analysis of a recently developed DSS called DailyDose is discussed. This is followed by the introduction of a method of improving personalized DSS recommendations through the integration of machine learning algorithms with the goal of predicting PPGR. Finally, future direction is provided on where this research can go next to further improve development of these systems to help those living with T1D improve PPGR.

Chapter 2 - DailyDose DSS in an 8-week randomized controlled study

DailyDose is a DSS designed to help people living with T1D using MDI therapy improve their glucose control by providing weekly recommendations on modifications to carbohydrate ratios, correction factors, and long-acting insulin dosing [21]. It is built as an iPhone application which provides recommendations for changes to insulin dosing which are customized to each user based on CGM data, carbohydrate intake, and exercise. The recommendation engine used in DailyDose is a k-nearest neighbors (KNN) algorithm, which was trained on virtual patient data to predict up to four changes to therapy parameters and behavioral changes each week based on glycemic features calculated from past user data.

DailyDose was evaluated in an 8-week study. 25 adults living with T1D on MDI therapy were recruited, comparing 2-week baseline TIR prior to using DailyDose to TIR in the final 2 weeks using DailyDose [23]. Participants in the study used Dexcom G6 CGM (Dexcom Inc, San Diego, CA), InPen (Medtronic plc, Minneapolis, MN) for tracking bolus insulin, and Clipsulin (Glooko Inc, Mountain View, CA) for tracking basal insulin. While there were no significant differences found between final 2-week TIR and baseline 2-week TIR, post-hoc analysis of data showed improvement in the weeks following times when participants followed recommendations given by DailyDose compared to weeks when participants did not follow the recommendations. This study and the results are discussed further below and were published in Castle et al. 2022 [23].

Methods

Data in the study were collected from user devices and logged in the DailyDose app. Participant demographics can be found in Table 1. Data collected by DailyDose included insulin doses from InPen and Clipsulin Bluetooth-enabled insulin dosing devices, along with CGM and meal bolus calculators from within the app. Acceptance or rejection of every recommendation for changes in correction factors, carbohydrate ratios, and long-acting insulin for each participant was tracked weekly. A binary adherence variable was created based on whether or not participants followed recommendations provided by the DSS when dosing their meal, correction or long acting insulin. If participants did not follow these recommendations, a subsequent recommendation would then be generated asking them to follow dosing guidance. Based on the presence of this new recommendation, it can be determined if they were taking the advice of the DSS or not. The change in the percent TIR was the primary outcome, calculated as the difference between the last two weeks and the first two baseline weeks of the study. We did a secondary analysis whereby we calculated the difference in TIR between weeks to determine if there was an increased TIR on weeks following acceptance of recommendations provided by the app.

Using the Stata version 16.1 software, a mixed effects regression model was fitted comparing differences in % TIR at the end of the study to baseline %TIR. Post-hoc, an additional mixed effects model was fit to compare differences in %TIR between weeks when participants accepted recommendations compared with weeks when they did not follow recommendations. Weeks were partitioned to compare those when participants had accepted and followed 50% or fewer of that week's recommendations versus those who had accepted and followed 50% or more. This grouping was selected to test whether accepting and following the majority of recommendations week-to-week resulted in improved TIR compared to otherwise not accepting and following recommendations.

Table 1 - Participant demographics, DailyDose assessment

Demographic	
Total participants, N	25
Mean age, years	35.8
Biological sex, N (%)	14 (56) F 11 (44) M

Ethnicity (self-identified), N (%)	White 22 (88) Asian 1 (4) Black/African American 2 (8)
HbA1c, %	8.2
Prior CGM use, N (%)	15 (60)

Results

There were no significant differences found in final TIR when compared to baseline TIR (P=0.25). However, post-hoc analysis revealed that when comparing weeks when at least 50% of recommendations were accepted and followed compared with weeks when they were not followed, % TIR was found to significantly improve by an average of 6.3% (P=0.001, CI 2.5%-10.1%) [23].

The distribution of changes in %TIR week-to-week is shown in Figure 1. The blue boxplot illustrates the weekly change in %TIR for weeks when participants did not accept and follow any recommendations provided by the app. The orange boxplot illustrates the weekly change in %TIR for weeks when participants accepted and followed some recommendations, but not all of them. Finally, the third boxplot in green represents the weekly change in %TIR for weeks when participants accepted and followed all recommendations.

Adherence to bolus calculator recommendations was also analyzed post-hoc. We found that out of 6694 boluses, the integrated bolus calculator was utilized 81.9% of the time. 67.5% of boluses delivered were within 0.5 units of the bolus calculator recommendation, while 14.7% were more than 0.5 units above this value, and 17.8% were more than 0.5 units below the recommendation. By analyzing bolus calculator adherence and TIR using an additional mixed effects model, we found a 16.9% increase in TIR in weeks that participants consistently delivered boluses within 0.5 units of recommended values compared to weeks in which the bolus calculator was not used at all (P=0.012). Importantly, usage of the calculator was not associated with increased risk of low glucose (P=0.159).

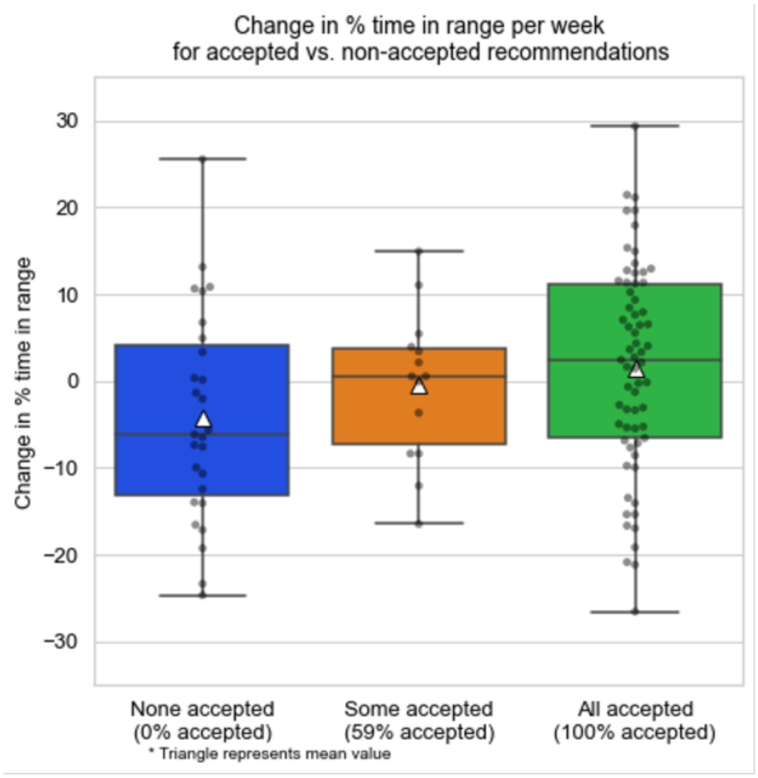

Figure 1 - Change in %TIR versus acceptance of recommendations [28]

Discussion

While overall there were no improvements in %TIR compared to baseline, in post-hoc analysis we found significant improvement in weeks following when participants accepted and followed recommendations. In addition, analysis of bolus calculator adherence showed that following the bolus calculator consistently led to substantial increases in %TIR compared to when the bolus calculator was not followed. Interviews with participants after the study indicated that a possible barrier to recommendation acceptance was simply not understanding why the recommendation was proposed in the first place and why it would be beneficial to follow the recommendation. Explainability in DSS predictions may help participants understand and accept recommendations for DSSs. Other potential barriers to acceptance of recommendations included not having adequate time to review and understand the recommendations. Incorporating this feedback in future design may lead improvements in overall rates of acceptance and adherence to DSS recommendations.

Chapter 3 - Forecasting postprandial glucose response using nutrient information and balanced random forests

Administration of correct prandial insulin boluses remains a persistent challenge for individuals regardless of insulin treatment modality. Carbohydrates are the primary macronutrient which cause blood glucose concentration to rise and are typically the only macronutrient used by bolus calculators. However, it is well documented that meals with high fat content delay gastric emptying by affecting rate of glucose absorption [11]. In addition, meals with high protein content can result in elevated blood glucose [13]. In the absence of reliable bolus calculation methods which include fats and proteins, there may be an opportunity to utilize macronutrient information in combination with CGM and insulin data to forecast components of the PPGR using machine learning models.

Glucose forecasting itself is a widely explored topic. There have been many different approaches which generally fall under data-driven modeling, physiological modeling, and a mix of both [29]. Data-driven modeling includes various statistical and machine learning algorithms, such as deep neural networks, gradient boosted trees, random forest, and least squares regression [30, 31, 32, 33]. These algorithms utilize current and past CGM data to forecast future glucose values over various prediction horizons, from 15 minutes up to several hours in the future. Various additional features may be included, such as delivered insulin, carbohydrates, time of day, exercise events, and other related features to the patient's current state. Rather than learning relationships between input features and future glucose values, physiological modeling approaches utilize systems of differential equations to describe glucose-insulin/glucagon dynamics [34, 35]. While not performing at the same level as state-of-the-art ML in terms of predictive accuracy, an advantage to using physiological models lies in their direct interpretability. Machine learning models can run the risk of learning correlations between insulin and blood glucose which violate known relationships and could potentially put individuals using these forecasting algorithms at risk [36]. When interpretability is difficult or not possible due to the model's complexity or black-box nature, a best practice is to instead use tools which allow explainability in the model's predictions [24]. Explainability allows for insight into which features or parameters drives the output of the model. Some examples of explainability algorithms are SHAP [37], LIME [38], and permutation feature importance, which was originally proposed by Breiman and Cutler in a technical report for random forest [39].

A persistent challenge in glucose forecasting is prediction of low glucose and very high glucose due to the infrequent occurrence of these events in real-world data relative to the time most people spend in target range or in high glucose. This challenge is not unique to glucose forecasting. When defined in the context of any model learning from imbalanced data, such as classification, regression, or clustering, the modeling task can be referred to as imbalanced learning [40]. For most regression tasks, mean square error (MSE) or mean absolute error is computed over the target and predictions values. Each sample in the target variable distribution has equal weighting when evaluating errors in prediction and providing feedback to update learned model parameters. As a result, patterns for more commonly observed values are optimized, which can lead to poorer performance for infrequently observed examples [41]. There have been several proposed strategies to address imbalanced regression problems. One strategy involves partitioning the target distribution into regions of common and uncommon outcomes [42]. Examples which lead to common outcomes may be under-sampled in the preprocessing phase, and those which lead to less common outcomes may be over-sampled. One disadvantage to this approach is the possibility of increased likelihood of overfitting as the additional examples are simply duplicated and may not provide additional meaningful information [43]. The Synthetic Minority Over-sampling Technique for Regression (SMOTER) algorithm has been developed as an alternative to create new synthetic examples based on the original data to address this problem [44]. However, these examples are created through linear interpolation of existing samples and can also present the risk of producing examples without additional meaning, especially in complex high dimensional feature spaces. Another proposed method is to augment existing samples by adding a small amount of Gaussian noise (GN) to create additional synthetic examples [42]. It has also been proposed that SMOTER and GN be combined into an algorithm called SMOGN [45]. The idea behind SMOGN is that it may be beneficial to reduce the risk of creating less meaningful new examples through interpolation by switching strategy when samples are very far apart from each other. GN is applied to the samples in the distant case, allowing for switching between both SMOTER and GN depending on proximity in output space as determined by k-nearest neighbors. Finally, rather than manipulating the training data directly, it is also possible to redefine the loss function in a way which penalizes prediction error for extreme examples more heavily than commonly observed examples [46].

In terms of glucose forecasting for glycemic ranges, both random forest and transformer-based deep learning models have achieved high accuracy in predicting low glucose and high glucose, evaluated on data from participants living with T1D [47]. In a recent review of glucose forecasting algorithms, it was found that random forest, gradient boosted trees, and neural network models yielded highest accuracy on prediction horizons up to 45 minutes with CGM data as input [48]. A study by Mosquera-Lopez, et al. also studied the use of long-short-term-memory (LSTM) neural networks to forecast blood glucose with high accuracy over 30-minute and 60-minute prediction horizons [33]. A significant contribution of this study was the introduction of the glucose variability impact index (GVII), a method which quantifies how glucose variability may impact the accuracy of algorithm predictions. It was found that GVII was highly correlated with RMSE ($R \geq 0.64$, $P < 0.001$). Finally, in another study, it was noted that the random forest algorithm may be better suited to learn complex patterns related to low glucose detection given temporal features [49].

Forecasting PPGR using nutrient information and other factors can be viewed as a special case of glucose forecasting. In work by Zeevi, et al. [25], various interactions between gut microbiome, physiological factors, food nutrition, and PPGR are studied on a large cohort of 800 participants living without diabetes. In this study, a gradient-boosted regression model was trained on numerous meal features, such as carbohydrates, fats, protein, alcohol and caffeine content, sodium, total calories, fiber, sugar, and more. Physiologic features for each participant were also included, along with CGM data, exercise, defecation routine, and information from clinical tests. The trained model was found to predict iAUC highly correlated with actual iAUC from the PPGRs ($R=0.68$). A major finding from this work was that while the iAUC of the PPGR is reproducible within the same individual consuming the same meal consistently, it varies significantly between individuals even when they consume the same foods. This suggests the need to train personalized models for forecasting the PPGR. Analysis of feature importance via partial dependence plots revealed several nutritional interactions with iAUC. Carbohydrates were found to be positively correlated with iAUC. Higher fat content in meals relative to amount of carbohydrates was observed to lower iAUC, potentially due to delayed gastric emptying. It was observed that fiber content in the meal raise iAUC for that meal, but more fiber consumed over the past 24 hours was associated with overall lower iAUC. In addition, differences in gut microbiome were found to contribute to whether foods resulted in a positive or negative outcome. Specifically, some bacterium associated believed to ferment carbohydrates and fiber were observed to lower iAUC, and some bacterium thought to associate with higher risk of obesity were associated with higher iAUC. Following Zeevi, et al.'s work, another study was conducted in 2019 on a cohort of 327 midwestern Americans to investigate if these findings would generalize beyond the Israeli population [50]. A gradient-boosted trees regression model was trained with the same input features. It was found that once again, the predicted iAUC was highly correlated (R=0.62) with actual iAUC. Root mean square error (RMSE) was also computed to be 14.82 mg/dL*h. While this work is promising for those living without diabetes, it was ultimately not tested or validated on a cohort of individuals living with T1D. Other studies have also worked to predict PPGR in participants with gestational diabetes but without microbiome data [51, 52]. In addition to iAUC, rise in blood glucose and maximum blood glucose were predicted with high correlation as observed previously.

Recent work by Annuzzi, et al. utilized artificial neural networks to predict postprandial blood glucose values in people with T1D with the goal of investigating nutritional influence [53]. To explore relationships between input variables and PPGR, an AI explainability algorithm called SHAP was leveraged [54]. In this study, 25 individuals were recruited for one week. Detailed food diaries were recorded for every meal, from which macronutrient information were extracted in addition to fiber, overall calories, and glycemic index. Glycemic features related to prior half hour CGM readings were created, including statistical characteristics of the measurements such as mean, median, standard deviation, kurtosis, skew, peak-to-peak, minimum, and maximum. Meal-time insulin boluses were included as features, along with smaller boluses recently delivered by the pump. With these features as inputs, the neural network was trained to forecast postprandial blood glucose values with up to a two-hour prediction horizon. SHAP was then applied and used to analyze feature importance. They found the features with the highest importance to be the prior 30 minutes of CGM readings, followed by meal carbohydrates, glycemic load, fats, and recently delivered bolus.

Better understanding and anticipation of the effects of a meal on an individual's PPGR prior to consumption could greatly benefit users of DSSs and AID devices for people living with T1D. In this chapter, forecasting various aspects of the PPGR is explored using meal data recorded from a cohort of 364 participants living with T1D from the type 1 diabetes in exercise (T1Dexi) initiative [55]. Machine learning models were trained to predict netAUC, minimum blood glucose, and maximum blood glucose over a three-hour prediction horizon. These models were trained on input features, including various CGM statistics, select demographics, insulin dosing, and physician reported meal macronutrients. Random forest was chosen as the primary model algorithm due to its high performance on small tabular data sets along with increased interpretability [56]. In order to try to overcome the challenge of predicting less common observations of low glucose and high glucose, secondary balanced random forest regression models were trained by modifying the bootstrap sample for each tree to oversample less common glycemic excursions while under-sampling common events. In addition, to overcome the difficulty extrapolating beyond frequently observed training examples, regression models were fit to learn the best weighting of predictions from each model and combine them in an effort to reduce residual error. The ultimate goal of this work is to apply the predictions of these models to help those living with T1D to better anticipate PPGR in automated meal detection for use in AIDs [57] and in improved smart bolus calculators for use in decision support settings [21]. More accurate and personalized prandial insulin bolus calculations may improve glucose outcomes.

