
EVALUATION OF MULTIPLE MACHINE LEARNING ALGORITHMS AND GLYCEMIC VARIABILITY INDICES FOR PREDICTIONS OF GLUCOSE LEVELS IN TYPE 2 DIABETES MELLITUS BASED ON CONTINUOUS GLUCOSE MONITORING AND ACTIVITY DATA

Annemijn Hoff (s2538830)

Bachelor thesis committee: dr. Kilian Kappert, dr. Maryam Amir Haeri, prof. dr. Goos Laverman

University of Twente, Bachelor Biomedical Technology

Biomedical Signals and Systems

July 12, 2023

Glycemic events and glycemic variability play a key role in the onset of complications in Diabetes Mellitus type 2 (DM2). Predicting these can be an important tool in self-management, and so preventing complications relating to DM2. Using Machine Learning, predictive models can be made. Continuous glucose monitoring (CGM) and activity data from the DIALECT (Diabetes and Lifestyle Cohort Twente) dataset were used. Features based on time, CGM, activity data, and clinical information were extracted. Various glycemic variability features were included, such as Mean Amplitude of Glycemic Excursion (MAGE), J-index, Time in Range (TIR), and High and Low Blood Glucose Index (HBGI and LBGI). Models for predicting the next glucose level were made using the following machine learning algorithms: Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and XGBoost (Extreme Gradient Boosting). These models were compared with regard to prediction accuracy with the following metrics: Clarke Error Grid Analysis, root mean square error (RMSE), and mean absolute error (MAE). First, personal models were made based on data from three patients with different variability. Variability was assessed by calculating the average daily risk range (ADRR). Second, population models were made based on the full data. All models were compared with a baseline: assuming the previous glucose value as the current value. Results show LR performed best with personal models and RF performed best with population models. Feature importance showed that the most important feature categories in most models were *previous glucose measurements* and *basic glucose calculations*. Glycemic variability features were not very high ranked, except in the personal LR models, and SD features in the population models.

Glykemische gebeurtenissen en glykemische variabiliteit zijn een belangrijk onderdeel in het ontstaan van complicaties bij patienten met Diabetes Mellitus type 2 (DM2). Deze zaken voorspellen kan helpen bij individueel diabetesmanagement, en kan zo complicaties voorkomen. Met machine learning kunnen voorspellende modellen worden gemaakt. Hiervoor zijn Continue glucose monitoring (CGM) en activiteit data van de DIALECT (Diabetes and Lifestyle Cohort Twente) dataset gebruikt. Verschillende features gebaseerd op tijd, CGM, activiteit data en klinische informatie zijn geëxtraheerd. Een aantal glykemische variabiliteit indexen zijn geïncludeerd, zoals Mean Amplitude of Glycemic Excursion (MAGE), J-index, Time in Range (TIR), en High en Low Blood Glucose Index (HBGI and LBGI). Voorspellingsmodellen zijn gemaakt met de volgende machine learning algoritmes: Linear Regression (LR), Decision Tree (DT), Random Forest (DT), en XGBoost (Extreme Gradient Boosting). Deze modellen werden vergeleken aan de hand van de volgende maatstaven: Clarke Error Grid analyse, root mean square error (RMSE), en mean absolute error (MAE). Er zijn persoonlijke modellen gemaakt op basis van data van drie patienten met verschillende variabiliteit. Deze variabiliteit is bepaald aan de hand van de average daily risk range (ADRR). Ook zijn er populatiemodellen gemaakt op basis van de volledige data. LR modellen bleken het beste bij de persoonlijke modellen, en RF was het beste bij de populatiemodellen. Daarnaast werd bepaald hoe belangrijk de features waren bij het maken van de modellen. De belangrijkste feature categorieën waren *vorige glucose metingen* en *simpele glucose berekeningen*. Glykemische variabiliteit indices bleken minder belangrijk, behalve bij de persoonlijke LR modellen, en SD bij populatiemodellen.

1 Introduction

Diabetes Mellitus (DM) is a chronic, metabolic condition characterized by inadequate control of blood glucose levels that can be caused by genetic factors and lifestyle, or due to other endocrinopathies. [1] There are many subtypes of diabetes, with types 1 and 2 being the most common types. Type 1 (DM1) is a result of defective insulin secretion and type 2 (DM2) is due to resistance to insulin action. [2] It is a very common condition, with a prevalence of 1 in 16 in the Netherlands. Out of these patients, approximately 9 % have type 1. [3] Complications are classified as microvascular or macrovascular. [2] Microvascular complications include neuropathy (damage to the nervous system), nephropathy (damage to the renal system), and retinopathy (damage to the retina). Macrovascular complications include cardiovascular disease, peripheral vascular disease, and stroke. Peripheral vascular disease can lead to injuries that do not heal which can lead to amputation, usually in the lower extremities. Cardiovascular disease, renal disease, and stroke are common causes of morbidity and mortality in DM patients. [1, 2] Treatment of DM2 includes lifestyle interventions and glucose management using insulin, metformin, and sulfonylureas. Regular screenings are necessary to manage and prevent complications. [1]

Glycemic events, also defined as hypo- (blood glucose < 3.9 mmol/L) and hyperglycemia (> 10 mmol/L) are very important risk factors for DM2 complications. [2, 4] Glycemic (or glucose) variability, which is defined as the fluctuation of blood glucose levels over time, has more recently been investigated and is recognized as a component of glucose control. [5] Intensive glucose control has been shown to decrease the risk of microvascular complications in DM2 patients, and high glucose variability has been associated with coronary artery disease. [6] Blood glucose dynamics are affected by many factors, such as carbohydrate intake, insulin, stress, physical activity, illness, alcohol, and smoking. [7] HbA1c (glycosylated hemoglobin) has always been the golden standard for the assessment of long-term glycemic control. [8] This metric provides an average of glucose levels for a period of 2-3 months. [9] However, HbA1c is not sensitive to rapid variations in glucose levels [10] and is not able

to predict severe hypoglycemic events accurately. [8, 9] Glycemic excursions (hypo- or hyperglycemic) and variability are a factor in short- and long-term complications of diabetes that are not predicted by HbA1c. [11] For example, glycemic variability has been suggested to have a role in diabetic retinopathy, independent of HbA1c. This is specifically the case for hyperglycemic risk assessment and fluctuation indices of glycemic variability. [11] Other indices of variability have been shown to be associated with cardiovascular, neuropathic, or retinopathic complications and lower quality of life. [10]

Glycemic variability can be approached as the fluctuation of glucose levels, the risk of glycemic events, or as low glucose control. Many different glycemic variability, risk assessment, and glucose control parameters have been proposed.

For glycemic variability, common parameters used are the SD (Standard Deviation of the mean), CV (Coefficient of Variation), MAGE (Mean Amplitude of Glucose Excursion), CONGA (Continuous Overall Net Glycemic Action), MODD (Mean Of Daily Differences), IQR (Inter-Quartile Range) and TIR (Time In Range).

The SD is a simple calculation and can give a rough estimate of variability. However, SD and SD-related indices do not give a different weight to minor or major swings. For DM2, an adequate SD is the mean blood glucose divided by 3. [12]

CV is the standard deviation divided by the mean of glucose concentration. Blood glucose profiles are considered highly variable if $CV > 36\%$. [13] CV has been used in clinical practice for the evaluation of glycemic variability. [14]

MAGE is the mean of glucose values exceeding one SD from the 24-hour mean and is a commonly used within-day variability indicator that can be calculated with CGM or self-measured blood glucose data as long as there are at least 7 measurements per day. [15]

CONGA is a measure of within-day variability and is calculated by taking the SD of the differences between the current observation and the observation n hours ago. n is usually 1, 2, or 4 hours. No normal ranges are available. [12]

MODD is a between-day index of variability and is

calculated by taking the mean of the difference between glucose values taken on two consecutive days. [16]

IQR is the difference between the 75th and 25th percentiles and has been used in clinical practice to evaluate glycemic variability. [14]

TIR indicates the time spent in the target range (between 3.9 and 10 mmol/L), and though it is not strictly a variability index, it is correlated and has been shown to be associated with diabetic complications. [10] Related indices are TAR (Time Above Range) and TBR (Time Below Range).

For risk assessment, commonly used parameters are LBGI (Low Blood Glucose Index), HBGI (High Blood Glucose Index), ADRR (Average Daily Risk Range)

LBGI and HBGI assess the risk of hypo- and hyperglycemic excursions respectively, relative to a period of time. [8, 12] These values can be based on self-measured blood glucose data or CGM data. An overall Blood Glucose Risk Index (BGRI) or Index of Glycemic Control (IGC) can be calculated from these values by adding them together, which indicates the risk of extreme glycemic values. [12] LBGI has repeatedly been shown to be an excellent predictor of severe hypoglycemia and HBGI is closely related to HbA1c and risk for hyperglycemia. [17]

ADRR is a measure of extreme glucose values and an indicator for assessing the daily risk of hypo- and hyperglycemia. [10, 12, 18] It is calculated using two to four weeks of self-measured blood glucose data, and requires at least 3 measurements a day. [12, 18] However, this measure is more sensitive to extreme variability than to within-range variability.

For glucose control, commonly used parameters are J-index and GMI (Glucose Management Indicator).

The J-index is an indicator of glucose control and is calculated using the mean and SD. Ideal control is a value between 10 and 20, good control is between 20 and 30, poor control is between 30 and 40, and inadequate control is over 40. [12]

GMI is an estimate of HbA1c calculated with at least 10 days of CGM data. Due to the way it is measured, there tend to be slight differences between the values. [19]

Measurements of blood glucose are usually done by

finger-prick test. For DM1 or unregulated DM2 continuous glucose monitoring (CGM) can be used. The most common technique relies on measuring glucose levels in interstitial fluid in a subcutaneous system. [5] This way, a daily glucose profile can be obtained and give a precise overview of glycemic fluctuations.

Prediction of blood glucose levels based on CGM data using machine learning or time-series approaches have been developed over the last few years. Many model types based on CGM or blood glucose data have been researched for predictions of glucose levels in DM1 and DM2 patients. [20–22] Common data-driven model types are neural networks such as artificial (ANN) or deep neural networks (DNN) or long short-term memory (LSTM) [13, 20, 23–26], decision trees (DT) and ensemble methods such as random forests (RF) and XGBoost (Extreme Gradient Boosting) [20, 23–29], support vector machines (SVM) or support vector regression (SVR) [13, 20, 23–26, 29, 30], (linear) regression (LR) [20, 24, 28, 31], and time-series-based approaches such as Auto Regressive Integrated Moving Average (ARIMA). [20, 23, 25–27, 29, 30]

Predictions of future glucose levels are useful for diabetes self-management since action can be taken based on future values instead of current values, which decreases the risk of hyper- and hypoglycemic events. [13, 32]

Population models of DM2 patients are limited due to inter- and intra-patient variability. Each individual DM2 patient has distinct glucose dynamics, which are not captured by population-based models. [33] Personal models are based on information from only one patient, which is less data.

This study focuses on developing personal and population models with machine learning algorithms LR, DT, RF, and XGBoost, comparing their performance and evaluating the importance of glycemic variability indices as features.