Methods

Data Sourcing and Demographics

Data for modeling were sourced from the T1DEXI study [55], which aimed to analyze and understand glycemic responses of individuals living with T1D during exercise. Participants in the study participated in a 4-week data collection period during which CGM, insulin, exercise (heart rate and accelerometry) were recorded using a Verily fitness watch and food data were self-reported and validated using food photography (remote food photography method or RFPM) [58] and nutritionist scoring. The results of this study indicated a significant interaction between glycemic response and exercise. Specifically, it was found that those who participated in aerobic exercise experienced a large drop in glucose (-18 ± 39 mg/dL), followed by interval exercise (-14 ± 32 mg/dL), and resistance exercise (-9 ± 36 mg/dL) [55]. Exercise improved TIR on active days compared with sedentary days, but active days were also found to be associated with slightly higher rates of low glucose. While the primary focus of this study was to capture glucose changes during different types of exercise, there are many examples of PPGRs with clinician provided nutrient analysis of meals in this data set. Analysis of these data offers rare insight into PPGRs of those living with T1D under real-world settings. In this work, we utilized the time-series data collected during the study from CGM, insulin pumps, fitness watches, and nutrient information. Specifically, we examined a subset of PPGRs during which subjects consumed meals but did not exercise in the following three hours after the start of the meal due to the possible interaction of exercise on the PPGR. Overall demographics for the T1DEXI study are provided in Table 2.

Table 2 - T1DEXI Study Demographics [55]

Demographic	
Total participants, N	503
Age, years	36.7 (14)
Biological sex, N (%)	367 (73) F 136 (27) M
Weight, lbs	161.5 +/- 30.6
Ethnicity (self-identified), N (%)	White 459 (91.4) Do not wish to answer/Don't know 13(2.6) Asian 10 (2) Black/African American 10 (2) More than one race 8 (1.6) American Indian/Alaskan Native 2 (0.4)
HbA1c, %	6.7 +/- 0.77
Duration of diabetes, years	17.9 +/- 13
CGM use, N (%)	Dexcom 423 (88.5) Medtronic 42 (8.8) Abbott 13 (2.7)
Insulin modality, N (%)	Closed loop 225 (44.7) SAP 189 (37.6) MDI 89 (17.7)

Out of the 503 participants with available data, 364 had sufficient data for meal analysis and modeling. A PPGR was considered eligible for analysis if there was sufficient data in the PPGR as discussed further in the following pre-processing section. Multiple daily injection (MDI) users were excluded from this analysis due to issues with dosing accuracy. After applying the exclusion criteria, 4493 meals remained for ML training and analysis. Demographics for this subset of participants are shown in Table 3.

Table 3 - T1DEXI Study Demographics – Meal Analysis Subset

Demographics	
Total participants, N	364
Age, years	36.5 +/- 13.8
Biological sex, N (%)	271 (75) F 93 (25) M
Weight, lbs	161.3 +/- 29.2
Ethnicity (self-identified), N (%)	White 333 (91.5) Asian 9 (2.5) Black/African American 7 (1.9) Do not wish to answer/Don't know 8 (2.3) More than one race 5 (1.4) American Indian/Alaskan Native 2 (0.6)
HbA1c, %	6.6 +/- 0.72
Duration of diabetes, years	18.1 +/- 12.9
CGM use, N (%)	Dexcom 312 (88.9) Medtronic 34 (9.7) Abbott 5 (1.4)
Insulin modality	Closed loop 195 (53.6) SAP 169 (46.4)
Total meal events, N	15640
Meal exclusion/inclusion (reason)	4493 – Included 5549 – Excluded (Bolus during PPGR) 2733 – Excluded (Exercise during PPGR) 2641 – Excluded (Insufficient length) 143 – Excluded (Missing CGM)

Selection of target variables

We predicted three components of the PPGR over a three-hour horizon following the start of the meal: netAUC, maxBG, and minBG. NetAUC, as previously defined, represents both positive and negative changes in the PPGR relative to the glucose at the start of the meal. The maxBG and minBG are the respective maximum and minimum blood glucose values over the PPGR prediction window. In many other studies, iAUC is used as a metric to quantify PPGR [25, 50, 51, 52]. However, iAUC fails to describe drops in BG levels in the PPGR which are below the starting CGM. We observed many instances when glucose dropped following a meal. iAUC is not an effective variable for the many instances when substantial glucose drops occurred after a meal was consumed. Instead, we used netAUC as the outcome variable of interest. Prediction of both maxBG and minBG were also used as they are more interpretable glucose outcomes that a person with T1D can use to better understand how their insulin dosing will impact the glucose response.

Pre-processing and feature extraction

Some features analyzed in prior work by Zeevi, et al. [25] were included as inputs in the models that we developed. These meal features include macronutrients for each meal (carbohydrates, protein, fats), alcohol, caffeine, fiber, sodium, sugar, and total calories. In addition, baseline HbA1c from blood testing, participant demographics (sex, weight, height, BMI), total carbohydrates consumed in the 3, 6, and 12 hours leading to the meal, total fiber consumed in the past 12 and 24 hours, and calories consumed in the prior 2, 3, 6, and 12 hours. Carbohydrates-to-fat ratio were computed for each meal and used as features in the model. A small value (1e-12) was added to all fat amounts to allow for the calculation in the case of zero carbs or fats.

Glucose and time of day features were used as inputs as outlined in [57]. These features are insulin on board (IOB), total daily insulin requirement (TDIR), participant age, insulin-to-carbohydrate ratio, and ratio of IOB relative to TDIR. TDIR was estimated for each participant based on median of insulin summed over all days while in the study. The insulin-to-carbohydrate ratio (ICR) was determined by a physician and adjusted as necessary. However, we did not have this data available for participants in the study, so a feature that estimates the ratio of short-acting insulin administered relative to the average amount administered was instead calculated for each meal consumed by each participant. We named this feature 'CR_ratio' and computed with the following method. Since the carb ratios may differ for a participant based on the time of day, meal events from each participant were split into several time windows: 11pm to 7am, 7am to 11am, 11am to 4pm, and 4pm to 11pm. For each participant and each meal window, average carb-to-insulin ratios were calculated by dividing each meal carbohydrate amount by amount of insulin delivered for the meal and computing the mean. This resulted in an average carb ratio for each participant in each meal window (CR_avg). Then, for an individual meal event, the ratio of carbohydrates to insulin was calculated (CR_t0). The meal CR_ratio was then computed by dividing CR_avg by CR_t0 to determine if more insulin or less insulin was taken for that meal than normal. This method has a limitation of relying on participant-estimated carbohydrates, which may introduce additional uncertainty in model predictions. Finally, for each dose of basal and bolus insulin, the feature for IOB assumed a linearly decay of insulin boluses given over a four-hour period. A table of features and calculations is provided in Table 4 and distributions of features in provided in Figure 4.

In order to create feature vectors for training the machine learning algorithm, meal times were located for each subject as recorded during the study in food logs. While user entered macronutrient estimations were available, we instead used the nutrient estimation as reviewed and confirmed by nutritionists via RFPM. This is because, as previously mentioned, user estimated macronutrient content can be less accurate when compared to clinician estimations.

For each meal record, an event window was constructed spanning between two to three hours following a meal and preceding the meal, based on availability of data. To reduce noise, some events were filtered out based on possible disturbances occurring. Specifically, meal events were excluded from analysis if exercise occurred during the meal window, the meal window was less than two hours, a bolus was taken more than 30 minutes after consumption of the meal, or if more than 75% of CGM data was missing after linear interpolation was applied when there were CGM gaps of less than or equal to 20 minutes. To allow for more meal events for training, a meal event was not excluded if a second insulin bolus occurred within the first 30 minutes of consuming a meal. If a bolus was taken 30 minutes after the meal began, the meal time was set to be the time of bolus delivery. For this case, features and targets were also calculated beginning at the time of bolus delivery.

Table 4 - Input features and calculations

Feature Name	**Calculation**
Meal carbohydrates content (g)	As evaluated by RFPM
Meal protein content (g)	As evaluated by RFPM
Meal fat content (g)	As evaluated by RFPM
Meal alcohol content (oz)	As evaluated by RFPM
Meal caffeine content (mg)	As evaluated by RFPM
Meal fiber content (g)	As evaluated by RFPM
Meal sodium content (mg)	As evaluated by RFPM
Meal sugar content (mg)	As evaluated by RFPM
Total meal calories (Kcal)	As evaluated by RFPM
Sum of carbohydrates (g) consumed prior to meal over: - 3 hours (36 samples) - 6 hours (72 samples) - 12 hours (144 samples)	$\sum_{i=0}^{N} carbohydrates_i$ $N \in \{36, 72, 144\}$
Sum of fiber (g) consumed prior to meal over: - 12 hours (144 samples) - 24 hours (288 samples)	$\sum_{i=0}^{N} fiber_i$ $N \in \{144, 288\}$
Sum of calories (Kcal) consumed over the previous:	

- 2 hours (24 samples) - 3 hours (36 samples) - 6 hours (72 samples) - 12 hours (144 samples)	$\sum_{i=0}^{N} calories_i$ $N \in \{24, 36, 72, 144\}$
Participant HbA1c	As reported in baseline demographics survey
Paricipant sex (M/F)	As reported in baseline demographics survey
Participant age (years)	As reported in baseline demographics survey
Participant weight (lbs)	As reported in baseline demographics survey
Participant height (in)	As reported in baseline demographics survey
Participant BMI	As reported in baseline demographics survey
Glucose at the time of prediction [57]	CGM_k
Time series of glucose measurements corresponding to one-hour worth of data sampled at 5-minute intervals before the prediction time [57]	CGM_{k-h} $h \in \{1,2,3, \dots, 12\}$
Glucose rate of change (GROC) at the time of prediction [57]	$\frac{CGM_k - CGM_{k-1}}{\Delta t}$
Average GROC during the hour prior to prediction [57]	$\frac{1}{13}\sum_{h=0}^{12} GROC_{k-h}$
Count of GROC values over the hour prior to prediction that are greater than pre-defined thresholds [57]	$\sum_{h=0}^{12} f(GROC_{k-h})$ $f(GROC_{k-h}) = \begin{cases} 1, if\ GROC_k > th \\ 0, otherwise \end{cases}$ $th \in \{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 6.0\}$
Average glucose over one hour prior to prediction time [57]	$\frac{1}{13}\sum_{h=0}^{12} CGM_{k-h}$
Average glucose calculated from two hours to one hour prior to prediction time [57]	$\frac{1}{12}\sum_{h=13}^{24} CGM_{k-h}$

Difference in average glucose during the last half hour vs. the preceding half hour [57]	$\frac{1}{6}\sum_{h=7}^{12} CGM_{k-h} - \frac{1}{7}\sum_{h=0}^{6} CGM_{k-h}$
Average difference between glucose over 30 minutes prior to prediction with respect to the glucose value exactly 30 minutes prior to prediction time [57]	$\frac{1}{6}\sum_{h=0}^{5}(CGM_{k-h} - CGM_{k-6})$
Binary value, set to 1 if the ROC at prediction time is greater than 5 mg/dL/min [57]	$\begin{cases} 1, & if\ GROC_k > 5 \\ 0, & otherwise \end{cases}$
Binary value, set to 1 if the ROC at prediction time is greater than 7 mg/dL/min [57]	$\begin{cases} 1, & if\ GROC_k > 7 \\ 0, & otherwise \end{cases}$
Time of day (hour, 0-23) [57]	$\cos\left(2\pi \frac{hour}{24}\right)$ $\sin(2\pi \frac{hour}{24})$

Model selection

Random forest [59] regression models were trained for each specified target variable. While some of the other publications choose to train gradient boosted trees [25, 50], random forest was selected due to increased interpretability and similar predictive accuracy in early experiments. One weakness with tree-based algorithms, however, is difficulty extrapolating beyond examples observations in the data [60]. This problem was encountered when predicting less common observations (e.g. low glucose) in the tails of target distributions as shown in Figure 2. In an attempt to correct this, inspiration was drawn from other work adapting random forest to learn from imbalanced data in classification contexts [61, 62]. These works point out that a bootstrap sample may not contain any of the less common observations in practice, which increase the difficulty learning to predict. In *Using Random Forest to Learn Imbalanced Data* [61]*, a* modification to random forest is suggested as a potential solution, which increases the ratio of minority examples observed by each tree through the application of a sampling heuristic such as stratified bootstrapping, over and under-sampling, or balancing both the number of minority and majority samples. To adapt this for the regression context, the “balanced random forest” is implemented by computing a histogram over the target distribution and drawing bootstrap samples uniformly from each bin. This ensures that each tree in the forest is trained on a subsample containing rarer examples.

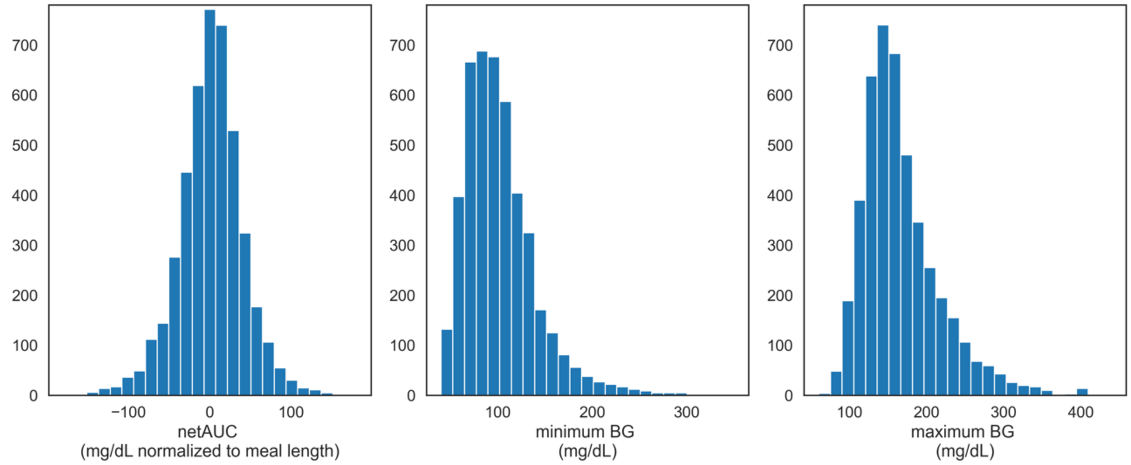

Figure 2 - Target distributions

Figure 3 shows the results of utilizing the balanced random forest algorithm on netAUC. While balancing the random forest helped to overcome some limitations predicting at the lower and upper ranges, as shown in decreased bias in quartiles 1 and 4, the bias was not eliminated. This also came at the cost of decreased accuracy for the under-sampled majority examples residing in quartiles 2 and 3, for which the standard random forest had higher accuracy. In order to try to leverage the accuracy of the random forest model and the balanced random forest model in different quartiles, we employed two different strategies to combine the predictions of each classifier: linear regression and Multivariate Adaptive Regression Splines (MARS) [63].

The bivariate linear regression model for combining random forest predictions is defined as follows:

$$y = b + a_1x_1 + a_2x_2$$

Equation 2 - Bivariate linear regression

where x_1 represents the prediction of random forest with standard sampling and x_2 represents the prediction with the balanced random forest with uniform sampling to ensure adequate representation of the minority observations. The coefficients a_1 and a_2 are weights that are learned during training. Linear regression fits a plane in this case, attempting to minimize the mean square error between the combined model predictions and the ground truth data.

We explored other ways of combining outputs of the random forest and the balanced random forest. A Multivariate Adaptive Regression Splines (MARS) model was trained for the same task of combining the outputs of the two models. The advantage of using the MARS algorithm lies within its flexibility to adapt to non-linear regions in the data through the use of hinge functions, which may be helpful when assigning higher weight to one region of the target distribution versus another. This results in piecewise linear regression. The specific python implementation of the algorithm used in this work is referred to as EARTH [64].

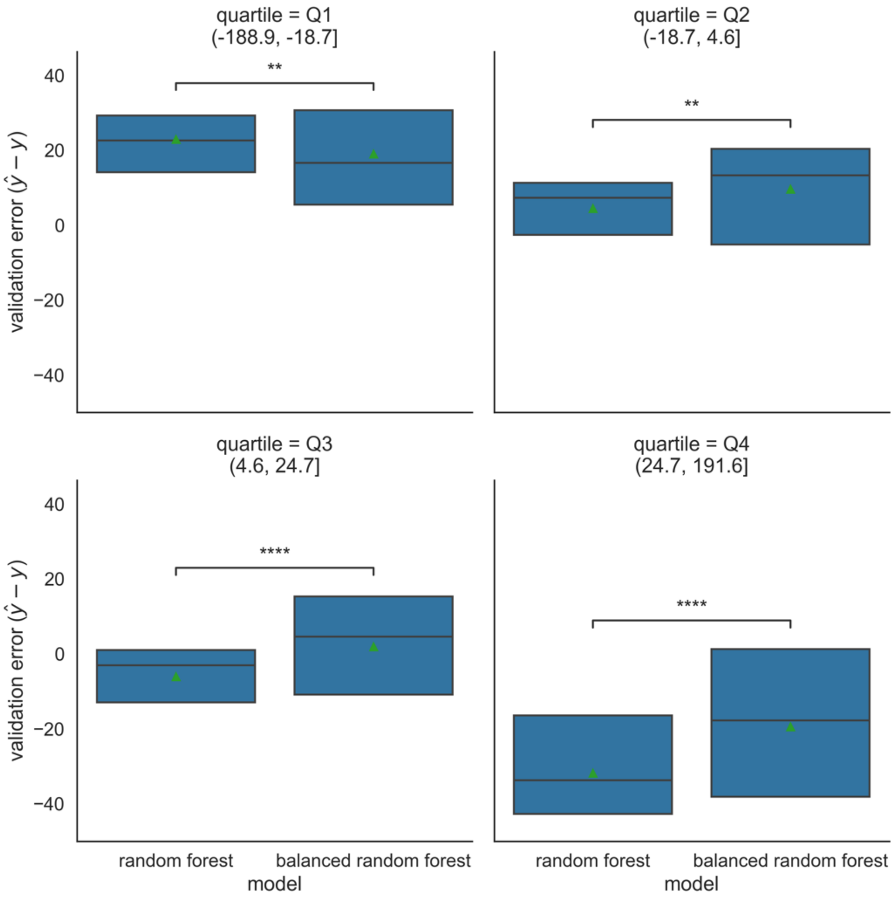

Figure 3 - Quartile bias in random forest versus balanced random forest

Hyperparameter tuning and model validation

In preparation for model training, 4493 meal events were separated randomly into training, validation, and test sets such that each participant was represented in only one of the sets of data. After splitting, the training partition contained 72% of the data, while the remaining 15% and 13% were allocated to the validation and test sets, respectively. The distributions of features in each set are shown in Figure 4. Bayesian hyperparameter tuning was then performed on the training and validation sets to select values for maximum tree depth, minimum leaf samples, minimum samples to split a node, and number of trees. The search was executed for 100 different hyperparameter combinations for each model utilizing the Weights and Biases software

[65]. Machine learning models were implemented using the python version 3.8 software and the scikit-learn framework, and visualizations were generated using matplotlib and seaborn libraries [66, 67, 68]. The best hyperparameters were found by tuning and minimizing the mean squared error of the predictions (Table 5). The results of training each individual random forest model and combining their predictions through regression models are shown in the following subsections.

Table 5 - Hyperparameter tuning search ranges and final values

		Best values found		
Hyperparameter	Search range	NetAUC	MinBG	MaxBG
Number of trees	Minimum: 5 Maximum: 1000	301	32	634
Minimum leaf samples	Minimum: 1 Maximum: 10	10	7	9
Minimum samples to split a node	Minimum: 2 Maximum: 10	10	9	7
Maximum tree depth	Minimum: 1 Maximum: 10	10	4	10

Finally, during analysis it was found that simply combining the predictions of each model did not necessarily lead to a single model which provides good accuracy across all regions of each target distribution. However, several models consistently provide better accuracy in some quartiles. For each target, models were ranked based on which performed best in each quartile of the target distribution. We developed a quartile-based selection method which utilized predictions from each model to estimate which quartile in the target distribution the true value was within. Specifically, the output of the models were averaged to obtain the estimate. Based on which quartile the estimate fell within, the prediction from the best performing model in that quartile was selected as the final output. This algorithm is detailed in Equation 3.

For an input meal feature vector:

```
1. Obtain predictions $y_{\text{pred}}$ by running through each model
       $y_{\text{pred}} = \{\hat{y}_{\text{rf}}, \hat{y}_{\text{balanced\_rf}}, \hat{y}_{\text{combination\_linear}}, \hat{y}_{\text{combination\_MARS}}\}$
2. Compute the mean of $y_{\text{pred}}$, $\bar{y}_{\text{pred}}$
       $\bar{y}_{\text{pred}} = \frac{1}{4}\sum_{i=0}^{3} y_{\text{pred}_i}$
3. Use $\bar{y}_{\text{pred}}$ to estimate which quartile in the target distribution the true target value may reside
    a. If $\bar{y}_{\text{pred}}$ is greater than or less than target observation max/min values, set to max/min
    b. For each (upper_bound, lower_bound, quartile_number) in quartile range:
        i. If $\text{upper\_bound} \geq \bar{y}_{\text{pred}} \geq \text{lower\_bound}$
               Return quartile_number
4. Look up best model for quartile_number and return that prediction
```

Equation 3 - Quartile-based selection algorithm

Boxplots showing model performance in each quartile for NetAUC are shown in Figure 5. Wilcoxon signed-rank testing is performed for all pairs using the *statannotations* library [69], only those with significant differences are shown as indicated by asterisks. Repeated measured are accounted for by grouping by subject and computing the mean error in each quartile. For NetAUC, the predictions of the linear model were best in terms of mean error for quartiles 1 through 3. In quartile 1, there were no significant differences between alternative models and all were significantly better than standard random forest. In quartile 2, "combination – MARS" and "combination – linear" yielded lowest median ME and were not significantly different from each other. In quartile 3, the linear regression model provided the best ME. Finally, in quartile 4, the balanced random forest was ranked the best model due to its ME closest to zero.

An alternative view is shown in Figure 6 as a lineplot. Each colored dot represents median ME for the corresponding model in that quartile. The error bars represent the IQR for the ME in that quartile. It is important to note that there is systemic bias in predictions of the model. The positive bias in Q1 and high negative bias in Q4 indicate that predictions are trending towards the distribution mean.

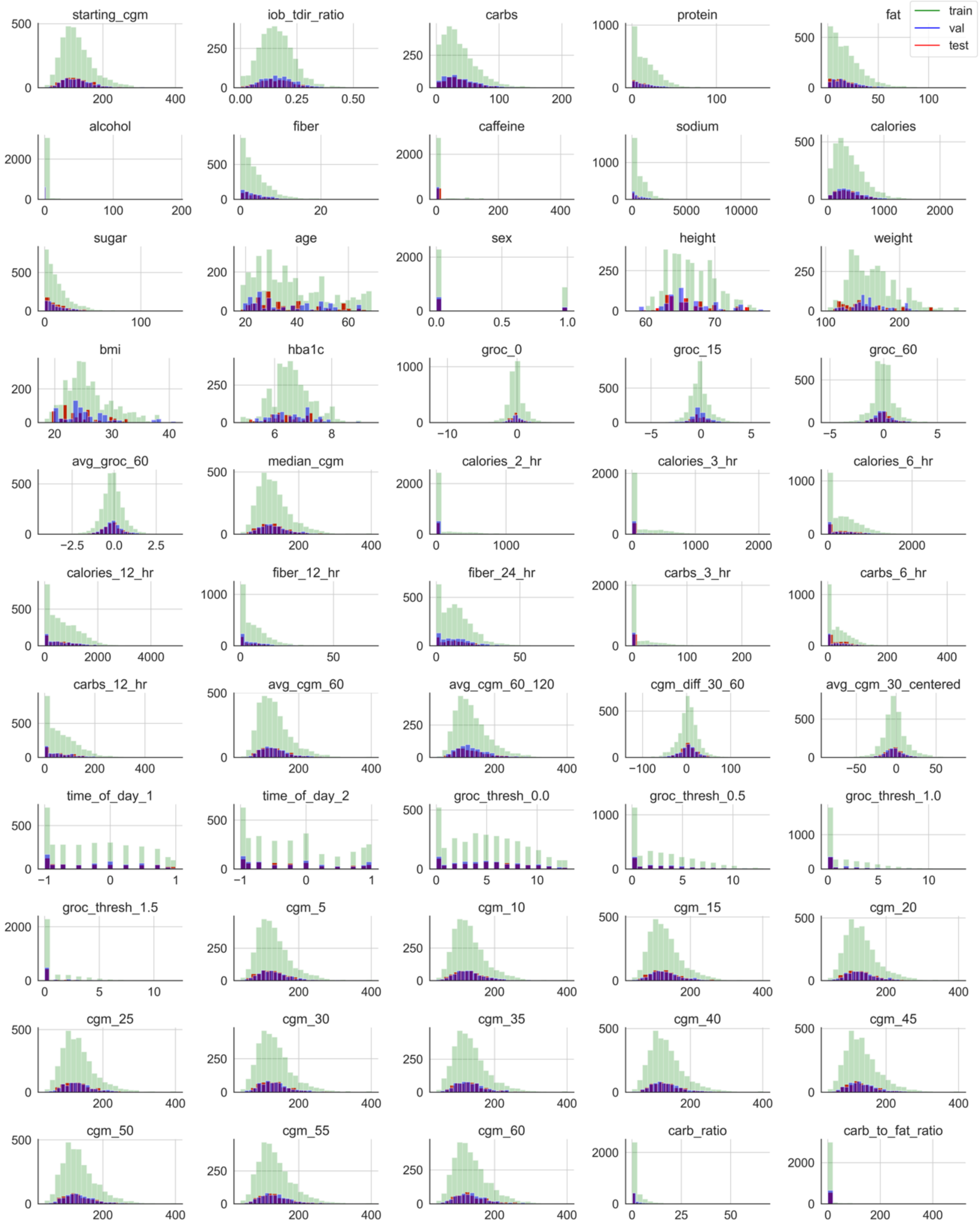

Figure 4 - Input feature distributions

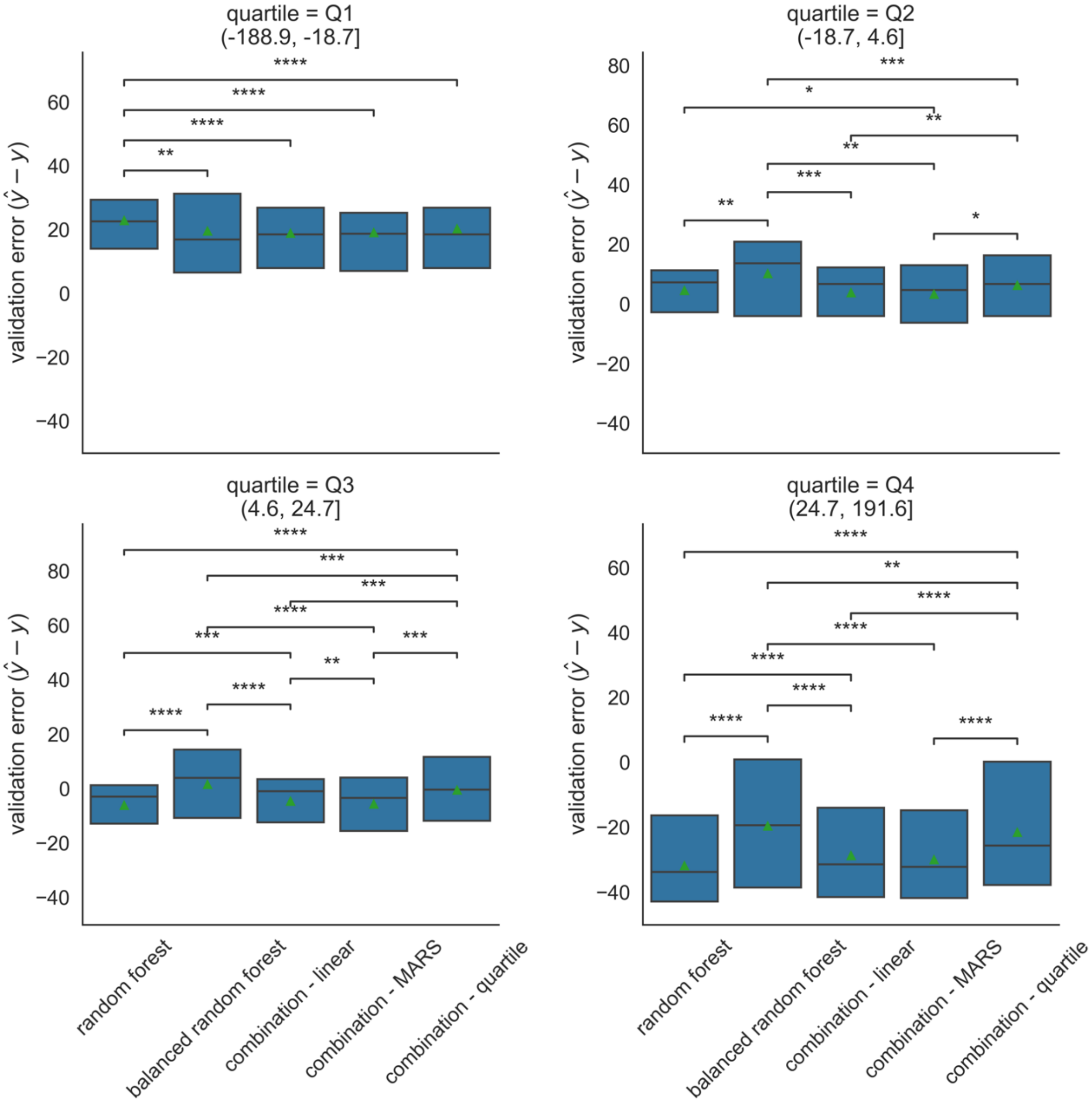

Figure 5 - NetAUC validation mean error across quartiles

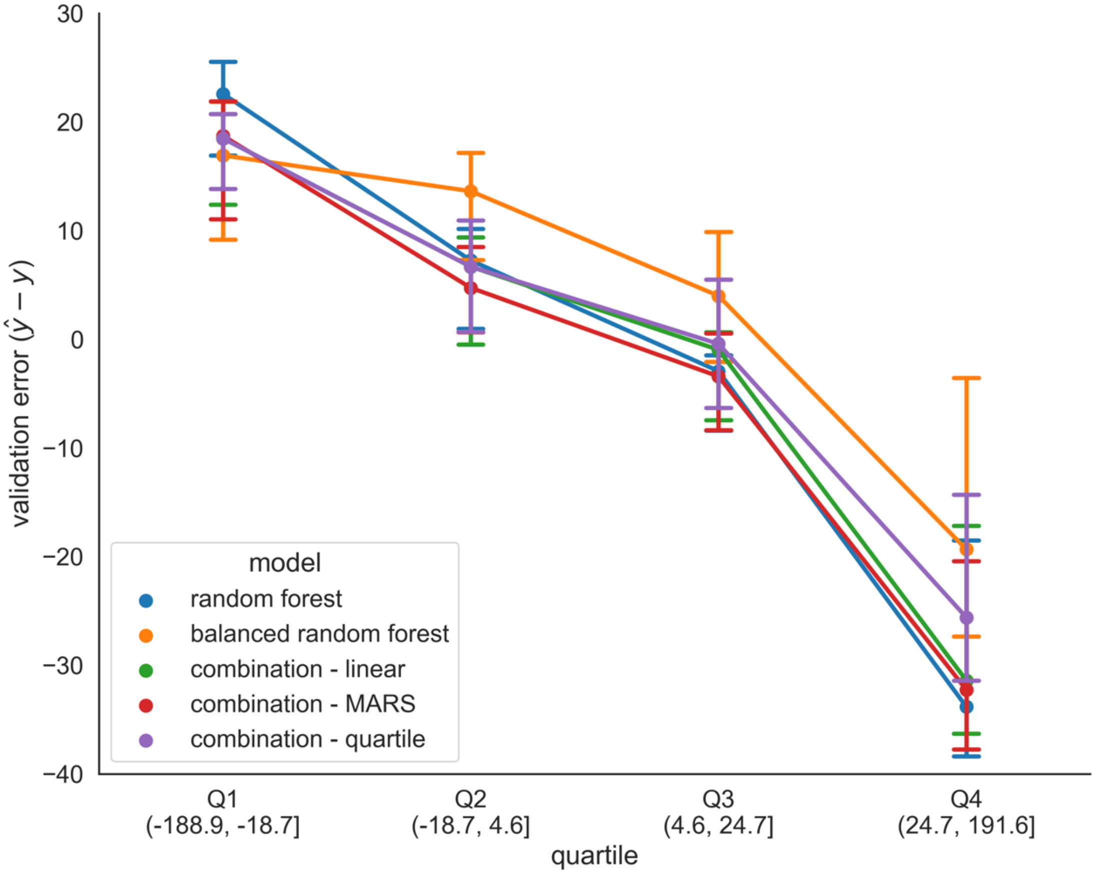

Figure 6 - Lineplot of NetAUC validation mean error across quartiles

The same pattern is followed for MaxBG and MinBG in the following Figure 7-Figure 10 with a summary of the best models in each quartile provided in Table 6. For MaxBG, quartile-based selection strategy appears to work well in balancing the difference in predictions between models across quartiles. The same systemic bias remains as observed in NetAUC modeling.