2 Methods

2.1 Database

CGM and activity data originate from “Diabetes and Lifestyle Cohort Twente” (DIALECT), an observational

cohort study performed in the Ziekenhuis Groep Twente (ZGT) hospital (Almelo and Hengelo, the Netherlands). The goal of DIALECT is to investigate the effects of lifestyle and dietary habits on outcomes in DM2 patients in specialist care. [34] Measurements were done under free-living conditions for two weeks, and data was collected using Abbott FreeStyle Libre (for CGM data, measurements every 15 minutes), a Fitbit wristband (for activity data, measurements every minute), and a food diary. Participants were not able to view their CGM or activity data. Written informed consent was obtained from all patients before participation. The study was approved by the local institutional review boards (Medisch Ethische Toetsingscommissie reg. nos. NL57219.044.16 and 1009.68020) and is registered in the Netherlands Trial Register (NTR5855).

2.2 Study population

Patients were selected according to the inclusion criteria below. Figure 1 shows a flowchart of the effects on the number of subjects due to steps taken in the exclusion of patients.

1. Must have CGM and activity data.
2. CGM and activity data must overlap.
3. The overlap interval is at least 7 days.
4. Gaps between CGM measurements must be no larger than 4 hours.

Certain subjects ($n = 36$) required parts of data with large gaps to be removed before they were eligible for inclusion.

Personal models are made for subjects of different glycemic variabilities to evaluate machine learning algorithm performance on these smaller datasets and to evaluate whether higher variability impacts prediction accuracy.

For personal models, three subjects were selected based on their ADRR values calculated over their full CGM data. ADRR is a measure of the risk of glycemic events. An ADRR below 20 is considered a low risk for glycemic events, between 20 and 40 is considered a moderate risk, and above 40 is a high risk. [12]

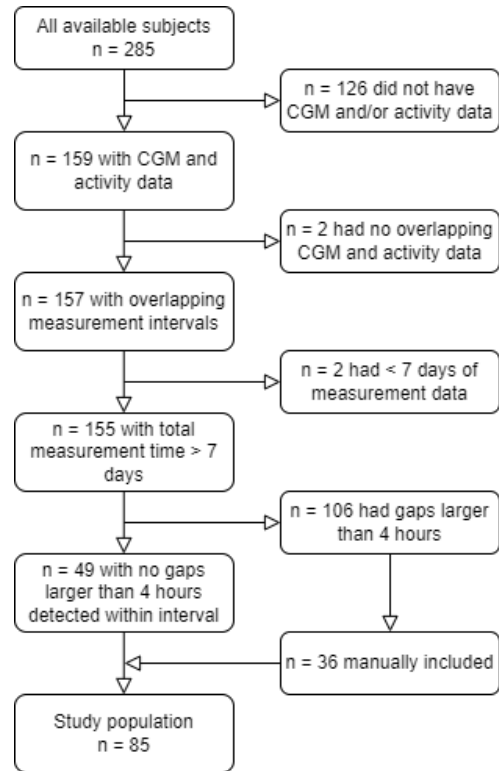


Figure 1: Flowchart documenting inclusion of study population.

Subjects with different variabilities were selected by taking the subject with the lowest, median, and highest ADRR, as long as each subject had a comparable amount of data.

2.3 Machine Learning Algorithms

Multiple ML algorithms were applied to the features selected, namely Linear Regression (LR), Decision Tree (DT) Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

2.3.1 Linear Regression

LR is an algorithm that aims to learn a linear model to predict the target variable. It determines the best fit by using least-squares optimization. [35]

2.3.2 Decision Tree

DT is a supervised machine learning method, that can be used for classification or regression. A DT model is built with a tree-like structure with nodes. Each node is a test of an attribute, and each branch descending from the node corresponds to a value for the attribute. Leaf nodes represent a decision that is made based on the input variables. [36]

2.3.3 Random Forest

RF is an ensemble technique that can be used for classification or regression. RF uses bagging (bootstrap aggregation), which means many samples are taken from the original dataset and fitting models to aggregates of these samples. This model fits multiple DT in parallel and uses majority voting in classification or averages in regression to determine the outcome. [35]

2.3.4 XGBoost








XGBoost is a tree-based ensemble technique that uses gradient boosting. [37] Boosting is a technique that gives weights to each training based on the error. This way, the model can learn from errors from previous iterations and become more accurate during training. [35] Gradient boosting is an optimization, where the model is trained in an additive manner, so new DTs are added to the model based on the lowest error. [38] XGBoost incorporates regularization to prevent overfitting, can handle missing data, and uses hardware more efficiently. [37, 38]

2.4 Feature extraction

Since data-driven machine learning algorithms are used, features were extracted so more data is provided to the algorithms.

66 features were extracted from CGM and activity data, and from clinical patient information. These are shown in table 6 in section A. Features were based on literature [23, 24, 30] and expert advice.

Table 1: Table of categories of features used, color-coded.

Color	Feature category	Amount
	Time	5
	Previous glucose measurements	8
	Basic glucose calculations	16
	Glycemic variability indices	12
	Baseline subject characteristics	13
	Medication	10
	Activity	2

The extracted features were further categorized into different groups, as can be seen in table 1.

For feature extraction for predictions of a target variable, it is important to keep in mind that calculations should be done on previous values, and not include the current value or future values. For example, taking the average of values for an hour needs to be considered as the previous hour, otherwise current or future values could be included in the calculations.

5 time-based features were extracted, consisting of the day of the week, weekend, hour, part of the day, and meteorological season.

41 CGM-based features were extracted. Out of these, 8 were previous glucose levels, 15 were basic glucose calculations, and 13 were glycemic variability indices.

Glucose values of 1-6 measurements ago and glucose values of 1 hour and 24 hours ago are included as features.

For basic glucose calculations, the mean, standard deviation, minimum, and maximum were calculated over a previous time interval. A rolling window was used for calculations for the past time interval, excluding the current glucose value. The mean and SD were calculated over different window sizes, including 1-6 hours and 24 hours. Shorter SD windows can give insight into more recent glycemic variations since larger windows are less sensitive to major swings in glucose values. These short SD windows are included in basic glucose calculations since most metrics are usually calculated over a longer period of time. The minimum and maximum glucose values of the previous 24 hours and the slope between the two previous measurements were also included as a feature. Though the slope is directly related to the previous glucose values, it is considered a basic glucose calculation in this study.

Glycemic variability indices (SD, IQR, LBGI, HBGI, CV, J-index, TIR (and others, such as time above/below range) and MAGE were calculated using a rolling window of 24 hours.

23 features from clinical information were extracted, including 13 baseline characteristics and 10 features regarding medication use. Baseline characteristics include age, sex, BMI, years since diagnosis, HbA1c, among others. Medication features include the use and dosage of metformin and sulfonylurea-derivatives, and of slow-

acting, fast-acting, and mix insulin.

2 activity-based features were extracted: steps taken in the past hour and steps taken in the past 24 hours, which were calculated using a rolling window.

ML algorithms can only take numerical inputs, so after extracting features, dummy variables were made of certain factorized features, such as part of day and season.

2.5 Feature Selection

To evaluate and select features for use in the ML algorithms, the Boruta algorithm will be used. Since most features are extracted based on similar information, a pairwise correlation matrix is expected to show a high correlation. For each feature category, a correlation matrix is made, and the features with a correlation over 90% were reported.

The Boruta algorithm is a wrapper around the RF algorithm. [39] It creates shadow variables of each variable, which is a shuffled version of the original variable. It runs an RF classifier and computes the importance of all variables. If a variable has significantly lower importance than the maximum importance of a shadow variable, it is deemed unimportant. This way, the Boruta algorithm will automatically determine the importance of the features used. [39]

2.6 Training and validation strategy

Random sampling with preset seeds is used to create testing (80 %) and training sets (20 %). 10-fold cross-validation is used on the training data. 10-fold cross-validation has often been used in other studies using data-driven machine-learning approaches to blood glucose dynamics. [21] The training data is split into 10 parts, and 9 out of those parts are used for training, while the last part is used to validate the model. Each time a different part is used for validation. Out of these 10 folds, the best model is the final output.

Model performance is influenced by the hyperparameter values. For each model, there are different hyperparameters involved that can be tuned. For each model type, a baseline was done by taking the caret package's default values for the hyperparameters. Next, a random

search is done of hyperparameters, of a long tune length. For each iteration, a model is made with random values for each hyperparameter, and the tune length determines the number of iterations. Depending on the performance of the model (measured by the Root Mean Square Error (RMSE) on the training set), a grid is made and the model is trained again using the grid specified for the hyperparameters. For the random search, the tune length is made shorter after determining a grid. This is due to the run time of the random search. Especially with RF and XGBoost, this process can take a long time and a lot of resources due to the many hyperparameter combinations. For these algorithms, a long tune length is done initially to find appropriate values to establish a possible grid.

To achieve more reliable results, each model was trained on 10 different data splits. This means that the models were trained 10 times, each time with a different randomized data split into training and testing sets. For each iteration, the best model per condition (so per model type and dataset) is taken and evaluated. Out of all these models, the best model overall per condition is chosen and further evaluated.

The exact tuning steps are documented in the appendix, see appendix B.

2.7 Evaluating model performance

RMSE, Median Average Error (MAE), and Clarke Error Grid Analysis were used to evaluate model performance. RMSE and MAE are common metrics to evaluate regression models, while Clarke Error Grid Analysis is often used to evaluate predictions of glucose levels. [21]

Clarke Error Grid Analysis was developed to evaluate the clinical accuracy of blood glucose monitoring systems, and it compares the predicted glucose value to the actual value and places it in a scatterplot as shown in figure 2. [40] Clarke Error Grid Analysis is also used to evaluate the clinical accuracies of glucose forecasting algorithms. [13] The zones indicate different accuracies of predicted glucose levels and represent the severity of clinical errors due to action or inaction on these predicted values. [13, 40]

A. Predicted values are within 20% of reference, or hy-

glycemic values (below 3.9 mmol/L). Clinically accurate, so clinically correct treatment decisions are made based on these values.

- B. Predicted values are over 20% of reference. Taking action based on these values would not be dangerous to the patient.
- C. Action on predicted values would cause overcorrection: actual glucose levels will fall below 3.9 mmol/L or rise above 10 mmol/L, causing a glycemic event.
- D. Predicted values are in the target range (between 3.9 and 10 mmol/L) but actual values are outside the target range, failure to treat would potentially be dangerous.
- E. Predicted values are opposite to reference value (hyperglycemic predicted values are actually hypoglycemic), and acting on these values would cause erroneous treatment, which is potentially dangerous.