Table 6 - Best models selected by quartile for each target distribution

Model	Q1	Q2	Q3	Q4
NetAUC	Combination - linear	Combination - linear	Combination – linear	Balanced random forest
MaxBG	Combination - MARS	Combination - linear	Combination – linear	Balanced random forest
MinBG	Combination - linear	Combination - linear	Balanced random forest	Balanced random forest

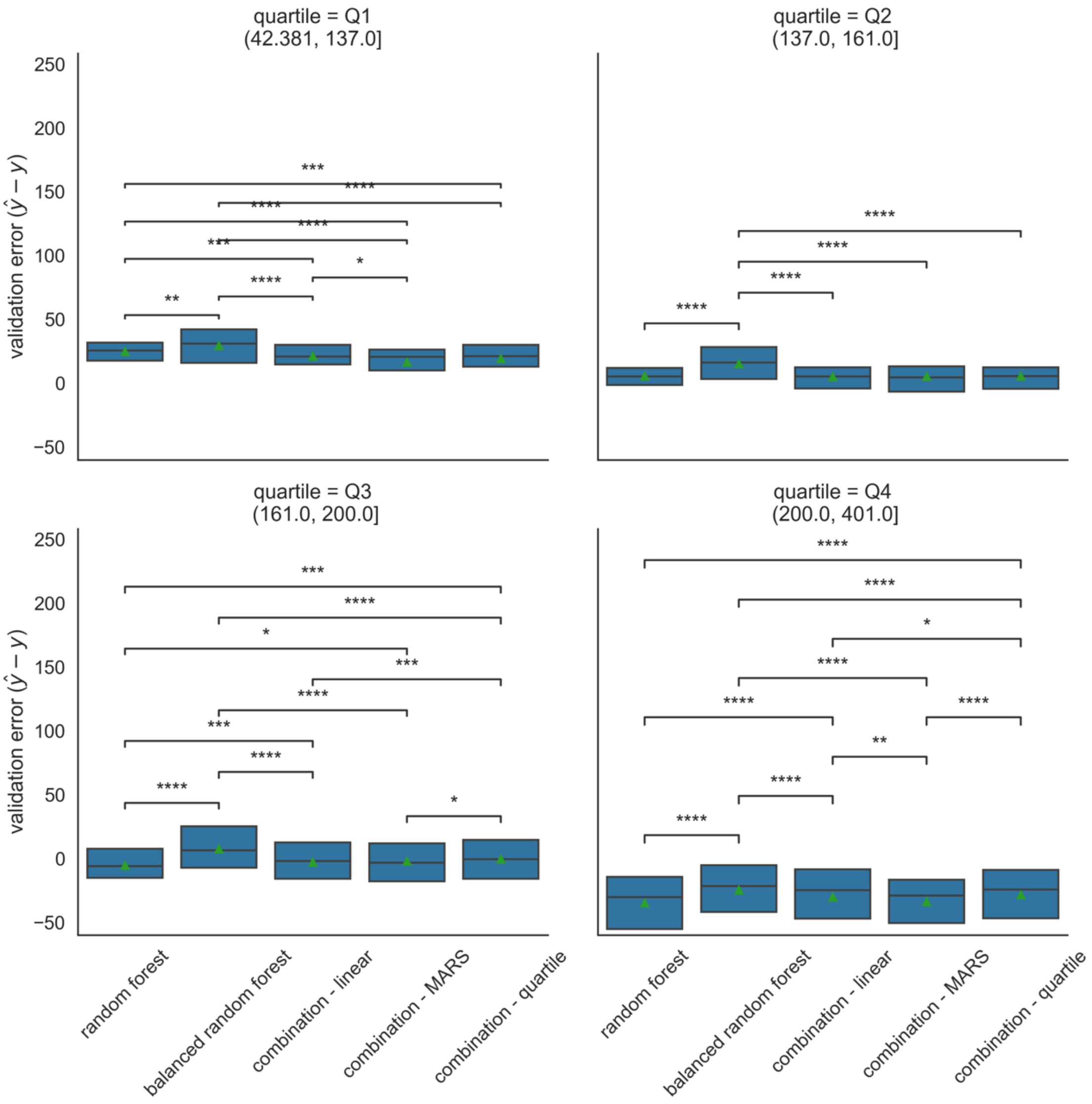

Figure 7 - MaxBG validation mean error across quartiles

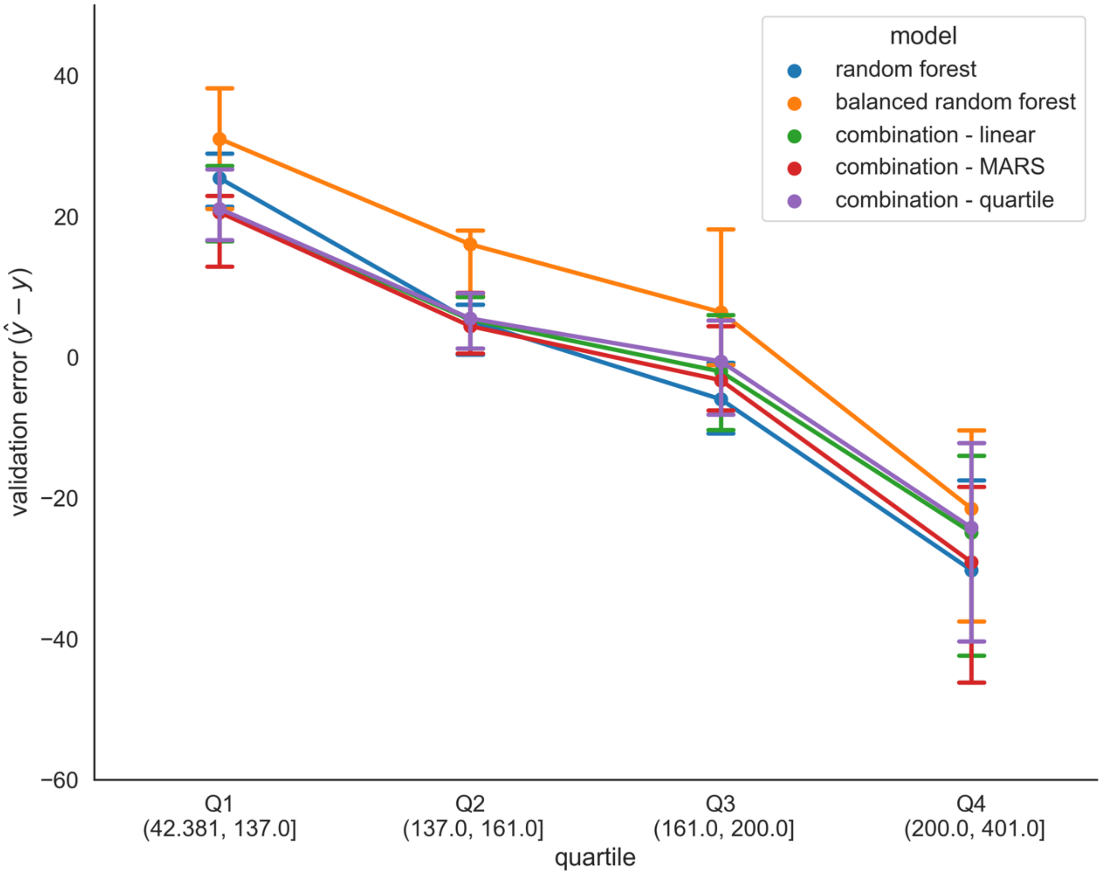

Figure 8 - Lineplot of MaxBG validation mean error across quartiles

Finally, the same pattern is once again observed for MinBG prediction. The quartile-based selection method worked well for this task in balancing the models in quartiles 2 through 4, but ultimately bias still remained. Predictions in quartile 1 could be potentially dangerous given the large error at a low blood glucose range.

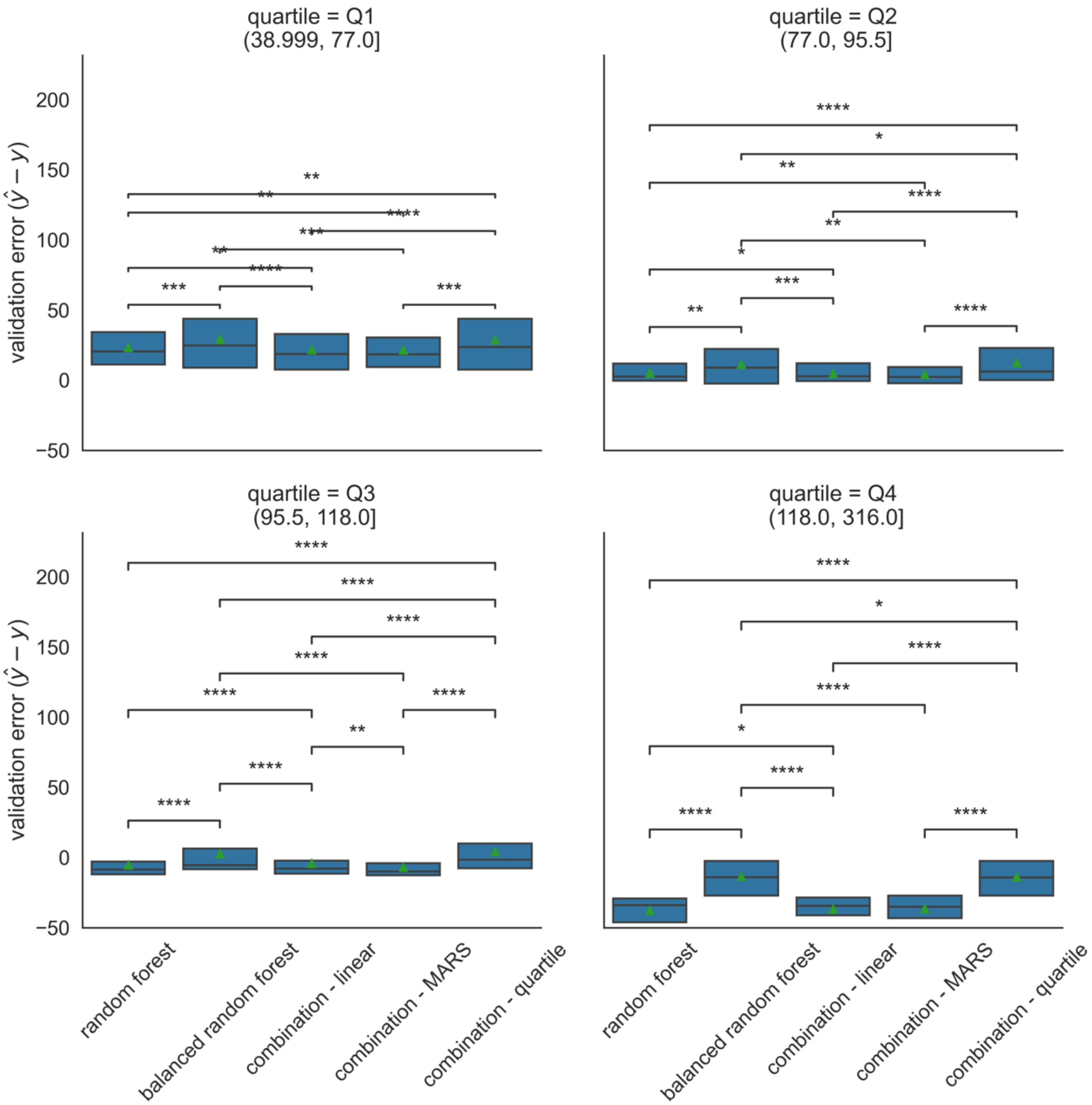

Figure 9 - MinBG validation mean error across quartiles

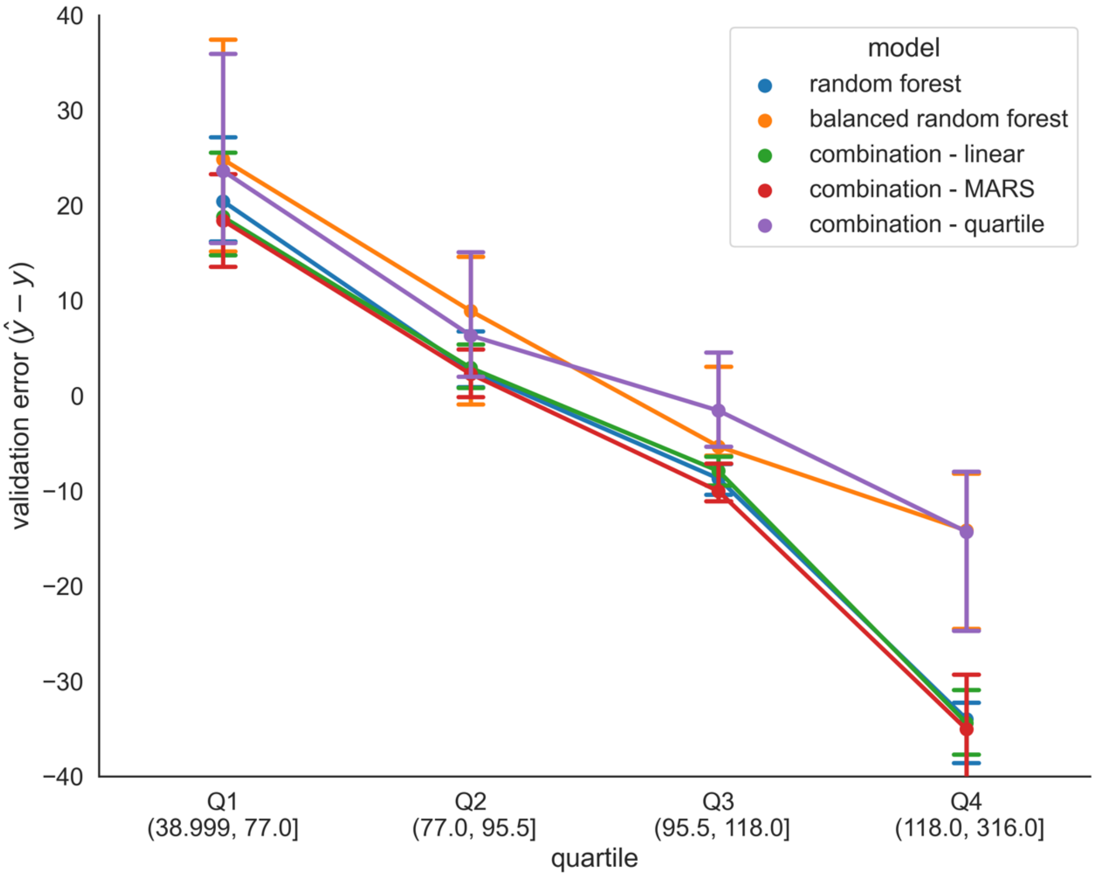

Figure 10 - Lineplot of MinBG validation mean error across quartiles

Results

Following validation, the training and validation data sets are combined and applied to train final models based on the strategy selected in model selection and validation. Final model mean error is aggregated per subject on the holdout sets for each forecasted component of PPGR and shown in the following Figure 16. Generally, the models which were ranked as best in validation were also best in holdout. For NetAUC, the quartile-based selection method improves bias across the first, third, and fourth quartile. There is no significant change in bias in the second quartile, matching accuracy with the standard random forest model. In prediction of maximum blood glucose, the quartile-based selection method also resulted in improved mean error relative to random forest in all quartiles. Finally, for prediction of minimum blood glucose, the results of quartile-based selection closely followed the balanced random forest as also observed in the validation set. Across all models systemic bias remained as can be observed in positive bias in the first quartile steadily decreasing to negative bias in the fourth quartile.

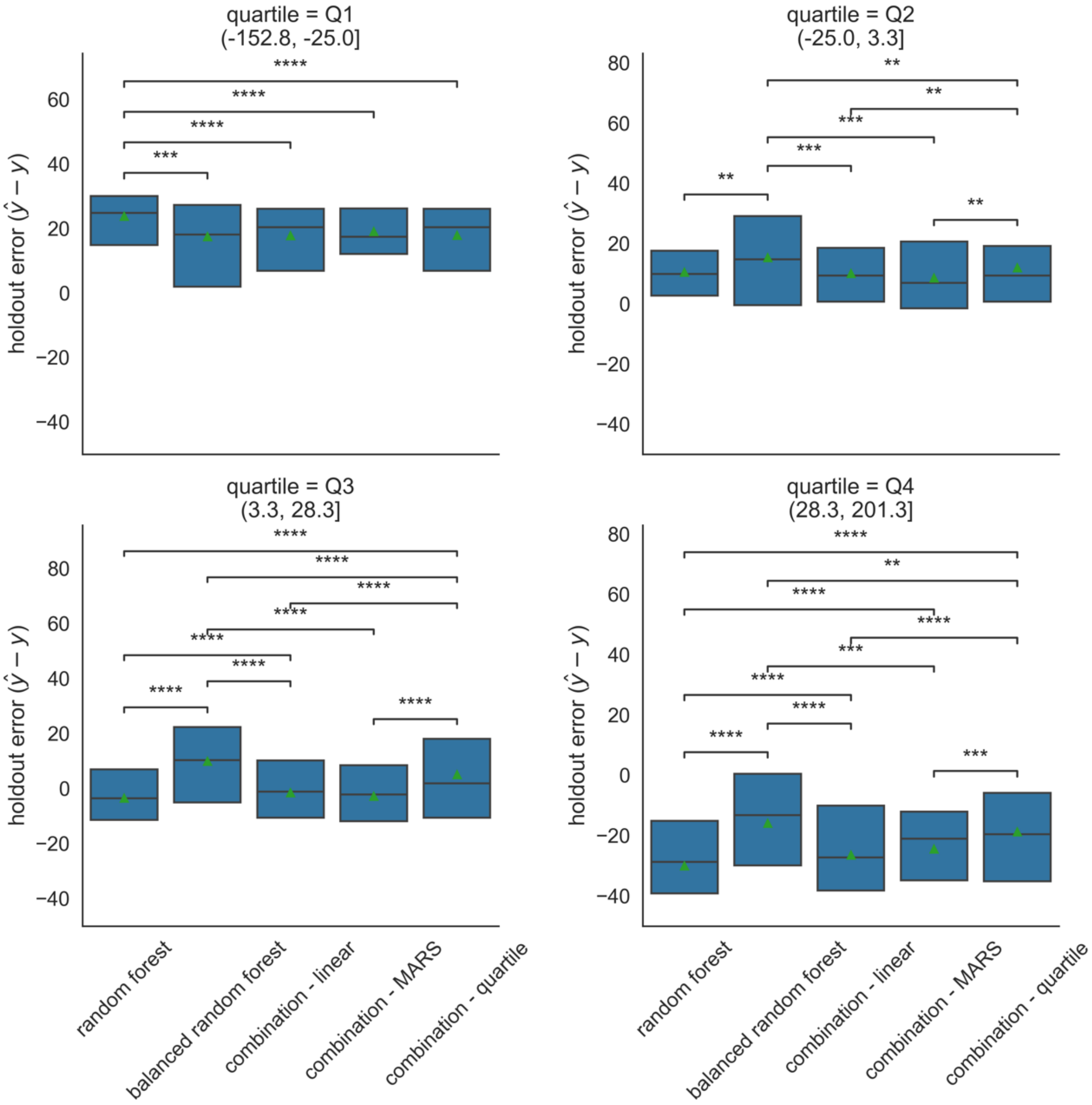

Figure 11 - NetAUC holdout mean error across quartiles

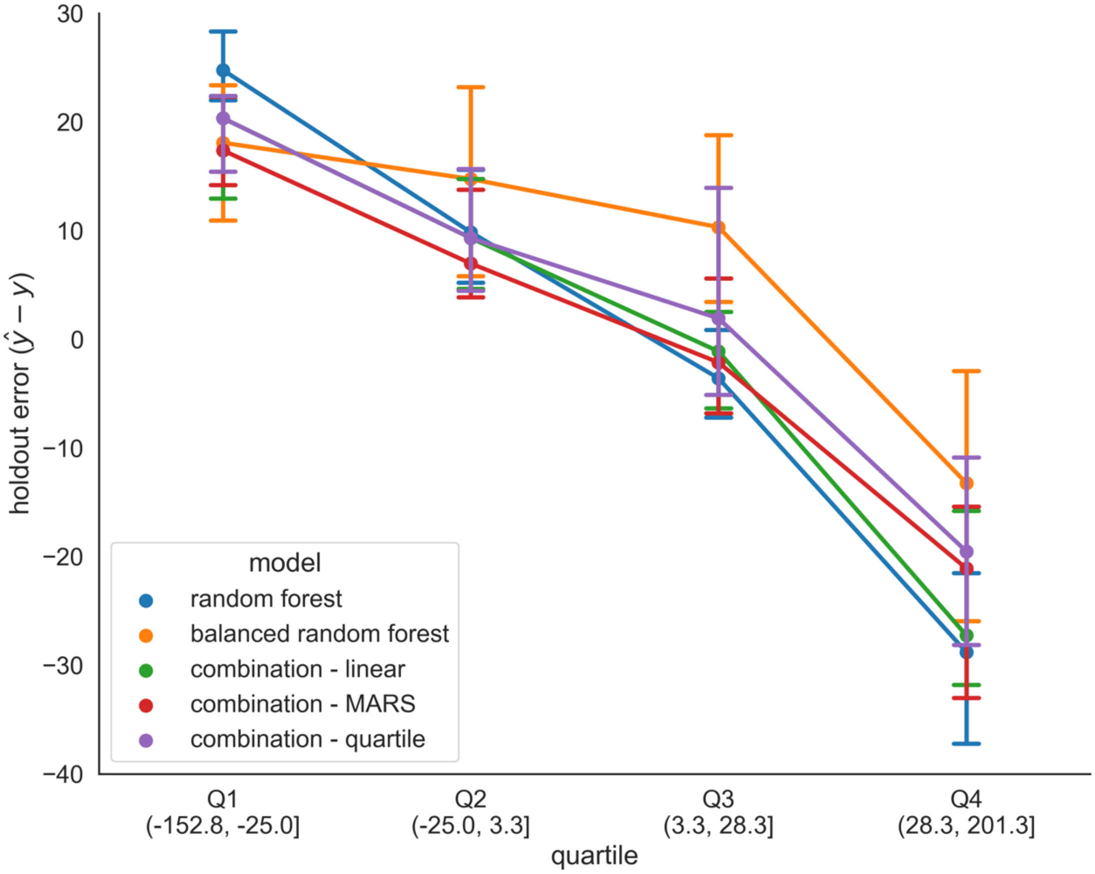

Figure 12 - Lineplot of NetAUC holdout mean error across quartiles

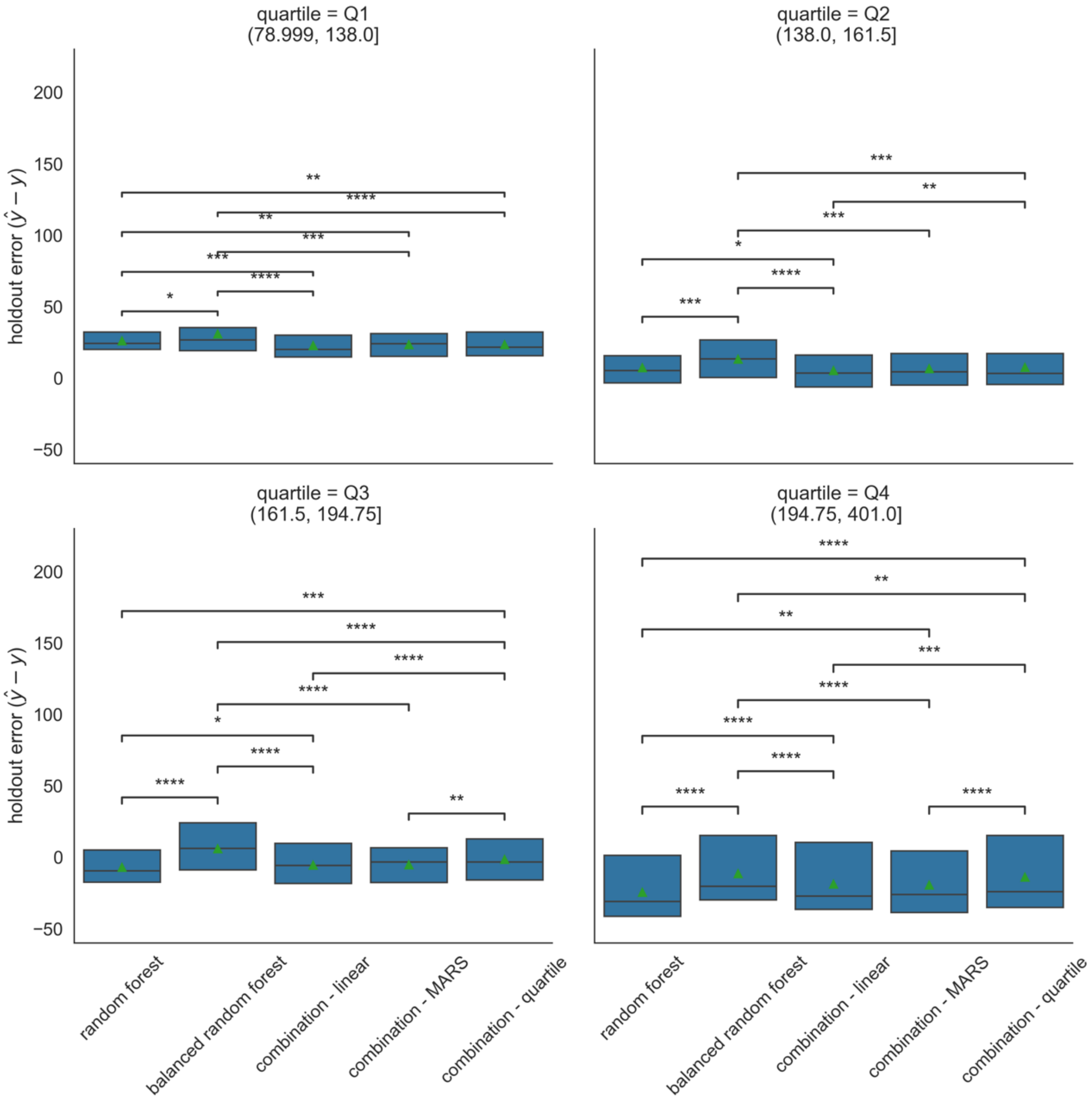

Figure 13 - MaxBG holdout mean error across quartiles

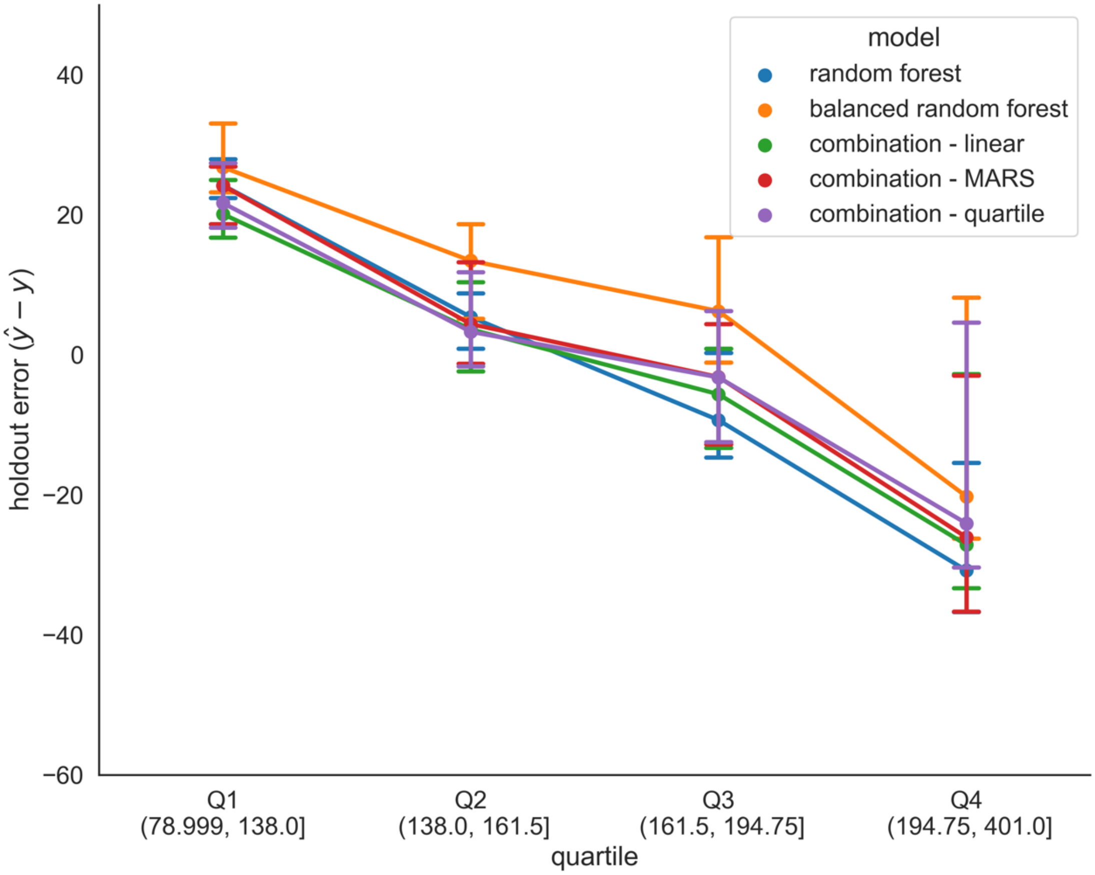

Figure 14 - Lineplot of MaxBG holdout mean error across quartiles

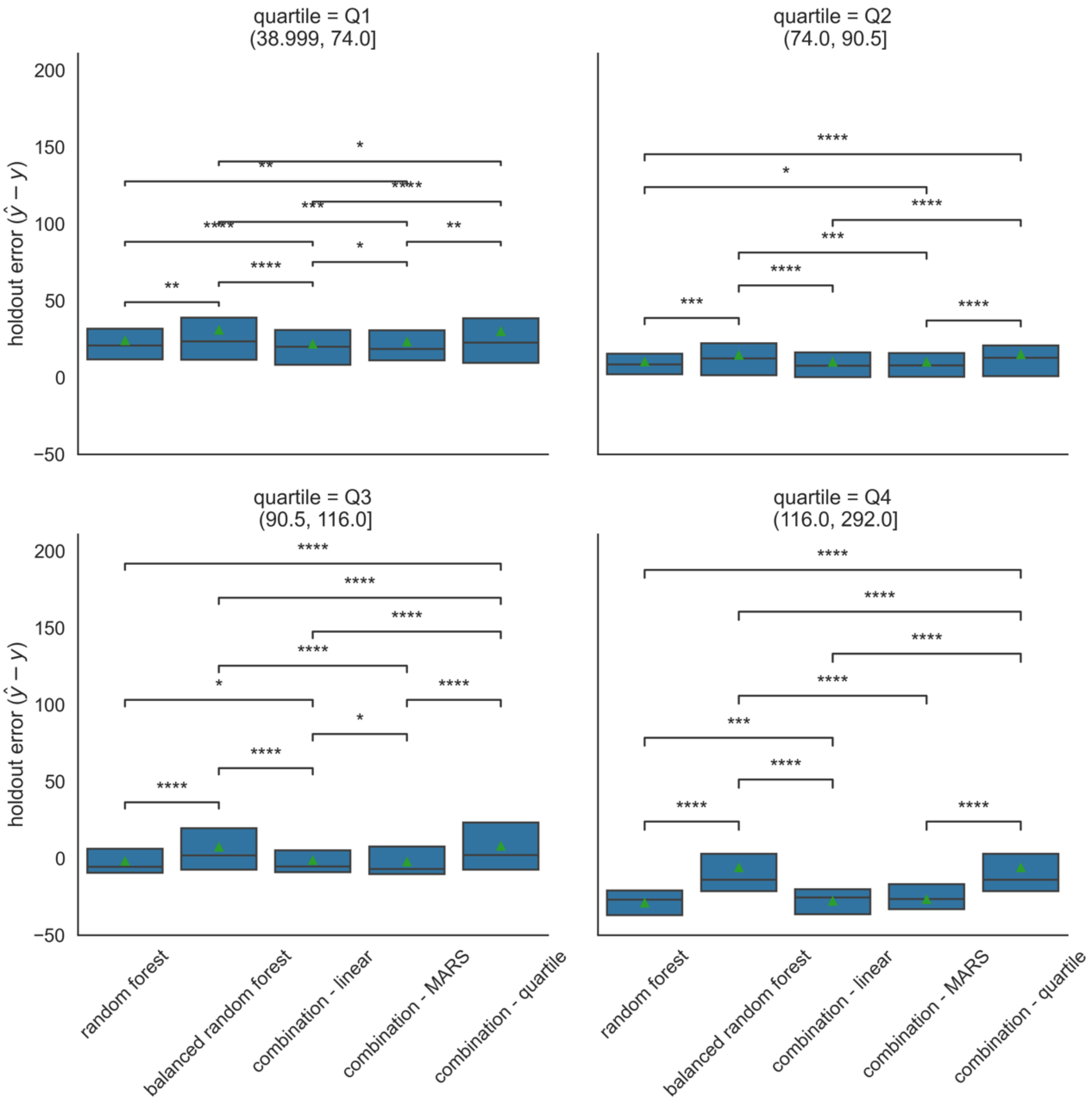

Figure 15 - MinBG holdout mean error across quartiles

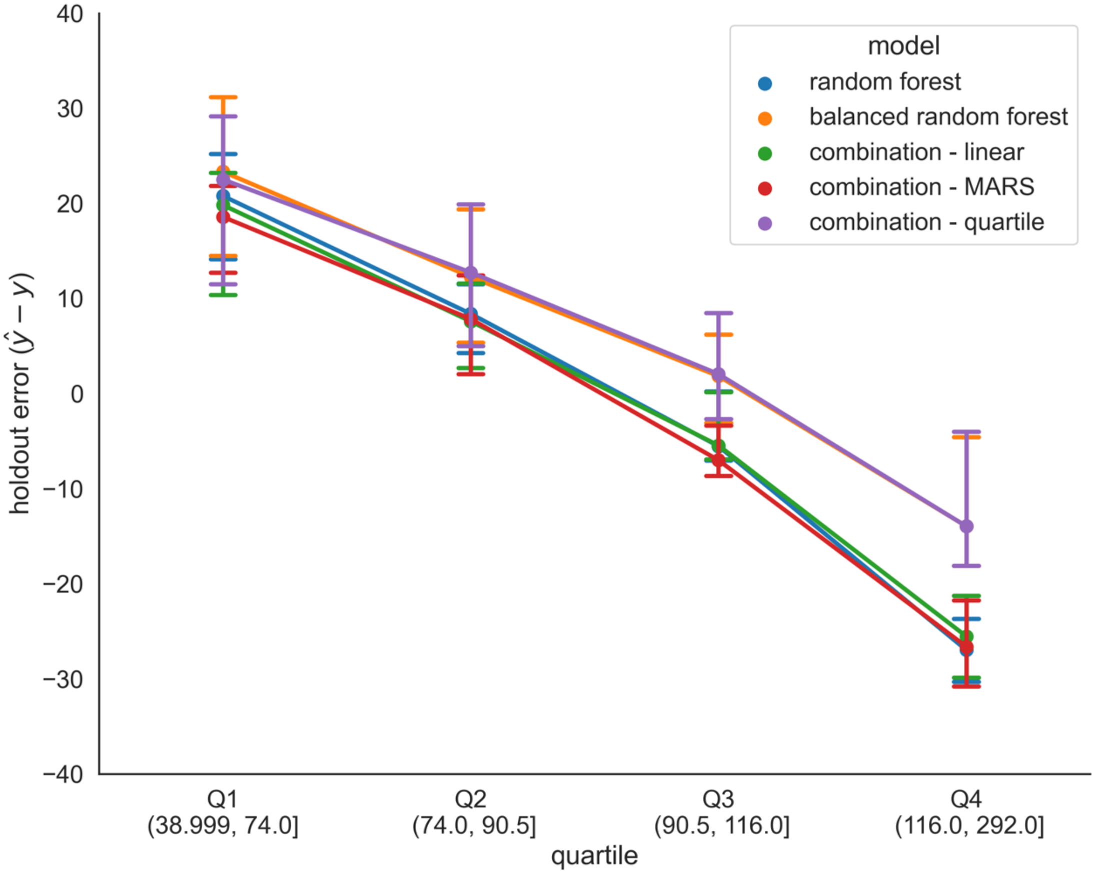

Figure 16 - Lineplot of MinBG holdout mean error across quartiles

Feature importance was also plotted for each random forest model using the standard scikit-learn implementation. Permutation importance was chosen rather than mean decrease in impurity as it is more robust to misleading results due to high cardinality of features. Each feature is shuffled over several iterations to observe the mean decrease in accuracy when the model's relationship between input and its target is broken. The top four feature importances found through this method are listed in Table 7.

Table 7 - Top 4 feature importances as calculated by permutation method

Model	1st important feature	2nd important feature	3rd important feature	4th important feature
NetAUC	Groc_0	Cgm_20	Cgm_30	Hba1c
MaxBG	Starting_cgm	Hba1c	Groc_0	Carbs
MinBG	Starting_cgm	Iob_tdir_ratio	Groc_0	Cgm_5

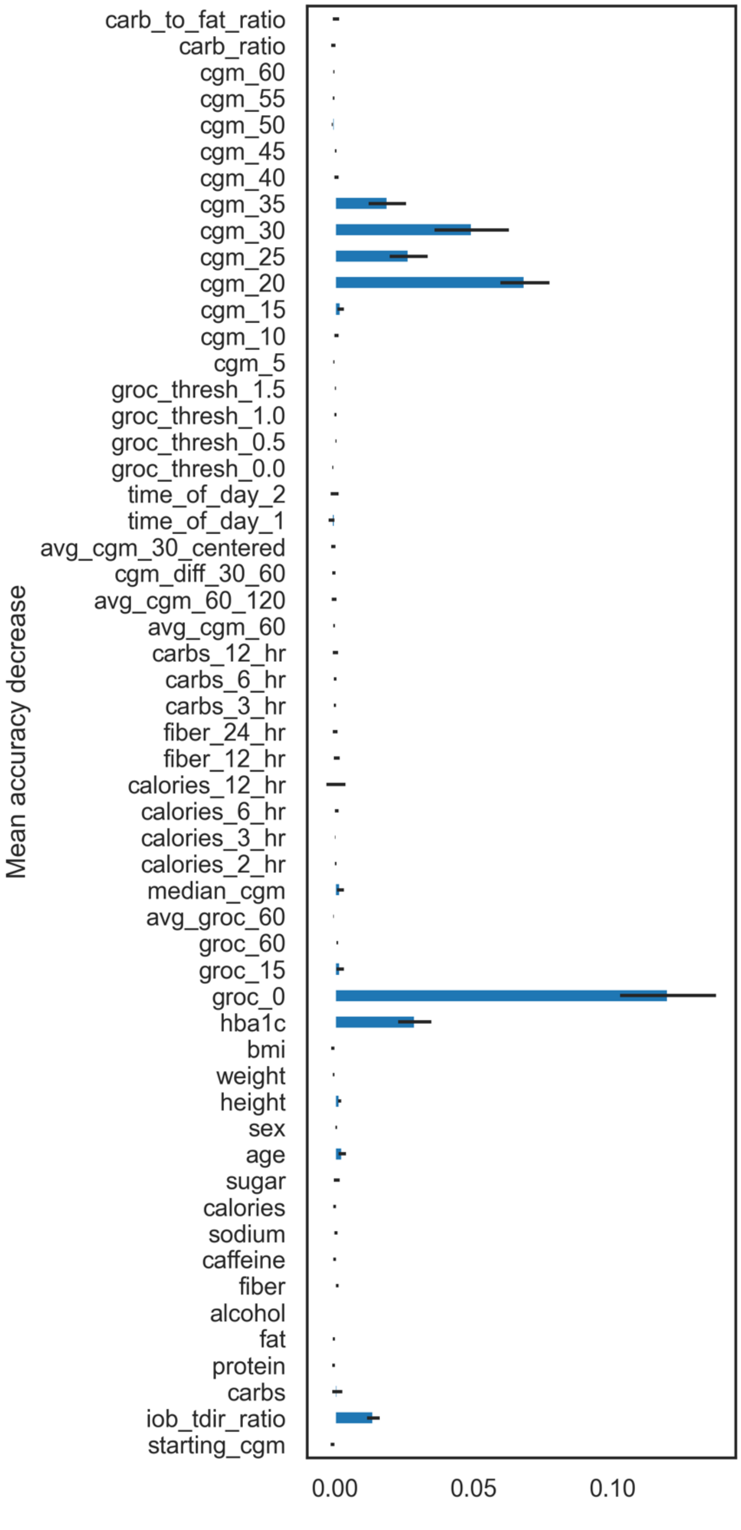