Zones A and B are considered clinically acceptable, while zones C, D, and E are considered clinically significant errors. [40]

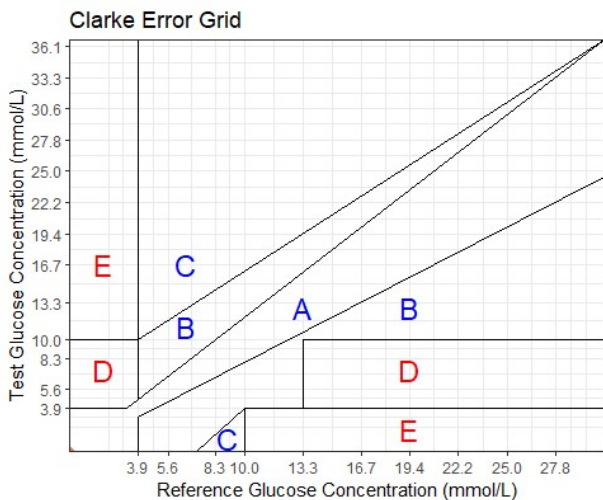


Figure 2: Empty Clarke Error Grid

Aside from these metrics, the performance is also compared to a baseline: the previous glucose value. For the baseline, the current glucose value is used as the reference, and the previous glucose value as the prediction.

This is to see whether assuming the previous value for the prediction gives accurate results and to compare these results to the model outcomes. For the baseline, the RMSE, MAE, and Clarke Error Grids are determined for each run. A Student’s t-test was performed to compare each metric of each model to the baseline metrics. To evaluate which model was best based on the average of all runs, a t-test was done based on the metrics to evaluate whether the best model was significantly better than the other models.

2.8 Feature importance

With the best models per condition, the feature importance is determined. For LR, it is determined by the absolute value of the t-statistic for each model parameter. [41] DT determines importance by attributing a reduction in the mean squared error to each variable for each split. This is then summed per variable. [41] With RF, the importance is determined in a similar way as Boruta, since Boruta is a RF wrapper. [39] First, the mean squared error is determined for the model, and then a variable is permuted. This is done by shuffling the values, and the differences are averaged and normalized by the error. [41] XGBoost feature importance is based on the total gain of each feature’s splits to the prediction of the model. [42]

2.9 Software

R Statistical Software (v4.3.0, R Core Team 2023) was used. For data processing, feature extraction and tidying, R packages ‘tidyverse’ (v2.0.0), ‘dplyr’ (v1.1.2), ‘data.table’ (v1.14.8), ‘lubridate’ (v1.9.2), ‘iglu’ (v3.4.2), and ‘runner’ (v0.4.3) were used. For feature selection, ‘ggplot2’ (v3.4.2), ‘ggcorrplot’ (0.1.4), ‘Boruta’ (v8.0.0), and ‘randomForest’ (v4.7-1.1) were used. For ML algorithm training ‘caret’ (v6.0-94) was used, using the methods ‘lm’ for LR, ‘rpart’ for DT, ‘ranger’ for RF, and ‘xgbTree’ for XGBoost. For evaluating models, ‘ggplot2’ (v3.4.2), ‘ega’ (v2.0.0), and ‘caret’ were used. The scripts used can be found in this repository: https://github.com/Annemijnh/Bachelor_Thesis

3 Results

3.1 Baseline characteristics

The baseline conditions of all included patients are presented in table 2.

Three patients with different variabilities were selected for personal model generation. For consideration of variability, ADRR was calculated for the full CGM data per subject. Subject 629 had an ADRR of 13.5, subject 688 had an ADRR of 63.2, and subject 708 had an ADRR of 29.9. There was a subject with a lower

ADRR, but there was considerably more data available for this subject.

The baseline characteristics of these specific patients can be seen in table 3.

3.2 Feature Selection

The Boruta Algorithm confirmed 78 attributes, meaning all features included were deemed important enough for inclusion. Based on this result, all features were used in training. Figure 3 shows the relative importance of the included features.

Table 2: Baseline characteristics of included patients

Baseline Characteristics	Value			
n	85			
Male	57 (67%)			
Current smoker	11 (13%)			
Past smoker	50 (59%)			
	mean	sd	min	max
Age	65	9	38	84
Years since diagnosis	15	9	0	39
Packyears	15	22	0	114
Alcohol units per month	14	19	0	81
Length (cm)	172	9	149	194
Weight (kg)	92	15	54	141
BMI (kg/m ²)	31	5	21	44
Waist circumference (cm)	111	12	83	139
Hip circumference (cm)	110	11	91	147
HbA1c (mmol/mol)	60.4	10.3	38	93
Fast insulin units (n = 45)	35	26	3	140
Mix insulin units (n = 7)	73	28	50	128
Slow insulin units (n = 45)	37	24	6	100
Metformin dosage (mg) (n = 68)	1500	700	500	3000
SU-derivative dosage (mg) (n = 26)	71	190	3	1000
Total measurement time (days)	11.4	2.4	7.0	14.3

Table 3: Baseline characteristics for personal model subjects

	Low-risk patient	Moderate-risk patient	High-risk patient
Age	45	71	65
Sex	Male	Male	Male
Smoking	Current, 15 pack years	Past, 15 pack years	Past, 34 pack years
Alcohol units per month	25	52	10
BMI (kg/m ²)	29.3	23.6	30.9
Years since diagnosis	0	15	10
HbA1c (mmol/mol)	38	56	77
Fast insulin units	0	24	36
Metformin dosage (mg)	500	2000	2000

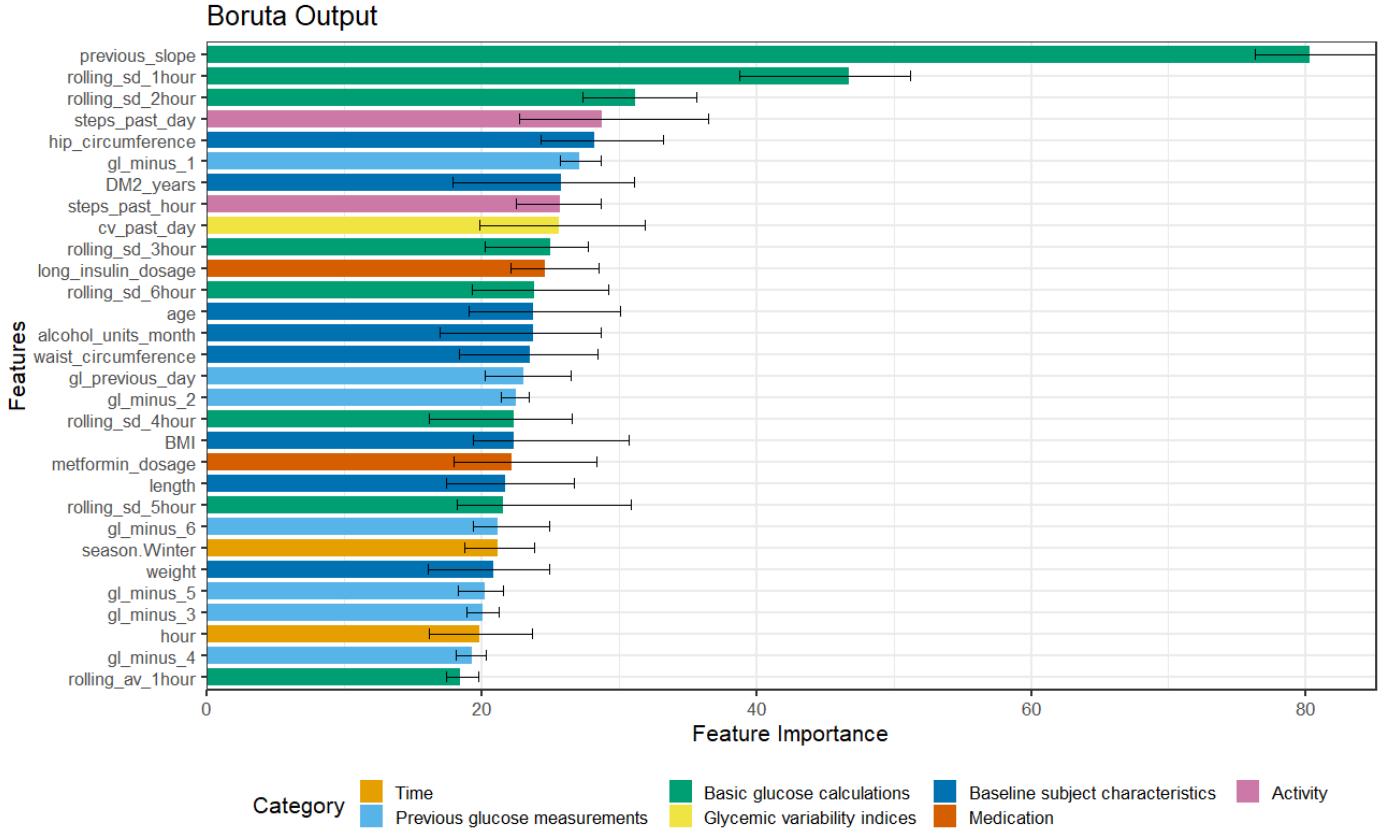


Figure 3: Boruta Algorithm results, only the top 30 features are shown.

The following features were correlated over 90%: the rolling mean of the 1-, 2-, 3-, 4-, 5-, and 24-hour windows, the rolling SD of the 5- and 6-hour windows, the previous glucose values of 1 to 6 measurements ago, the J-index, LBGI, and HBGI of the previous day, and the dosage of mix insulin. The correlation matrices per feature category are in appendix C, see figure 8. This shows that previous glucose measurement features, some of the basic glucose calculation features (especially the different rolling averages and SD windows), and some of the glycemic variability features are heavily inter-correlated.

3.3 Personal models

3.3.1 Evaluation of model performance

The results of the model evaluation can be seen in table 4. Table 4 shows the RMSE, MAE, and percentages of predicted values in the Clarke Error Grid zones per model per patient, and figure 9 in appendix D shows the Clarke Error Grids for the best model per model type per patient.

For the low-risk patient, on average LR performed

best, since the RMSE is lowest ($RMSE = 0.47 \pm 0.03$ mmol/L), and most predicted values fall in zone A or B (CEG zone AB = 99.95 ± 0.15 %). RF is very close to LR, and there are no statistical differences in RMSE means ($p = 0.61$) or MAE means ($p = 0.49$). The best final model, however, could be LR or RF. RF could be considered the best since the RMSE is the same as LR ($RMSE = 0.44$ mmol/L), but the MAE is lower (LR was 0.35 mmol/L, RF was 0.31 mmol/L). On the other hand, the predicted values by LR all fall in zones A and B, while 99.53 % of predicted values by RF fall in zones A and B. However, this difference is very small, so both models are considered to be the best for this patient. All models outperformed the baseline ($p < 0.05$ with respect to RMSE), so assuming the previous glucose value is less accurate than using a model.

For the medium-risk patient, LR performed best on average ($RMSE = 0.49 \pm 0.03$ mmol/L) and as a final model ($RMSE = 0.42$ mmol/L), due to the RMSE and MAE being the lowest, and all predicted values are in zone A or B.