Figure 17 - NetAUC random forest feature importance computed via permutation method on holdout set

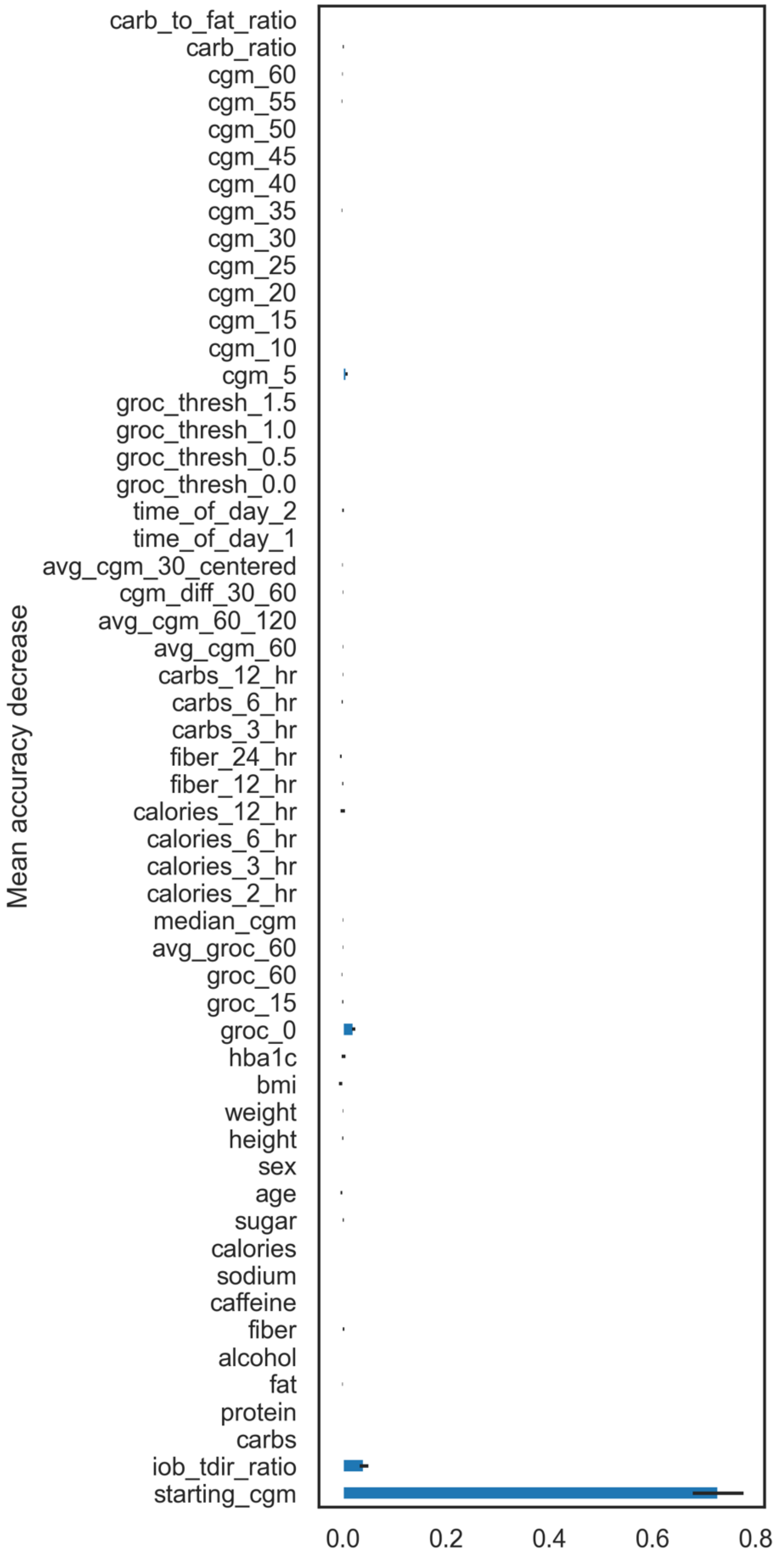

Figure 18 - MaxBG random forest feature importance computed via permutation method on holdout set

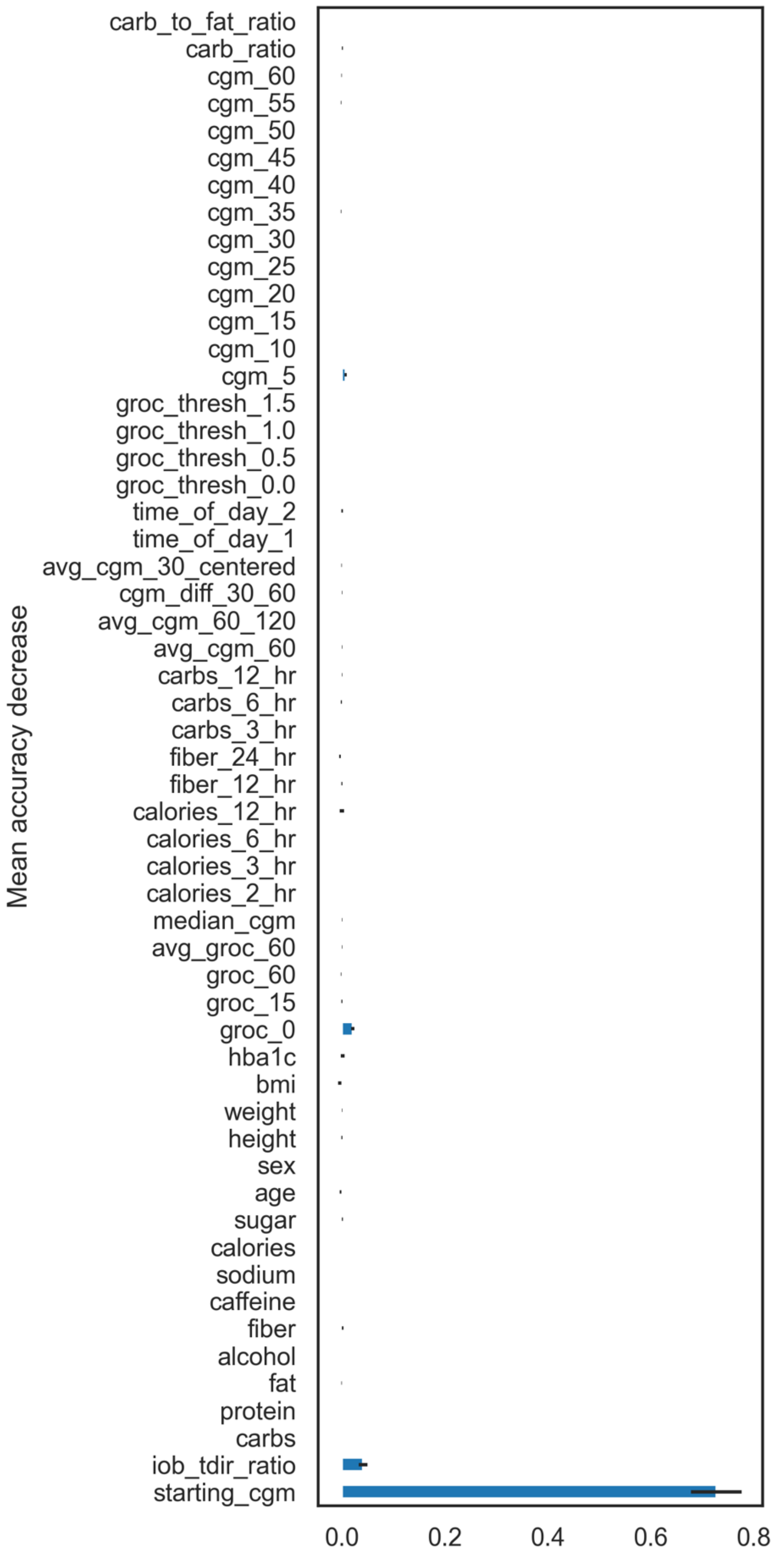

Figure 19 - MinBG random forest feature importance computed via permutation method on holdout set

Feature importance is an interesting diagnostic because it provides insight into which inputs the model finds to be informative when making a prediction given the training data. For NetAUC, the glucose rate of change at prediction time along with prior CGM readings and HbA1c were relied on by the random forest model the most. Other features, such as ratio of IOB to TDIR, height, and age were also important but to a lesser degree. In the MaxBG prediction model, starting CGM was by far the most important feature. HbA1c and glucose rate of change at prediction time contributed, but to a lesser degree. Carbohydrates consumed at the time of meal and calories over the past 12 hours appeared to also have a small impact. Similar feature importance was observed in MinBG prediction, with CGM reading at the time of the meal contributing the most. However, the IOB to TDIR ratio and prandial glucose rate of change were the next most important features. Many of the other features did not contribute substantially to predictive accuracy.

A useful tool to evaluate the safety of the maxBG and minBG predictions is the Parkes consensus error grid [70]. Predictions within the "A" region of the grid are considered to be accurate and safe. Predictions in the "B" region are less accurate but also generally considered to be safe. Regions C-E become progressively more dangerous [24]. In the following Figure 20 and Figure 21, the predictions generated by the quartile-based selection method are shown on this grid. For the maximum BG prediction task, 99.2% of predictions are within the safe regions of A and B. 0.7% of the data lie within the C range, which is slightly more dangerous as a higher BG level is predicted than the actual outcome. There are no predictions within the more dangerous D and E zones. Finally, for minimum BG prediction, 89.4% of predictions are within the safer range. 8.7% of predictions fall within the C zone and 1.5% are within the D zone. Predictions within these zones could be more dangerous to a person using these predictions in decision support settings if they lead to administration of larger insulin boluses when not indicated.

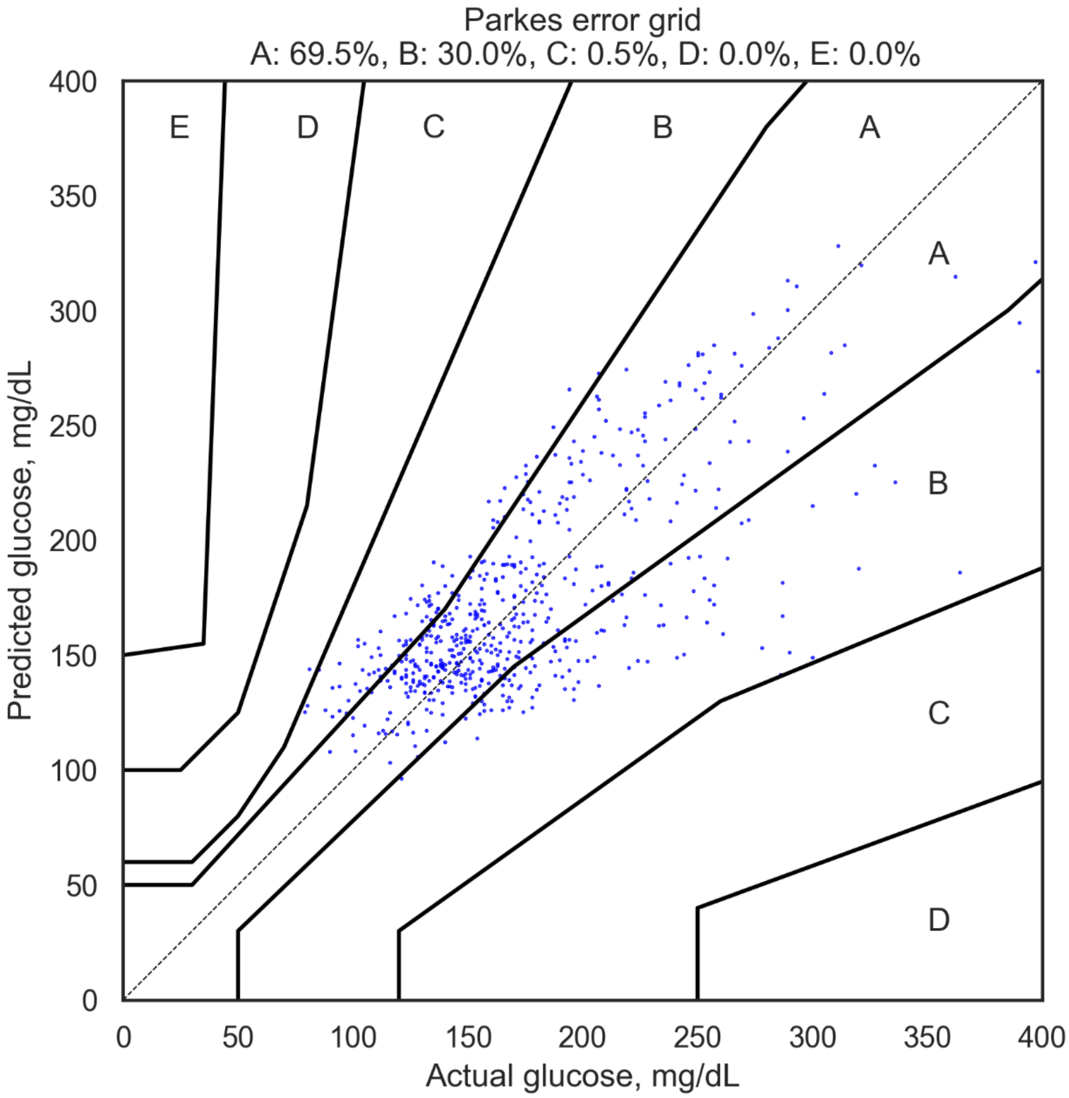

Figure 20 - MaxBG holdout set predictions - Parkes error grid

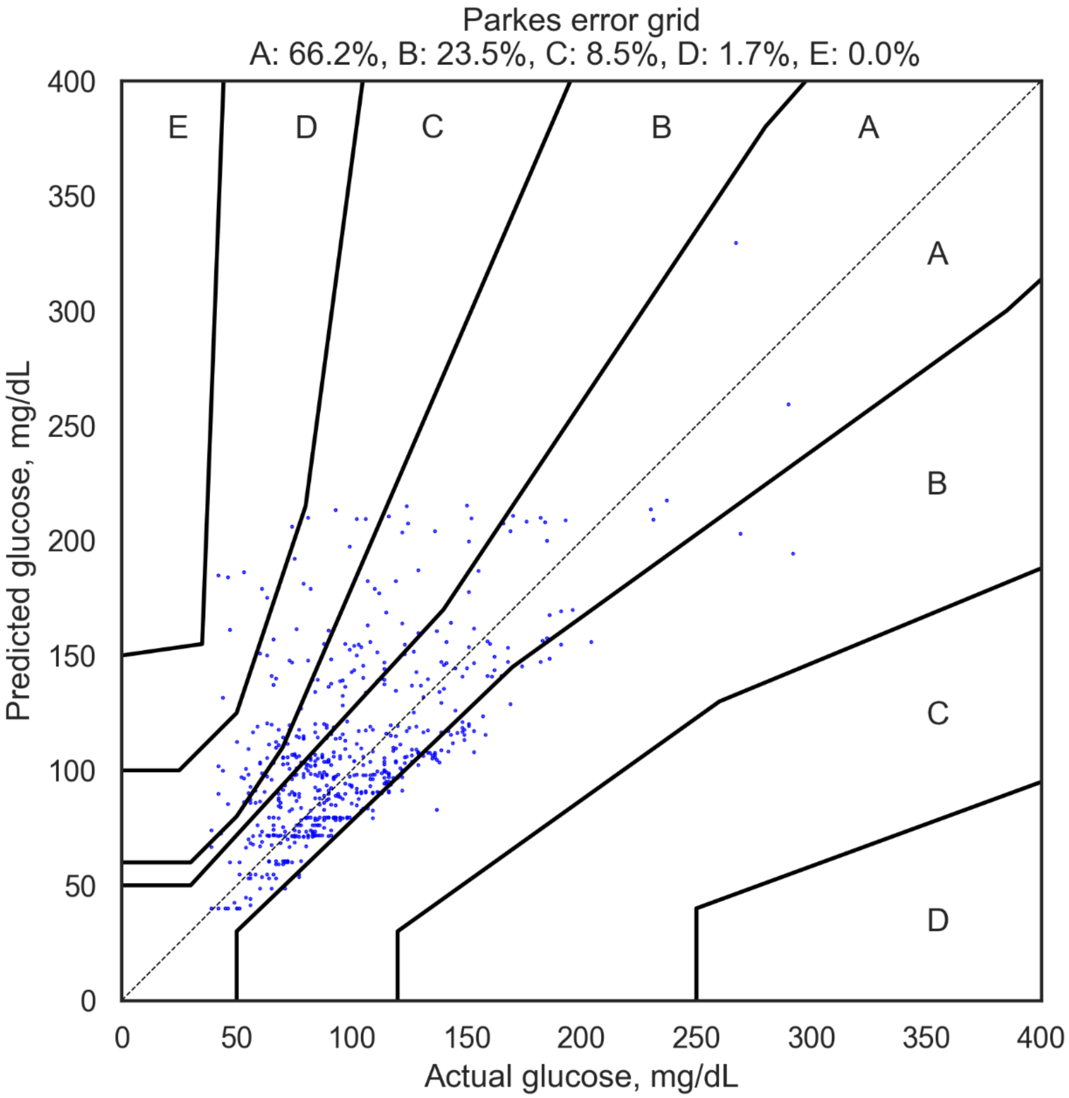

Figure 21 - MinBG holdout set predictions - Parkes error grid

Finally, in Table 8 the final holdout evaluation metrics are reported based on the final predictions of the quartile-based selection method. For all models, high correlation was observed (Pearson R > 0.61). The MaxBG model had the highest coefficient of determination ($R^2 = 0.47$), while the MinBG model had the lowest score ($R^2 = 0.09$). The low score for MinBG implies that the model is insufficient to fully capture the variance in the data, and that more examples or features may be required to improve performance.

Table 8 - Holdout set evaluation metrics

Model	**ME ± std (mg/dL)**	**RMSE ± std (mg/dL)**	**R**	**R^2**
NetAUC	4.68 ± 17.59	20.24 ± 12.64	0.64	0.34
MaxBG	2.98 ± 20.90	33.11 ± 13.68	0.70	0.47
MinBG	14.48 ± 22.59	26.36 ± 21.41	0.61	0.09

Discussion

Overall, we have shown that while there is bias in the predictions of the netAUC, MaxBG, and MinBG of the PPGR, using data balancing approaches significantly reduces this bias. These models may be used to inform decision support around calculation of meal insulin boluses to help people understand how their insulin dosing and macronutrient intake could impact their glucose response. While the maxBG predictions were clinically safe according to the Parkes error grid (Figure 20), use of the MinBG model could be unsafe to use in a clinical setting due to the high bias as observed in mean error (Table 8) and the Parkes error grid analysis (Figure 21). This process highlights the complexity of developing safe and robust augmentations to decision support systems. However, it is important to note that the results of analysis of feature importance can provide insight into patterns the model has learned, which can be iteratively used to revisit and understand the underlying data. In the subset of data included in this analysis, participants were not exercising or consuming additional food within the subsequent three hours following the meal. As the data were sourced from an exercise study initiative, this substantially reduced the amount of events available for the machine learning model to use for training. In addition, many meal events had boluses during the postprandial forecasting window which presented an additional unknown disturbance.

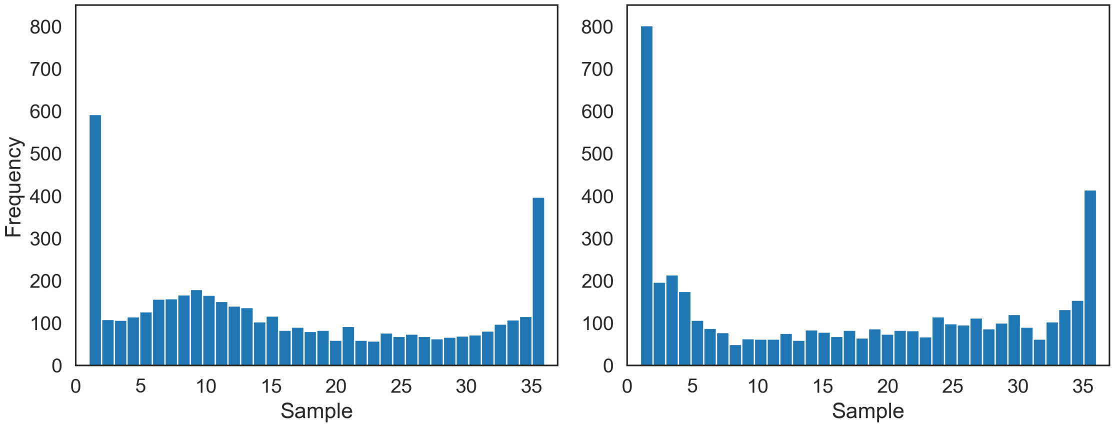

Figure 22 - Histogram of time-to-maximum blood glucose (5-minute sample number) in event window (left) Histogram of time-to-minimum blood glucose (5-minute sample number) in event window (right)

Examining feature importance from the models shows that for NetAUC, the glucose rate of change, prior CGM readings, and IOB to TDIR ratio were important to the prediction. The model appeared to leverage nutritional information but to a lesser extent. When examining feature importance for MaxBG and MinBG, the most important feature by far was the starting CGM, followed by IOB to TDIR for MinBG. For minimum BG, normalized insulin on board makes sense as a predictor as it will directly influence lowering of blood glucose over the prediction window. It is more surprising, however, that the starting CGM value is the primary influence of accuracy for these two models. Taking a closer look at the data itself helps to reveal why this may be. In Figure 22, time to maximum blood glucose and time to minimum blood glucose are plotted in terms of frequency and the 5-minute sample number for which the statistic occurred. For the majority of events, participants began eating when they had high blood glucose and after eating glucose dropped. The second highest majority of observations had glucose that increased consistently during the postprandial glucose prediction window. The third highest number of observations peaked slightly less than an hour into the event. Similarly, in the case of minimum blood glucose, many participants entered the meal event low or were declining to a low within the following 15 minutes. The second highest majority hit a low at the end of the three-hour postprandial glucose prediction window. It is possible that this pattern is picked up by the machine learning algorithm during the learning process. If this is truly the case, more data would not help solve the problem, and instead a different approach to learning from data would be needed. One such approach could involve clustering PPGRs based on whether or not they are expected to rise, fall, or exhibit other patterns. Based on the clusters, patterns in the data could be observed and classifiers trained to predict outcomes if the patterns are predictable with respect to input features.

Finally, data imbalance is a notorious and challenging problem in machine learning. While we attempted to correct for this imbalanced through balanced sampling techniques, the systemic bias ultimately remained and could not be overcome with the balanced random forest alone. This resulted in overall predictions from all models trending towards the distribution means. Evidence of this pattern can be observed in Figure 16 in which there is consistent positive bias in the first quartile, declining across each quartile to high negative bias in the fourth quartile. However, it is worth noting that the work did offer insight on how modifying the bootstrap sample can modestly boost predictive accuracy in the regression context. Specifically, the balanced random forest consistently decreased bias in the fourth quartile for each model. In addition, the quartile-based model selection method resulted in good compromise between the predictions of the collection of models. Despite these methods, ultimately the quality and distribution of data matters and can be difficult to overcome with imbalance correction alone.

Limitations

The limitations of this work should be discussed. As detailed in Methods, data are only included from participants of AID and open loop pump modalities. This is because MDI users needed to manually log when they dosed with long-acting and short-acting insulin, resulting in a lack of reliable data. In the future, devices such as Bluetooth-enabled smart pens may help to more accurately track doses. For AID modalities, the insulin pump automatically logs this information making it more accurate. Another limitation lies within the available data. Other studies have shown microbiome has a significant impact on PPGR [25, 50]. Such data were not available for this analysis, which may have limited the predictive accuracy of the ML models. These studies also involved standardized meals for participants in order to limit the variability introduced by different dietary choices. We also use nutritional information as evaluated by nutritionists, which will not be available in practice when used within an MDI or decision support system. When an individual's own estimation is to be relied on, this will introduce additional uncertainty in model predictions. Finally, the exclusion of meal events based on possible disturbances such as exercise, an additional meal, or bolus delivery resulted in a much narrower set of events for analysis than commonly encountered in daily life of participants.