Table 4: Evaluation of models, personal and population

Model type	Metrics	Low Risk Patient			Moderate Risk Patient			High Risk Patient			Full Population		
		Mean (SD)	Best model	Mean (SD)	Best model	Mean (SD)	Best model	Mean (SD)	Best model	Mean (SD)	Best model	Mean (SD)	Best model
LR	RMSE	0.47 (0.03) *	0.44	0.49 (0.03) *	0.42	0.53 (0.03) *	0.47	0.288 (0.003) *	0.283				
	MAE	0.35 (0.02) *	0.35	0.34 (0.02) *	0.30	0.38 (0.02) *	0.35	0.1367 (0.0008) *	0.1355				
	CEG zone AB %	99.95 (0.15)	100	100 (0)	100	99.7 (0.3)	100	99.986 (0.005) *	99.99				
	Tuning strategy	D8, G2	Default	D9, G1	Default	D6, G4	Default	D7, G3	Grid				
DT	RMSE	0.57 (0.04) *	0.51	0.62 (0.04) *	0.54	0.73 (0.04) *	0.67	0.315 (0.003) *	0.308				
	MAE	0.44 (0.03)	0.38	0.46 (0.03) *	0.40	0.54 (0.03) *	0.51	0.161 (0.002) *	0.159				
	CEG zone AB %	99.5 (0.6)	99.07	100 (0)	100	99.6 (0.3)	99.59	99.963 (0.008)	99.97				
	Tuning strategy	R5, G5	Grid	R7, G3	Grid	R7, G3	Grid	R1, G9	Grid				
RF	RMSE	0.48 (0.03) *	0.44	0.53 (0.04) *	0.43	0.58 (0.03) *	0.55	0.279 (0.003) *	0.274				
	MAE	0.36 (0.03) *	0.31	0.36 (0.02) *	0.30	0.41 (0.02) *	0.39	0.119 (0.001) *	0.119				
	CEG zone AB %	99.7 (0.3)	99.53	100 (0)	100	99.5 (0.4)	99.18	99.972 (0.011)	99.99				
	Tuning strategy	D1, R3, G6	Default	D3, R1, G6	Grid	D1, R5, G4	Random	D5, R4, G1	Default				
XGBoost	RMSE	0.50 (0.04) *	0.45	0.53 (0.05) *	0.44	0.60 (0.03) *	0.56	0.280 (0.003) *	0.274				
	MAE	0.37 (0.02) *	0.34	0.37 (0.02) *	0.33	0.45 (0.02) *	0.42	0.1283 (0.0009) *	0.1274				
	CEG zone AB %	99.7 (0.4)	100	100 (0)	100	99.7 (0.3)	100	99.973 (0.008)	99.99				
	Tuning strategy	D4, R5, G1	Grid	R5, G5	Random	D1, R4, G5	Grid	R10	Random				
Baseline	RMSE	0.62 (0.04)	0.57	0.58 (0.04)	0.49	0.68 (0.03)	0.63	0.386 (0.004)	0.381				
	MAE	0.45 (0.03)	0.40	0.40 (0.02)	0.35	0.49 (0.02)	0.46	0.179 (0.001)	0.178				
	CEG zone AB %	99.9 (0.2)	100	100 (0)	100	99.8 (0.2)	100	99.968 (0.006)	99.97				

RMSE = Root Mean Square Error, MAE = Mean Absolute Error, CEG = Clarke Error Grid, LR = Linear Regression, DT = Decision Tree, RF = Random Forest, and tuning strategy: D = Default, R = Random, G = Grid. The number after D, R, or G indicates the number of models that used this strategy. T-test of each model compared to baseline metrics: * : $p < 0.05$

Compared to RF and XGBoost, LR improved weakly significantly ($p < 0.1$). RF was very close with regard to the final model since only the RMSE is slightly higher (RMSE = 0.43 mmol/L). Only DT did not outperform the baseline and performed significantly worse ($p < 0.05$), the other models significantly outperformed the baseline when evaluating the RMSE and MAE. Since for each model, all predicted values are in zones A and B, there is no difference in performance.

For the high-risk patient, LR also performed best on average (RMSE = 0.53 ± 0.03) and as a final model (RMSE = 0.47), due to the RMSE and MAE being the lowest. This time, LR was clearly better as a final model, and there was a strongly significant difference when compared to the other models ($p < 0.01$). Again, DT did not outperform the baseline and did significantly worse regarding RMSE and MAE ($p < 0.05$).

The Clarke Error Grids are included in appendix D in figure 9. These show where the predicted values fall, and for most best models, the values are all in zone A or B. This can also be seen in table 4. The figure shows that DT has a larger spread of values, which is also reflected by the RMSE. Also, the values for the low-risk patient seem to be mostly in the euglycemic range (between 3.9 and 10 mmol/L), while the moderate-risk patient has more values in hypo- and hyperglycemic ranges. The high-risk patient also has more values in the higher range, even higher than the moderate-risk patient. Out of the clinically dangerous predicted values, all of them are in zone D, which are values that are actually outside of the target range but the model predicted them as inside the target range.

3.3.2 Evaluation of feature importance

In figure 4 the top 15 most important features are shown, categorized by the previously defined feature categories and by the variability per patient.

The LR feature importance shows a good representation of the different categories of features. Basic glucose calculations and previous glucose measurements are the most important feature categories. Especially the previous glucose value and the slope of the previous two values, since they are the most essential features. Of the other

previous glucose measurement features, only the glucose value for the previous day and previous hour were important. As for the basic glucose calculations, other than the previous slope, the average of two rolling windows (4 and 5 hours) and the minimum of the past day were relevant. Both activity features are in the top 15 and are most relevant to the moderate- and high-risk patients. For glycemic variability indices, CV, HBGI, and TIR are in the top 15, and these features seem most relevant for the moderate-risk patient. Interestingly, TIR is not important to the high-risk patient. Two time-based features were in the top 15, namely Wednesday and Thursday, and these seem to be most relevant for the low-risk patient.

For DT, only three categories are represented: basic glucose calculations, previous glucose measurements, and activity. The previous glucose value and the previous slope are the most important. The previous glucose measurements seem to be most important to the high-risk patient and consist of the previous 3 glucose measurements and the measurement from an hour ago. For the basic glucose calculations, many SD and mean windows were important. Both activity features were included, and are most relevant to the low-risk patient. No glycemic variability indices are included, however, SD does indicate a short-term variability.

With the feature importance of RF, one feature is most important to all three patients: the previous glucose measurement. 7 out of 8 features in the previous glucose measurements category are represented in the top 15. From the basic glucose calculations category, three rolling mean windows, the previous slope, and the SD for the previous hour are relevant. Both activity-based features are in the top 15, though from the 7th most important feature onward, there seem to be small differences in importance between the features. One time-based feature is shown: Sunday.

For XGBoost, as with RF, the most important feature is the previous glucose measurement. After the 5th or 6th top features, the differences in importance become much smaller. Within this top 5, the previous slope and average of two windows (1 and 2 hours) are included, as is the glucose value of 2 measurements ago. For the other features that are not in the top 5, more previous glucose

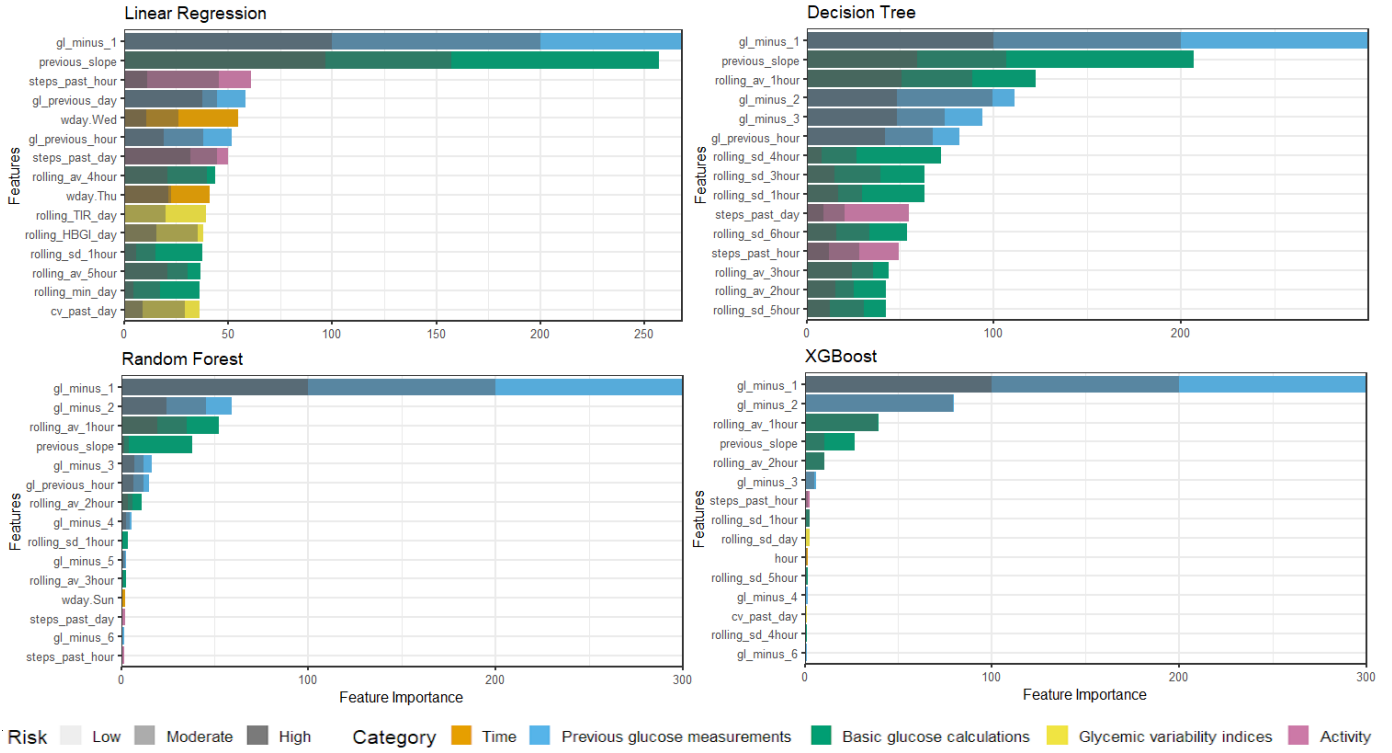


Figure 4: Feature importance of the personal models

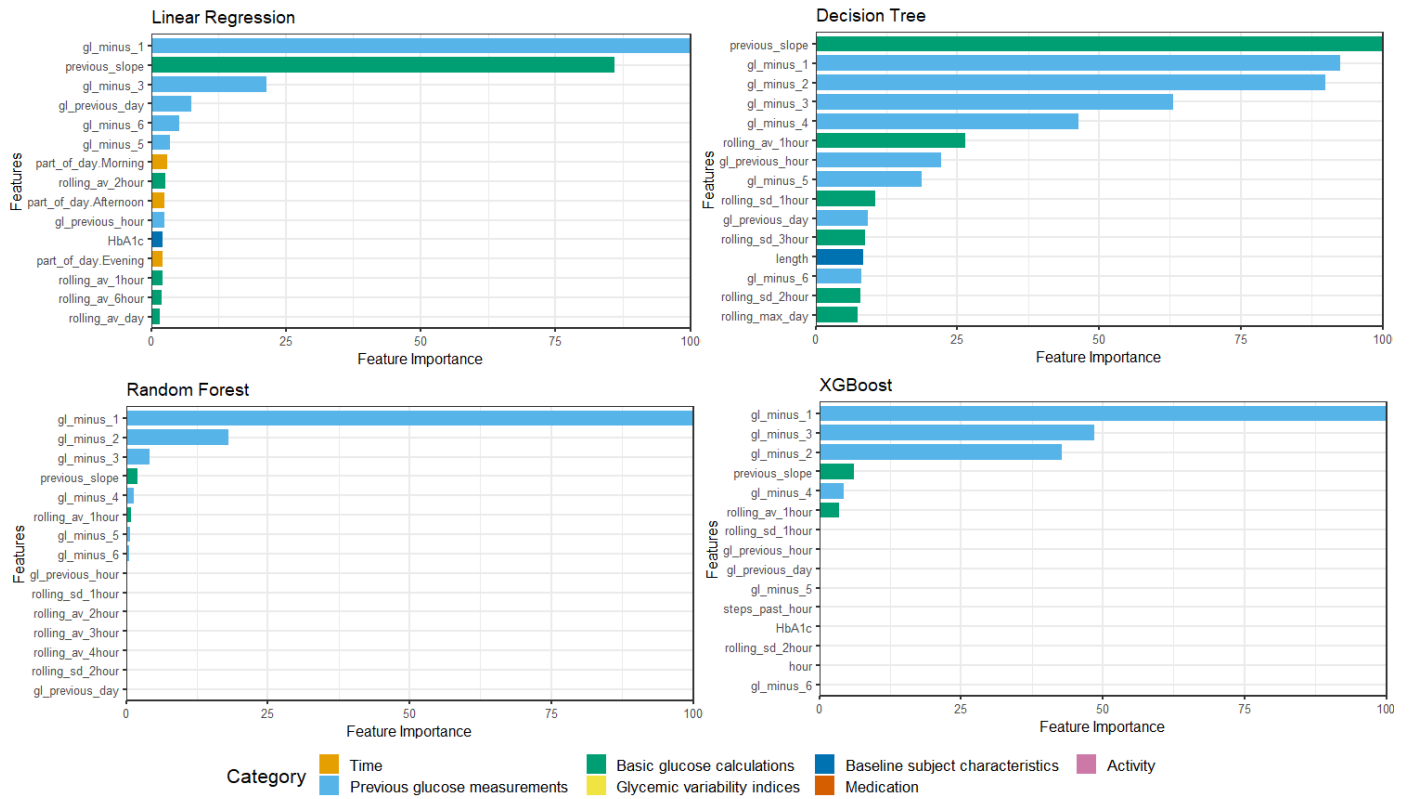


Figure 5: Feature importance of the population models

measurements are included, the rolling SD of 4 different windows (1, 4, 5, and 24 hours), one activity-based feature, one time-based feature, and, next to SD of 24 hours, CV as a glycemic variability index.

No baseline subject characteristics or medication features are included, this is due to the personal models deeming these features as constants, so they are not relevant when making personal models.

3.4 Population models

3.4.1 Evaluation of model performance

The same machine learning algorithms were used to create population models for the entire included patient population. The results of the model evaluation can be seen in table 4 and figure 6.

All models outperformed the baseline ($p < 0.05$) when considering RMSE and MAE. For the Clarke Error Grid zones, only LR was significant ($p < 0.05$).

On average, RF performed best when considering RMSE (0.279 ± 0.003 mmol/L) and MAE (0.119 ± 0.001 mmol/L). XGBoost and RF had the best final models when considering RMSE (RMSE = 0.274 mmol/L), but RF had a lower MAE (MAE = 0.119 mmol/L for RF, MAE = 0.1274 mmol/L for XGBoost). On average, RF outperformed XGBoost only when considering MAE ($p < 0.01$).

The Clarke Error Grids in figure 6 show that for all best models, predicted values mostly fall in zone A and B, as was seen in table 4. Though some models had significantly lower RMSE and MAE, all models are shown to predict mostly clinically safe values. Again, the only clinically unsafe zone that predicted values are in is zone D.

3.4.2 Evaluation of feature importance

Figure 5 shows a bar graph of the top 15 most important features per model type.

For LR, the previous glucose measurements and basic glucose calculations are the most important categories, and the previous glucose value and the previous slope are the most important features. Within the basic glucose calculations, only the rolling mean of 4 different windows seem to be in the top 15. Three time-based features are

included: all part-of-the-day features. One baseline subject characteristic is included: HbA1c. As discussed in the introduction, HbA1c is an indicator of the glycemic control of the previous months.

With the DT model, only three categories are important: previous glucose measurements (all features are included in the top 15), basic glucose calculations, and one baseline subject characteristic: length. From the basic glucose calculations, the previous slope, rolling average of the 1-hour window, SD of 1-, 2-, and 3-hour windows, and the maximum of the past day are relevant. Although these smaller SD windows are included in the basic glucose calculations category, they are an indication of short-term glycemic variability.

RF only had some features that are clear in the bar graph. The previous glucose values seem to be the most important, especially 1 and 2 measurements ago. In total, all features in the category of previous glucose measurements are included, and some basic glucose calculation features, such as the previous slope, the mean of the 1-, 2-, 3-, and 4-hour windows, and the SD of the 1- and 2-hour windows. Out of these, SD can indicate short-term glycemic variability, but these features barely show up in the bar graph, so are not as important as other features.

XGBoost only had 6 features that clearly show up in the bar graph, 4 of these are previous glucose measurements and the other two are the previous slope and the mean of the previous hour. Out of the other features, all other features from the previous glucose measurement category are present, and the SD of 1- and 2-hour windows, one activity-based feature (steps of the past day), one time-based feature (hour), and one subject-based feature (HbA1c).

Compared to the personal models, it seems certain features are less important. Activity features only showed up in the top 15 of the feature importance in XGBoost, and only steps of the past hour was relevant. Even then, this feature was not very important compared to the top 6 of most important features in XGBoost. SD and glycemic variability features seemed more important in personal models. DT had 3 different windows of SD in the top 15, though the personal models had more SD windows.

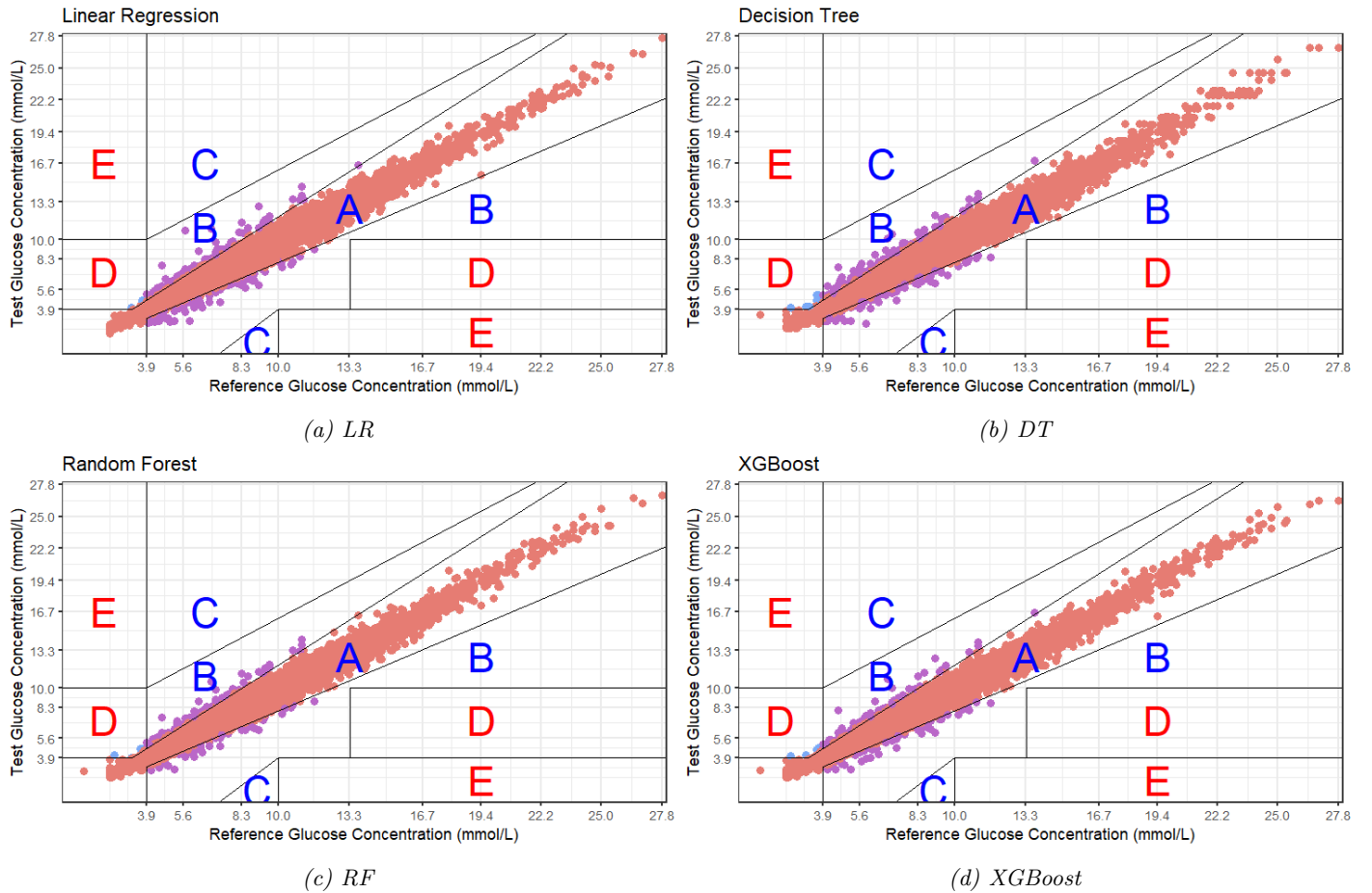


Figure 6: Clarke Error Grids of the population models

4 Discussion

4.1 Model performance

This study compared different ML algorithms for personal and population models and the importance of glycemic variability features in modeling. Earlier studies have compared different ML algorithms, including LR, DT, RF, and XGBoost, and personal and population models. This is the first study to research the importance of glycemic variability indices as features in ML models for personal and population models.

For the personal and population models, almost all models outperformed the baseline, which means using machine learning is an improvement in accurate predictions compared to assuming the previous glucose value. Although the error of the models compared to the baseline was lower, the performance of the baseline (taking the previous glucose value as a prediction) was clinically good, since the Clarke Error Grids show that the predicted values mostly fall in zone A and B, which means

these values are clinically accurate.

Out of the personal models in general, LR performed best, while out of the population models, RF performed best. It is surprising that XGBoost and RF did not perform as well as expected on the personal datasets. Since these algorithms did perform better with a larger dataset, namely the population set, a possible explanation is the amount of data these algorithms need to perform well. XGBoost and RF are able to handle small datasets well since they are ensemble methods, although the number of features included (55, not including the baseline characteristics and medication) increases the dimensionality of the dataset. This means either the number of observations should be increased or the number of features should be decreased to achieve higher accuracy for these algorithms on the smaller datasets. [43]

In general, the population models are more accurate since the RMSE and MAE were lower than in the personal models. Due to glycemic variability indices being included as features, it could be assumed that the individual variability of each subject is taken into account.