Chapter 4 - Conclusion

There still exists ample opportunity to continue work on improving DSSs. While overall use of the DailyDose app did not lead to improved TIR, post-hoc analysis revealed that adherence to provided recommendations leads to increased TIR. This is promising for future work aiming to improve overall rate of recommendation acceptance. The integration of an algorithm which is better able to anticipate personalized prandial insulin boluses given nutritional information could also be promising and help improve glycemic outcomes. Final model predictions in maximum blood glucose forecasting task yielded 99.5% safe predictions in Parkes error grid analysis, indicating potential usefulness in clinical settings. In the minimum blood glucose forecasting model, this was lowered to 89.7% of predictions increasing the risk of danger to participant. Attempts to improve model forecasting were consistently helpful in the fourth quartile, but ultimately systemic bias persisted throughout all models indicating that more work is to be done to improve performance. Through examining feature importance in the random forest, it was shown which factors were heavily utilized during prediction time. These features are useful in better understanding patterns in the data to build more robust models.

Future work

DSSs such as DailyDose may be improved with the integration of an algorithm which can successfully and accurately forecast components of PPGR. There are several benefits to including PPGR prediction into the app, including the ability to improve bolus calculation with personalized incorporation of meal macronutrients and the ability to change meal macronutrient profile to see how the delivered bolus and response might change. Another benefit of including PPGR models is increased explainability. Interviews with study participants revealed that many participants may not have understood recommendations or the benefit of the recommendations.

Removing this barrier may help increase acceptance of recommendations and possibly lead to better glycemic control.

More work is needed for a postprandial glucose forecasting algorithm to be integrated with a DSS such as DailyDose. First, it will be necessary to include data from MDI users in modeling. In addition, the systemic bias would need to be corrected as observed in the forecasting models presented here. Feature importance analysis revealed that the primary driver for minimum and maximum BG prediction was the starting CGM. It is possible that the starting CGM influences the learned bias, which may be alleviated with more data procured through a future study. However, more data would only be helpful if the time-to-max and time-to-min changes with more samples as well. Ultimately, different approaches to modeling may be required, such as a clustering technique which groups PPGRs based on trend and learns which features are associated with the PPGR. Other factors have also been observed in studies to influence PPGR, such as microbiome and standardized meals for participants [25, 50]. Controlling for meal variance could in turn also increase the accuracy of reported nutrients in a future study. Through additional study data and application of alternative modeling techniques, it may be possible to leverage these models in a way which improves quality of life for those living with T1D.

References

[1] Centers for Disease Control and Prevention, "National Diabetes Statistics Report," 2023. [Online]. Available: https://www.cdc.gov/diabetes/data/statistics-report/index.html.

[2] A. L. Burrack, T. Martinov and B. T. Fife, "T Cell-Mediated Beta Cell Destruction: Autoimmunity and Alloimmunity in the Context of Type 1 Diabetes," *Frontiers in endocrinology,* 2017.

[3] J. B. McGill and A. Ahmann, "Continuous Glucose Monitoring with Multiple Daily Insulin Treatment: Outcome Studies," *Diabetes technology & therapeutics ,* vol. 19, no. S3, p. S3–S12, 2017.

[4] Sherr, Jennifer L., et al., "Automated Insulin Delivery: Benefits, Challenges, and Recommendations. A Consensus Report of the Joint Diabetes Technology Working Group of the European Association for the Study of Diabetes and the American Diabetes Association," *Diabetes Care 1,* vol. 45, no. 12, pp. 3058-3074, 2022.

[5] M. E. Pauley, C. Berget and L. H. Messer, "Barriers to Uptake of Insulin Technologies and Novel Solutions," *Medical devices (Auckland, N.Z.),* vol. 14, p. 339–354, 2021.

[6] Foster, N. C., Beck, R. W., Miller, K. M., et al., "State of Type 1 Diabetes Management and Outcomes from the T1D Exchange in 2016-2018," *Diabetes technology & therapeutics,* vol. 21, no. 2, pp. 66-72, 2018.

[7] R. M. Bergenstal, "Understanding Continuous Glucose Monitoring Data," *Role of Continuous Glucose Monitoring in Diabetes Treatment,* 2018.

[8] S. Schmidt and K. Nørgaard, "Bolus calculators," *J Diabetes Sci Technol,* vol. 8, no. 5, pp. 1035-41, 2014.

[9] J. Lucier and R. S. Weinstock, "Type 1 Diabetes," 3 March 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK507713/. [Accessed 24 February 2024].

[10] P. E. Cryer, "Hypoglycemia in type 1 diabetes mellitus," *Endocrinology and metabolism clinics of North America,* vol. 39, no. 3, p. 641–654, 2010.

[11] S. J. Gillespie, K. D. Kulkarni and A. E. Daly, "Using Carbohydrate Counting in Diabetes Clinical Practice," *Journal of the American Dietetic Association,* vol. 98, no. 8, pp. 897-905, 1998.

[12] P. Boulby, R. Moore, P. Gowland and R. C. Spiller, "Fat delays emptying but increases forward and backward antral flow as assessed by flow-sensitive magnetic resonance imaging," *Neurogastroenterology and motility,* pp. 27-36, 1999.

[13] S. Laxminarayan, J. Reifman, S. S. Edwards, H. Wolpert and G. M. Steil, "Bolus Estimation—Rethinking the Effect of Meal Fat Content," *Diabetes technology & therapeutics,* vol. 17, no. 12, pp. 860-866, 2015.

[14] Paterson, M., et al., "The Role of Dietary Protein and Fat in Glycaemic Control in Type 1 Diabetes: Implications for Intensive Diabetes Management," *Current diabetes reports,* vol. 15, no. 9, p. 61, 2015.

[15] A. Wolpert, A. Atakov-Castillo, A. Smith and M. Steil, "Dietary fat acutely increases glucose concentrations and insulin requirements in patients with type 1 diabetes: implications for carbohydrate-based bolus dose calculation and intensive diabetes management," *Diabetes care,* vol. 36, no. 4, pp. 810-816, 2013.

[16] Smith T. A., Smart C. E., Fuery M. E. J., et al, "In children and young people with type 1 diabetes using pump therapy, an additional 40% of the insulin dose for a high-fat, high-protein breakfast improves post- prandial glycaemic excursions: a cross-over trial," *Diabet Med,* vol. 38, no. 7, 2021.

[17] S. Annan, L. Higgins, E. Jelleryd, T. Hannon, S. Rose, S. Salis, J. Baptista, P. Chinchilla and M. Marcovecchio, "ISPAD Clinical Practice Consensus Guidelines 2022: Nutritional management in children and adolescents with diabetes," *Pediatric diabetes,* vol. 23, no. 8, pp. 1297-1321, 2022.

[18] L. T. Meade and W. E. Rushton, "Accuracy of Carbohydrate Counting in Adults," *Clinical diabetes : a publication of the American Diabetes Association,* vol. 34, no. 3, pp. 142-147, 2016.

[19] Gillingham, M. B., et al, "Assessing Mealtime Macronutrient Content: Patient Perceptions Versus Expert Analyses via a Novel Phone App," *Diabetes technology & therapeutics,* vol. 23, no. 2, pp. 85-94, 2021.

[20] N. Tyler and P. G. Jacobs, "Artificial Intelligence in Decision Support Systems for Type 1 Diabetes," *Sensors,* vol. 20, no. 11, p. 3214, 2020.

[21] R. Nimri, M. Phillip and B. Kovatchev, "Decision Support Systems and Closed-Loop," *Diabetes technology & therapeutics,* vol. 24, no. S1, pp. S-58, 2022.

[22] Tyler, N. S., et al., "An artificial intelligence decision support system for the management of type 1 diabetes," *Nature metabolism,* vol. 2, no. 7, pp. 612-619, 2020.

[23] Unsworth, R, et al., "Safety and Efficacy of an Adaptive Bolus Calculator for Type 1 Diabetes: A Randomized Controlled Crossover Study," *Diabetes Technology & Therapeutics,* vol. 25, no. 6, pp. 414-425, 2023.

[24] Castle, J. R., et al., "Assessment of a Decision Support System for Adults with Type 1 Diabetes on Multiple Daily Insulin Injections," *Diabetes Technology & Therapeutics,* vol. 24, no. 12, pp. 892-897, 2022.

[25] Jacobs, P. G., et al, "Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls and opportunities," *IEEE reviews in biomedical engineering,* 2023.

[26] Zeevi, D., et al., "Personalized nutrition by prediction of glycemic responses," *Cell,* vol. 163.5, pp. 1079-1094, 2015.

[27] American Diabetes Association, "Postprandial Blood Glucose," *Diabetes Care,* vol. 24, no. 4, p. 775–778, 2001.

[28] M. Metwally, T. O. Cheung, R. Smith and K. Bell, "Insulin pump dosing strategies for meals varying in fat, protein or glycaemic index or grazing-style meals in type 1 diabetes: A systematic review," *Diabetes Research and Clinical Practice,* vol. 172, 2021.

[29] Castle, J. R., Espinoza, A.Z., Tyler, N.S., et al., "771-P: Acceptance of Decision Support Recommendations Improves Time in Range for People Living with Type 1 Diabetes on Multiple Daily Injections.," *Diabetes,* vol. 71, no. Supplement_1, p. 771–P, 2022.

[30] M. Munoz-Organero, "Deep Physiological Model for Blood Glucose Prediction in T1DM Patients," *Sensors(Basel, Switzerland),* vol. 20, no. 14, p. 3896, 2020.

[31] Prendin, F., Del Favero, S., Vettoretti, M., et al., "orecasting of Glucose Levels and Hypoglycemic Events: Head-to-Head Comparison of Linear and Nonlinear Data-Driven Algorithms Based on Continuous Glucose Monitoring Data Only," *Sensors,* vol. 21, no. 5, p. 1647, 2021.

[32] Syafrudin, M., Alfian, G.,Fitriyani, N.L., et al., "A Personalized Blood Glucose Prediction Model Using Random Forest Regression," *2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems,* pp. 295-299, 2022.

[33] A. Zale and N. Mathioudakis, "Machine Learning Models for Inpatient Glucose Prediction," *Current Diabetes Reports,* vol. 22, p. 353–364, 2022.

[34] 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. 16, no. 1, pp. 7-18, 2022.

[35] B. Huard and G. Kirkham, "Mathematical modelling of glucose dynamics," *Current Opinion in Endocrine and Metabolic Research,* 2022.

[36] Resalat, N., Youssef, J. E., Tyler, N., et al., "A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model," *PLOS ONE,* vol. 14, no. 7, p. e0217301, 2019.

[37] Prendin, F., Pavan, J., Cappon, G., et al, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP," *Scientific reports,* vol. 13, no. 1, p. 16865, 2023.

[38] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Proc. 31st Int. Conf. Neural Inf. Process. Syst.,* pp. 4768-4777, 2017.

[39] M. T. Ribeiro et al., ""Why should i trust you?“: Explaining the predictions of any classifier," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining,* pp. 1135-1144, 2016.

[40] L. Breiman and A. Cutler, "Technical report: Random forests manual v4: UC Berkeley," 2003. [Online]. Available: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf. [Accessed 24 February 2024].

[41] H. Haibo and M. Yunqian, "Imbalanced learning: foundations, algorithms, and applications," *Wiley-IEEE Press,* vol. 1, no. 27, p. 12, 2013.

[42] R. Ribeiro and N. Moniz, "Imbalanced regression and extreme value prediction," *Mach learn,* vol. 109, pp. 1803-1835, 2020.

[43] P. Branco, L. Torgo and R. P. Ribeiro, "Pre-processing approaches for imbalanced distributions in regression," *Neurocomputing,* vol. 343, pp. 76-99, 2019.

[44] Y. Yang, K. Zha, Y. Chen, H. Wang and D. Katabi, "Delving into deep imbalanced regression," *International Conference on Machine Learning,* pp. 18842-18851, 2021.

[45] L. Torgo, R. P. Ribeiro, B. Pfahringer and P. Branco, "Smote for regression," *In Portuguese conference on artificial intelligence,* pp. 378-389, 2013.

[46] P. Branco, L. Torgo and R. Ribeiro, "SMOGN: A Pre-Processing Approach for Imbalanced Regression," *Proceedings of Machine Learning Research,* vol. 74, pp. 36-50, 2017.

[47] R. Alejo, J. M. Sotoca, V. García and R. M. Valdovinos, "Back propagation with balanced MSE cost function and nearest neighbor editing for handling class overlap and class imbalance," *In International Work-Conference on Artificial Neural Networks,* pp. 199-206, 2011.

[48] R. Sergazinov, M. Armandpour and I. Gaynanova, "Gluformer: transformer-based personalized glucose forecasting with uncertainty quantification," *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 1-5, 2023.

[49] Liu, K., Li, L., Ma, Y., et al., " Machine Learning Models for Blood Glucose Level Prediction in Patients With Diabetes Mellitus: Systematic Review and Network Meta-Analysis," *JMIR Med Inform,* vol. 11, p. e47833 , 2023.

[50] Dave, D., DeSalvo, D. J., Haridas, B., et al., "Feature-Based Machine Learning Model for Real-Time Hypoglycemia Prediction," *J Diabetes Sci Technol,* vol. 15, no. 4, pp. 842-855, 2021.

[51] Mendes-Soares, H., et al., "Assessment of a Personalized Approach to Predicting Postprandial Glycemic Responses to Food Among Individuals Without Diabetes," *JAMA network open,* vol. 2, no. 2, 2019.

[52] Pustozerov, E. A., et al., "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus," *IEEE Access,* vol. 8, pp. 219308-219321, 2020.

[53] Pustozerov, E. A., et al., "The role of glycemic index and glycemic load in the development of real-time postprandial glycemic response prediction models for patients with gestational diabetes," *Nutrients,* vol. 12, no. 2, p. 302, 2020.

[54] Annuzzi, G., Apicella, A., Arpaia, P., Bozzetto, L., Criscuolo, S., et al., "Exploring Nutritional Influence on Blood Glucose Forecasting for Type 1 Diabetes Using Explainable AI," *IEEE journal of biomedical and health informatics,* 2023.

[55] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems,* vol. 30, 2017.

[56] Riddell, M. C., et al., "Examining the Acute Glycemic Effects of Different Types of Structured Exercise Sessions in Type 1 Diabetes in a Real-World Setting: The Type 1 Diabetes and Exercise Initiative (T1DEXI)," *Diabetes care,* vol. 46, no. 4, pp. 704-713, 2023.

[57] Xu, P., Ji, X., Li, M. et al., "Small data machine learning in materials science," *npj Comput Mater,* vol. 9, no. 42, 2023.

[58] Mosquera-Lopez, C., et al., "Enabling fully automated insulin delivery through meal detection and size estimation using Artificial Intelligence," *NPJ digital medicine,* vol. 6, no. 1, p. 39, 2023.

[59] Martin, C. K., Han, H., Coulon, S. M, et al., "British Journal of Nutrition," *A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method,* vol. 101, no. 3, p. 446–456, 2008.

[60] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[61] H. Zhang, D. Nettleton and Z. Zhu, "Regression-enhanced random forests," *arXiv preprint,* p. arXiv:1904.10416, 2019.

[62] C. Chen, A. Liaw and L. Breiman, "Using Random Forest to Learn Imbalanced Data," 2004.

[63] T. Vink, "Adjusting the bootstrap in Random Forest," 15 June 2022. [Online]. Available: https://timvink.nl/blog/post-balanced-trees/#balancing-class-weight. [Accessed 5 February 2024].

[64] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The annals of statistics,* vol. 19, no. 1, pp. 1-67, 1991.

[65] J. Rudy, "py-earth: A Python implementation of Jerome Friedman’s Multivariate Adaptive Regression Splines," 2013.

[66] L. Biewald, "Experiment Tracking with Weights and Biases," 2020. [Online]. Available: https://www.wandb.com/.

[67] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[68] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software,* vol. 6, no. 60, p. 3021, 2021.

[69] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering,* vol. 9, no. 3, pp. 90-95, 2007.

[70] Charlier, F., Weber, M., Izak, D. et al., "Statannotations (v0.6)," Zenodo, 2022.

[71] Parkes, J. L., 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.