This means the population models would work well, regardless of individual variability, possibly better due to more data being used in training. Another possibility is due to the amount of data. Since data-driven machine-learning algorithms are used, more data will give more precise models.

Van Doorn et al [26] used NN, SVR, and ARIMA to predict glucose values with CGM and accelerometer data, with prediction horizons of 15 and 60 minutes. At 15 minutes prediction horizon, the RMSE was 0.19 mmol/L for all patients ($n = 170$), and 0.29 mmol/L for DM2 patients ($n = 43$). [26] In a literature review by Woldaregay et al [21], RMSE reported at a prediction horizon of 15 minutes was between 0.14 and 0.66 mmol/L in DM1 patients. This study included NN, SVR, ARIMA, RF, and hybrid models. To compare, the RMSE in this study for the best personal models was 0.47, 0.49, and 0.53 mmol/L for the low-, moderate- and high-risk patients respectively, and for the best population model, the RMSE was 0.279 mmol/L. In both studies, NN performed best. Considering these outcomes, investigating NN for predicting glucose values is recommended. However, this does require some reconsideration regarding the number of features included. Another point from these studies is that more data improves prediction accuracy, as is also shown in this study.

4.2 Tuning

Tuning was done with default, random or grid values for hyperparameters. The initial tuning (see appendix B) showed that with LR, the strategy did not matter, with DT either random or grid would perform best, with RF grid performed best, and with XGBoost random performed best. In the final results (see also E), the best models for DT were tuned with a grid, though many best models per run used a random strategy. RF was more divided, and had each strategy represented in the final models, though default was less common in general, except with the population models. XGBoost had both random and grid, though the population model only used random. It is possible the random strategy underperformed in RF and XGBoost since the tune length of the random search strategy was initially set longer, but due

to a lack of resources, it was set shorter. Especially for algorithms (such as RF and XGBoost) that have many possible hyperparameter combinations, a longer tune length is recommended. Due to the nature of random tuning, there is a larger chance that more optimal tuning is found at longer tune lengths.

In this study, the unseen separate testing set was only used to evaluate models and result output. Especially in RF or XGBoost, higher *mtry* or *colsample.bytree* causes overfitting, although RF has another hyperparameter that should prevent overfitting, namely *min.node.size* (see also appendix B). XGBoost prevents overfitting with the *eta* and *min.child.weight* parameters, and *gamma*, though *gamma* was not used in tuning in this study. With DT there was only one hyperparameter to tune, *cp*, which determines the amount of pruning on the tree. Having a deeper and more complex tree means the model is also likely to overfit and the RMSE will be higher than estimated using the training set. Prevention of overfitting was done by using 10-fold cross-validation [35] and 10 runs to select the best model, but since DT generally had a higher error compared to other models and the baseline, it is possible that these models were overfitted. Especially on the smaller datasets and due to the dimensionality, RF and XGBoost are likely to overfit. It is recommended to lower the number of features or add more observations. [43]

There are ways to automatically tune models, using the ‘mlr3verse’ package in R. [44] This includes options for manual and automatic tuning, which could result in better hyperparameter optimization than was done in this study. This could also be applied for all runs of the training, since only the first run was used in tuning, and the other runs used the same tuning parameters as the first run.

With XGBoost only two hyperparameters were tuned, and this resulted in an improvement of RMSE when comparing it to the default tuning. However, random tuning outperformed the other strategies with one of the personal models and the population model. Since two out of the seven hyperparameters were investigated, it is recommended to investigate the other hyperparameters.

4.3 Feature importance

Regarding the feature importance, previous glucose measurements are clearly very important in modeling. However, these features are very heavily inter-correlated, especially the previous glucose values of 1 to 6 measurements ago and the rolling mean of short (1- to 5-hour) windows. Many of these features showed to be important in model creation. Although multicollinearity does not impact model performance, it does impact feature importance. [45] This makes it harder to interpret the actual importance.

LR determines importance using the coefficients, and if multicollinearity is present, the value of a coefficient will be lower if it is heavily correlated with other variables. RF uses permutation importance calculations, and with the presence of collinear variables, the importance is spread over the correlated variables. [46, 47] This very likely applies to DT and XGBoost as well, since these models are all tree-based. Since the splitrule ‘extratrees’ (Extremely randomized trees) is used in RF (see hyperparameters for the best models in appendix E), this issue should theoretically not be present for RF personal models. [46] The importance of features of this model could be assumed to be accurate. However, the population model used ‘variance’, so this implies that the importance of the highly correlated variables is spread among them.

Since the previous glucose measurements are highly correlated with each other (see appendix C), it is likely the feature importance calculations have underestimated their actual importance, so previous glucose measurements are likely even more important than is shown in this study. For basic glucose calculations, the rolling windows of the mean and SD are highly correlated between different window lengths, which means these features are also likely underestimated. Glycemic variability features also have high correlations, positive and negative, so it is likely some of these were underestimated in importance. The baseline subject and medication features had some correlated features, though only mix insulin was over 90%.

Multicollinearity could be detected with pairwise correlation, with a cut-off of 80 or 90 %. [48] Pairwise correlation does not necessarily indicate multicollinearity, so

other detection methods are recommended, such as calculating VIF (Variance Inflation Factor, where 10 or higher indicates multicollinearity) or PCA (Principal Component Approach, where small eigenvalues indicate a high chance of multicollinearity).

Multicollinearity can be mitigated by using different feature selection techniques. [48] For example, stepwise regression can be used with LR, which consists of forward selection (starting with no features, adding one at a time) or backward elimination (starting with all features, removing one at a time). For forward selection, the features are added that give the highest decrease in the residual sum of squares. For backward elimination, the removal of features depends on the lowest increase in the residual sum of squares. [48] With this amount of features, however, this type of selection would take a long time due to the number of possible combinations. Other options include Lasso, Elastic-net, or Ridge algorithms for feature selection. [48]

If multicollinearity is present, and features are removed, this could also impact model performance on the smaller (personal) datasets. As mentioned before, the higher dimensionality impacts the performance of RF and XGBoost, so if the amount of features is reduced, this will make RF and XGBoost more viable candidates for the personal models. On the other hand, the Boruta algorithm did not exclude any features, though Boruta does not consider the collinearity of variables. Correlation over 90% was mentioned in this study, but the highly correlated variables were not removed. To assess the importance of glycemic variability in models, it is recommended to investigate importance by making sure no multicollinearity is present with the method mentioned above, since this gives more robust models and an interpretable feature importance. [45]

4.4 Data processing and feature extraction

For feature extraction, a rolling window was often used. Although this gives very accurate calculations of the mean of the previous time interval, this requires more calculations to be done over many windows. Especially with more complex calculations, using a rolling window

could take a long running time. In larger datasets, a less intensive method could be used.

The inclusion of patients required gaps between measurements to be no larger than 4 hours. Glycemic variability calculations are less reliable when gaps between all measurements are over 2-4 hours for SD and CONGA and 1 hour for MAGE. [49] Since patients included with gaps only had occasional gaps that are up to 4 hours, most variability calculations could still be considered reliable. Another consideration is that lowering the maximum allowable gap size would cause fewer data to be included, and more manual processing would be necessary. On the other hand, with smaller gaps, CGM data could be preprocessed using interpolation, Kalman smoothing, or filtering. [7] Linear interpolation and Kalman smoothing have been shown to lower RMSE for some machine-learning models, including XGBoost. [50]

For all glycemic variability indices, the minimum interval to determine glycemic variability and adequate glucose control is 12 days, when considering SD and CV. [9] The recommended interval is 14 days. Since there were on average 11 days of data, the choice was made to only consider within-day variability per day. This is not a reliable indication of each patient's actual glycemic variability, though it can show short-term swings in glucose levels.

Not all glycemic variability indices were researched, partially due to not having enough data available to calculate all metrics. This was the case for ADRR and GMI. Other metrics, like CONGA and MODD, were not included due to high correlation with SD [51] and due to the complexity of the calculations involved.

Activity data was only based on steps, while other activities can also contribute to the blood glucose dynamics. Some wearables can automatically detect other types of activity, which could then be included as features. Another option is, like the dietary data, manual tracking, though this is less reliable. This could be done by weighing activity features on the level of physical activity. For example, walking could be considered low-level, while weightlifting could be considered a higher-level activity.

Other features that could be added are episodes of

hypo- or hyperglycemia. An episode could be defined as glucose levels above or below the target range for at least 2 measurements, or for 30 minutes.

Dietary data of some of the subjects were available, but in this study, this data was not considered. For inclusion of dietary features, one could consider carbohydrate intake or calorie intake, or create a model to estimate glucose absorption. [21]

Dynamic features that could be added are meal times and medication times. Fast-acting insulin can significantly lower glucose levels in a short amount of time, which is currently not reflected in the features since this study only looked into daily doses of medication. As with the glucose absorption model, one could consider different insulin uptake models, since different types of insulin will act differently with regard to lowering blood glucose values. Glucose response in interstitial fluid measurements have commonly a delay of five to ten minutes with respect to blood glucose measurements. [52] In practice, this is unlikely to have a substantial impact on diabetes management. When features of meal and insulin times are considered, one could consider taking this delay into account.

Other factors that influence the glucose levels of DM2 patients have not been included in this study, such as other diseases, sleep, and psychological state. [33] Certain conditions have been registered in DIALECT that could be included in future research. Sleep and psychological state could be considered for data collection. Sleep tracking is possible using a FitBit since this device is able to track the individual's sleep. The psychological state of a patient could be registered in a similar way to dietary data, by keeping a mood diary. This could be done with a daily mood or stress rating, or by registering different basic emotions (such as joy, sadness, fear, anger, and surprise) at different times of day.

4.5 Future perspective

Predictions of glucose levels can help improve diabetes self-management. This can be done by an alarm, that notifies DM2 patients of an oncoming glycemic event or shared decision-making based on future glucose values. [21] Since in this study, a short prediction horizon was

used (only the next value), this would not be as effective for an alarm as a longer prediction horizon, though longer prediction horizons will have a larger error. [7] The features that were used in this study focus heavily on the prediction of the next value. To investigate prediction horizons, the rolling window functions for features would need to incorporate more lag, and previous glucose measurements as a feature would need to be reconsidered. In other studies, prediction horizons from 15 minutes up to 2 hours have been investigated. [7]

5 Conclusion

In this study, four machine learning models were compared with regard to their performance in predicting the next glucose level in DM2 patients. On average, LR performed best for the low-risk, moderate-risk, and high-risk patients. RF and LR were the best final models for the low-risk patient, and LR was the best final model for the moderate- and high-risk patient. RF achieved the best population model. The most important feature categories were the previous glucose measurement and basic glucose calculations. Glycemic variability indices were not ranked highest in feature importance, some indices were in the top 15 for LR and XGBoost in the personal model, and some SD windows for DT, RF, and XGBoost in the population models.

Bibliography

- [1] Sapra A, Bhandari P. Diabetes Mellitus. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK551501/>.
- [2] Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of Diabetes and Diabetes-Related Complications. *Physical Therapy*. 2008 Nov;88(11):1254–1264. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3870323/>. doi:10.2522/ptj.20080020.
- [3] Centrum Gezondheid en Maatschappij van het Rijksinstituut voor Volksgezondheid en Milieu. Diabetes mellitus | Leefstijl en geslacht | Volksgezondheid en Zorg; 2022. Available from: <https://www.vzinfo.nl/diabetes-mellitus/leefstijl-en-geslacht>.
- [4] Kalra S, Mukherjee JJ, Venkataraman S, Bantwal G, Shaikh S, Saboo B, et al. Hypoglycemia: The neglected complication. *Indian Journal of Endocrinology and Metabolism*. 2013;17(5):819–834. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3784865/>. doi:10.4103/2230-8210.117219.
- [5] Breyton AE, Lambert-Porcheron S, Laville M, Vinoy S, Nazare JA. CGMS and Glycemic Variability, Relevance in Clinical Research to Evaluate Interventions in T2D, a Literature Review. *Frontiers in Endocrinology*. 2021;12. Available from: <https://www.frontiersin.org/articles/10.3389/fendo.2021.666008>. doi:10.3389/fendo.2021.666008.
- [6] Akasaka T, Sueta D, Tabata N, Takashio S, Yamamoto E, Izumiya Y, et al. Effects of the Mean Amplitude of Glycemic Excursions and Vascular Endothelial Dysfunction on Cardiovascular Events in Nondiabetic Patients With Coronary Artery Disease. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*. 2017 May;6(5). Publisher: Wiley-Blackwell. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5524064/>. doi:10.1161/JAHA.116.004841.
- [7] Woldaregay AZ, Årsand E, Botsis T, Albers D, Mamykina L, Hartvigsen G. Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes. *Journal of Medical Internet Research*. 2019 May;21(5). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6658321/>. doi:10.2196/11030.
- [8] McCall AL, Cox DJ, Crean J, Gloster M, Kovatchev BP. A Novel Analytical Method for Assessing Glucose Variability: Using CGMS in Type 1 Diabetes Mellitus. *Diabetes Technology & Therapeutics*. 2006 Dec;8(6):644–653. Publisher: Mary Ann Liebert, Inc., publishers. Available from: <https://www.liebertpub.com/doi/abs/10.1089/dia.2006.8.644>. doi:10.1089/dia.2006.8.644.

- [9] Danne T, Nimri R, Battelino T, Bergenstal RM, Close KL, DeVries JH, et al. International Consensus on Use of Continuous Glucose Monitoring. *Diabetes Care*. 2017 Dec;40(12):1631–1640. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6467165/>. doi:10.2337/dc17-1600.
- [10] Kusunoki Y, Konishi K, Tsunoda T, Koyama H. Significance of Glycemic Variability in Diabetes Mellitus. *Internal Medicine*. 2022 Feb;61(3):281–290. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8866772/>. doi:10.2169/internalmedicine.8424-21.
- [11] Sartore G, Chilelli NC, Burlina S, Lapolla A. Association between glucose variability as assessed by continuous glucose monitoring (CGM) and diabetic retinopathy in type 1 and type 2 diabetes. *Acta Diabetologica*. 2013 Jun;50(3):437–442. doi:10.1007/s00592-013-0459-9.
- [12] Frontoni S, Di Bartolo P, Avogaro A, Bosi E, Paolisso G, Ceriello A. Glucose variability: An emerging target for the treatment of diabetes mellitus. *Diabetes Research and Clinical Practice*. 2013 Nov;102(2):86–95. Available from: <https://www.sciencedirect.com/science/article/pii/S0168822713003227>. doi:10.1016/j.diabres.2013.09.007.
- [13] Daniels J, Herrero P, Georgiou P. A Multitask Learning Approach to Personalized Blood Glucose Prediction. *IEEE Journal of Biomedical and Health Informatics*. 2022 Jan;26(1):436–445. Conference Name: IEEE Journal of Biomedical and Health Informatics. doi:10.1109/JBHI.2021.3100558.
- [14] García Maset L, González LB, Furquet GL, Suay FM, Marco RH. Study of Glycemic Variability Through Time Series Analyses (Detrended Fluctuation Analysis and Poincaré Plot) in Children and Adolescents with Type 1 Diabetes. *Diabetes Technology & Therapeutics*. 2016 Nov;18(11):719–724. Publisher: Mary Ann Liebert, Inc., publishers. Available from: <https://www.liebertpub.com/doi/10.1089/dia.2016.0208>. doi:10.1089/dia.2016.0208.
- [15] Kovatchev B, Umpierrez G, DiGenio A, Zhou R, Inzucchi SE. Sensitivity of Traditional and Risk-Based Glycemic Variability Measures to the Effect of Glucose-Lowering Treatment in Type 2 Diabetes Mellitus. *Journal of Diabetes Science and Technology*. 2015 Oct;9(6):1227–1235. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4667308/>. doi:10.1177/1932296815587014.
- [16] McDonnell CM, Donath SM, Vidmar SI, Werther GA, Cameron FJ. A Novel Approach to Continuous Glucose Analysis Utilizing Glycemic Variation. *Diabetes Technology & Therapeutics*. 2005 Apr;7(2):253–263. Publisher: Mary Ann Liebert, Inc., publishers. Available from: <https://www.liebertpub.com/doi/10.1089/dia.2005.7.253>. doi:10.1089/dia.2005.7.253.
- [17] Kovatchev BP, Clarke WL, Breton M, Brayman K, McCall A. Quantifying Temporal Glucose Variability in Diabetes via Continuous Glucose Monitoring: Mathematical Methods and Clinical Application. *Diabetes Technology & Therapeutics*. 2005 Dec;7(6):849–862. Publisher: Mary Ann Liebert, Inc., publishers. Available from: <https://www.liebertpub.com/doi/abs/10.1089/dia.2005.7.849>. doi:10.1089/dia.2005.7.849.
- [18] Patton SR, Clements MA. Average Daily Risk Range as a Measure for Clinical Research and Routine Care. *Journal of Diabetes Science and Technology*. 2013 Sep;7(5):1370–1375. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3876383/>.
- [19] Bergenstal RM, Beck RW, Close KL, Grunberger G, Sacks DB, Kowalski A, et al. Glucose Management Indicator (GMI): A New Term for Estimating A1C From Continuous Glucose Monitoring. *Diabetes Care*. 2018 Sep;41(11):2275–2280. doi:10.2337/dc18-1581.
- [20] Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*.

- 2021 Dec;13:148. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8686642/>. doi:10.1186/s13098-021-00767-9.
- [21] Woldaregay AZ, Årsand E, Walderhaug S, Albers D, Mamykina L, Botsis T, et al. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial Intelligence in Medicine*. 2019 Jul;98:109–134. Available from: <https://www.sciencedirect.com/science/article/pii/S0933365717306218>. doi:10.1016/j.artmed.2019.07.007.
- [22] Tsihklaki S, Koumakis L, Tsiknakis M. Type 1 Diabetes Hypoglycemia Prediction Algorithms: Systematic Review. *JMIR Diabetes*. 2022 Jul;7(3):e34699. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9353679/>. doi:10.2196/34699.
- [23] Alexiou S, Draitsas E, Kocsis O, Moustakas K, Fakotakis N. An approach for Personalized Continuous Glucose Prediction with Regression Trees. In: 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM). Preveza, Greece: IEEE; 2021. p. 1–6. Available from: <https://ieeexplore.ieee.org/document/9566278/>. doi:10.1109/SEEDA-CECNSM53056.2021.9566278.
- [24] Elhadd T, Mall R, Bashir M, Palotti J, Fernandez-Luque L, Farooq F, et al. Artificial Intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (The PRO-FAST – IT Ramadan study). *Diabetes Research and Clinical Practice*. 2020 Nov;169:108388. Available from: <https://www.sciencedirect.com/science/article/pii/S0168822720306410>. doi:10.1016/j.diabres.2020.108388.
- [25] Zafar A, Lewis DM, Shahid A. Long-Term Glucose Forecasting for Open-Source Automated Insulin Delivery Systems: A Machine Learning Study with Real-World Variability Analysis. *Healthcare*. 2023 Mar;11(6):779. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10048652/>. doi:10.3390/healthcare11060779.
- [26] van Doorn WPTM, Foreman YD, Schaper NC, Savelberg HHCM, Koster A, van der Kallen CJH, et al. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. *PLoS ONE*. 2021 Jun;16(6):e0253125. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8224858/>. doi:10.1371/journal.pone.0253125.
- [27] Seo W, Park SW, Kim N, Jin SM, Park SM. A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study. *Computer Methods and Programs in Biomedicine*. 2021 Nov;211:106424. Available from: <https://www.sciencedirect.com/science/article/pii/S0169260721004983>. doi:10.1016/j.cmpb.2021.106424.
- [28] Zale AD, Abusamaan MS, McGready J, Mathioudakis N. Prediction of Next Glucose Measurement in Hospitalized Patients by Comparing Various Regression Methods: Retrospective Cohort Study. *JMIR Formative Research*. 2023 Jan;7:e41577. Available from: <https://formative.jmir.org/2023/1/e41577>. doi:10.2196/41577.
- [29] Rodríguez-Rodríguez I, Chatzigiannakis I, Rodríguez JV, Maranghi M, Gentili M, Zamora-Izquierdo M. Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through Machine Learning Techniques. *Sensors*. 2019 Jan;19(20):4482. Number: 20 Publisher: Multidisciplinary Digital Publishing Institute. Available from: <https://www.mdpi.com/1424-8220/19/20/4482>. doi:10.3390/s19204482.
- [30] Bunescu R, Struble N, Marling C, Shubrook J, Schwartz F. Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression. In: 2013 12th International Conference on Machine Learning and Applications. vol. 1; 2013. p. 135–140. doi:10.1109/ICMLA.2013.30.

- [31] Marcus Y, Eldor R, Yaron M, Shaklai S, Ish-Shalom M, Shefer G, et al. Improving blood glucose level predictability using machine learning. *Diabetes/Metabolism Research and Reviews*. 2020;36(8):e3348. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/dmrr.3348>. doi:10.1002/dmrr.3348.
- [32] Frandes M, Timar B, Timar R, Lungeanu D. Chaotic time series prediction for glucose dynamics in type 1 diabetes mellitus using regime-switching models. *Scientific Reports*. 2017 Jul;7(1):6232. Number: 1 Publisher: Nature Publishing Group. Available from: <https://www.nature.com/articles/s41598-017-06478-4>. doi:10.1038/s41598-017-06478-4.
- [33] Contreras I, Oviedo S, Vettoretti M, Visentin R, Vehí J. Personalized blood glucose prediction: A hybrid approach using grammatical evolution and physiological models. *PLOS ONE*. 2017 Nov;12(11):e0187754. Publisher: Public Library of Science. Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187754>. doi:10.1371/journal.pone.0187754.
- [34] Beernink JM, Oosterwijk MM, Khunti K, Gupta P, Patel P, Boven JFMv, et al. Biochemical Urine Testing of Medication Adherence and Its Association With Clinical Markers in an Outpatient Population of Type 2 Diabetes Patients: Analysis in the DIAbetes and LiFEstyle Cohort Twente (DI-ALECT). *Diabetes Care*. 2021 Jun;44(6):1419. Publisher: American Diabetes Association. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8247497/>. doi:10.2337/dc20-2533.
- [35] Marsland S. *Machine Learning: An Algorithmic Perspective*. 2nd ed. Chapman and Hall/CRC; 2014. Available from: <https://www.taylorfrancis.com/books/9781466583337>. doi:10.1201/b17476.
- [36] Mitchell TM. *Machine Learning*. McGraw-Hill series in computer science. New York: McGraw-Hill; 1997.
- [37] Duckworth C, Guy MJ, Kumaran A, O’Kane AA, Ayobi A, Chapman A, et al. Explainable Machine Learning for Real-Time Hypoglycemia and Hyperglycemia Prediction and Personalized Control Recommendations. *Journal of Diabetes Science and Technology*. 2022 Jun. Publisher: SAGE Publications Inc. doi:10.1177/19322968221103561.
- [38] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785–794. ArXiv:1603.02754 [cs]. Available from: <http://arxiv.org/abs/1603.02754>. doi:10.1145/2939672.2939785.
- [39] Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *Journal of Statistical Software*. 2010 Sep;36:1–13. doi:10.18637/jss.v036.i11.
- [40] Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating Clinical Accuracy of Systems for Self-Monitoring of Blood Glucose. *Diabetes Care*. 1987 Sep;10(5):622–628. doi:10.2337/diacare.10.5.622.
- [41] Kuhn M. 15 Variable Importance | The caret Package; 2019. Available from: <https://topepo.github.io/caret/variable-importance.html>.
- [42] Chen T, Benesty M, Tang Y, He T, Khotilovich V, Cho H, et al.. XGBoost Parameters — xgboost 1.7.6 documentation; 2022. Revision 36eb41c9. Available from: <https://xgboost.readthedocs.io/en/stable/parameter.html>.
- [43] Xu F, Zhao H, Zhou W, Zhou Y. Cost-sensitive regression learning on small dataset through intra-cluster product favoured feature selection. *Connection Science*. 2022 Dec;34(1):104–123. doi:10.1080/09540091.2021.1970719.
- [44] Lang M, Schratz P. *mlr3verse: Easily Install and Load the ‘mlr3’ Package Family*; 2023. Available from: <https://CRAN.R-project.org/package=mlr3verse>.
- [45] Kutner MH, editor. *Applied linear statistical models*. 5th ed. The McGraw-Hill/Irwin series operations and decision sciences. Boston: McGraw-Hill Irwin; 2005.

- [46] Parr T, Turgutlu K, Csiszar C, Howard J. Beware Default Random Forest Importances; 2018. Available from: <http://explained.ai/decision-tree-viz/index.html>.
- [47] Tološi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011 Jul;27(14):1986–1994. Available from: <https://doi.org/10.1093/bioinformatics/btr300>. doi:10.1093/bioinformatics/btr300.
- [48] Chan JYL, Leow SMH, Bea KT, Cheng WK, Phong SW, Hong ZW, et al. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*. 2022 Jan;10(8):1283. Available from: <https://www.mdpi.com/2227-7390/10/8/1283>. doi:10.3390/math10081283.
- [49] Baghurst PA, Rodbard D, Cameron FJ. The Minimum Frequency of Glucose Measurements from Which Glycemic Variation Can Be Consistently Assessed. *Journal of Diabetes Science and Technology*. 2010 Nov;4(6):1382–1385. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3005048/>.
- [50] Acuna E, Aparicio R, Palomino V. Analyzing the Performance of Transformers for the Prediction of the Blood Glucose Level Considering Imputation and Smoothing. *Big Data and Cognitive Computing*. 2023 Mar;7(1):41. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. Available from: <https://www.mdpi.com/2504-2289/7/1/41>. doi:10.3390/bdcc7010041.
- [51] Rodbard D. Glucose Variability: A Review of Clinical Applications and Research Developments. *Diabetes Technology & Therapeutics*. 2018 Jun;20(S2):S2–5. Available from: <https://www.liebertpub.com/doi/full/10.1089/dia.2018.0092>. doi:10.1089/dia.2018.0092.
- [52] Rebrin K, Sheppard NF, Steil GM. Use of Subcutaneous Interstitial Fluid Glucose to Estimate Blood Glucose: Revisiting Delay and Sensor Offset. *Journal of Diabetes Science and Technology*. 2010 Sep;4(5):1087–1098. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2956819/>.
- [53] Kuhn M. 6 Available Models | The caret Package; 2019. Available from: <https://topepo.github.io/caret/available-models.html>.
- [54] Boehmke B, Greenwell B. Chapter 11 Random Forests | *Hands-On Machine Learning with R*. Taylor & Francis Group; 2020. Available from: <https://bradleyboehmke.github.io/HOML/random-forest.html>.
- [55] Probst P, Wright M, Boulesteix AL. Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining and Knowledge Discovery*. 2019 May;9(3). Available from: <http://arxiv.org/abs/1804.03515>. doi:10.1002/widm.1301.
- [56] Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*. 2017;36(8):1272–1284. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7212>. doi:10.1002/sim.7212.

List of Figures

1	Flowchart documenting inclusion of study population.	4
2	Empty Clarke Error Grid	7
3	Boruta Algorithm results, only the top 30 features are shown.	9
4	Feature importance of the personal models	12
5	Feature importance of the population models	12
6	Clarke Error Grids of the population models	14
7	Example of Moderate-risk patient: grid with 2 splitrules, <i>mtry</i> vs RMSE.	30
8	Correlation matrices of each feature group, except activity-based features	32
9	Personal model Clarke Error Grids of the best models. The columns indicate the different patients and the rows indicate the different model types	33

List of Tables

1	Table of categories of features used, color-coded.	5
2	Baseline characteristics of included patients	8
3	Baseline characteristics for personal model subjects	8
4	Evaluation of models, personal and population	10
5	List of abbreviations	24
6	Included features, definition, source, and type of data	25
7	Result of different tuning strategies	31
8	Hyperparameter values for LR	34
9	Hyperparameter values for DT	34
10	Hyperparameter values for RF	34
11	Hyperparameter values for XGBoost	35

List of Abbreviations

Table 5: List of abbreviations

Abbreviation	Definition
DM2	Diabetes Mellitus type 2
BMI	Body Mass Index
HbA1c	Glycated hemoglobin
CGM	Continuous Glucose Monitoring
SD	Standard Deviation
LBGI	Low Blood Glucose Index
HBGI	High Blood Glucose Index
CV	Coefficient of Variation
MAGE	Mean Amplitude of Glycemic Excursions
CONGA	Continuous Overall Net Glycemic Action
MODD	Mean of Daily Differences
GMI	Glucose Management Indicator
ADRR	Average Daily Risk Range
TIR	Time In Range
TAR	Time Above Range
TAHR	Time Above High Range
TBR	Time Below Range
TBLR	Time Below Low Range
IQR	Inter Quartile Range
LR	Linear Regression
DT	Decision Tree
RF	Random Forest
XGBoost	Extreme Gradient Boosting
NN	Neural Network

A Appendix: List of Features

Table 6: Included features, definition, source, and type of data

Abbreviation	Definition	Source	Type	Category	Color
wday	Day of the week, Monday to Sunday	Time	Factor	Time	Orange
weekend	Whether the current date is a day in the weekend	Time	Factor	Time	
hour	Hour of day	Time	Integer	Time	
part_of_day	Part of the day, divided into Morning (06:00 - 12:00), Afternoon (12:00 - 18:00), Evening (18:00 - 00:00) and Night (00:00 - 06:00)	Time	Factor	Time	
season	Meteorological season, divided into Spring, Summer, Autumn and Winter	Time	Factor	Time	
gl	Glucose measurement in mmol/L. Target value for predictions	CGM	Numeric	Target variable	Blue
gl_minus_1	Glucose measurement 1 measurement ago	CGM	Numeric	Previous glucose measurements	
gl_minus_2	Glucose measurement 2 measurements ago	CGM	Numeric	Previous glucose measurements	
gl_minus_3	Glucose measurement 3 measurements ago	CGM	Numeric	Previous glucose measurements	
gl_minus_4	Glucose measurement 4 measurements ago	CGM	Numeric	Previous glucose measurements	
gl_minus_5	Glucose measurement 5 measurements ago	CGM	Numeric	Previous glucose measurements	
gl_minus_6	Glucose measurement 6 measurements ago	CGM	Numeric	Previous glucose measurements	
gl_previous_hour	Glucose measurement 1 hour ago	CGM	Numeric	Previous glucose measurements	
gl_previous_day	Glucose measurement 24 hours ago	CGM	Numeric	Previous glucose measurements	
rolling_av_1hour	Rolling window mean of glucose measurements of the previous hour (not including current measurement)	CGM	Numeric	Basic glucose calculations	
rolling_av_2hour	Rolling window mean of glucose measurements of the previous 2 hours	CGM	Numeric	Basic glucose calculations	
rolling_av_3hour	Rolling window mean of glucose measurements of the previous 3 hours	CGM	Numeric	Basic glucose calculations	
rolling_av_4hour	Rolling window mean of glucose measurements of the previous 4 hours	CGM	Numeric	Basic glucose calculations	

rolling_av_5hour	Rolling window mean of glucose measurements of the previous 5 hours	CGM	Numeric	Basic glucose calculations
rolling_av_6hour	Rolling window mean of glucose measurements of the previous 6 hours	CGM	Numeric	Basic glucose calculations
rolling_sd_1hour	Rolling window standard deviation of glucose measurements of the previous hour	CGM	Numeric	Basic glucose calculations
rolling_sd_2hour	Rolling window standard deviation of glucose measurements of the previous 2 hours	CGM	Numeric	Basic glucose calculations
rolling_sd_3hour	Rolling window standard deviation of glucose measurements of the previous 3 hours	CGM	Numeric	Basic glucose calculations
rolling_sd_4hour	Rolling window standard deviation of glucose measurements of the previous 4 hours	CGM	Numeric	Basic glucose calculations
rolling_sd_5hour	Rolling window standard deviation of glucose measurements of the previous 5 hours	CGM	Numeric	Basic glucose calculations
rolling_sd_6hour	Rolling window standard deviation of glucose measurements of the previous 6 hours	CGM	Numeric	Basic glucose calculations
rolling_av_day	Rolling window mean of glucose measurements of the previous 24 hours	CGM	Numeric	Basic glucose calculations
rolling_min_day	Minimum glucose value of the previous 24 hours	CGM	Numeric	Basic glucose calculations
rolling_max_day	Maximum glucose value of the previous 24 hours	CGM	Numeric	Basic glucose calculations
previous_slope	Slope of the two previous measurements	CGM	Numeric	Basic glucose calculations
rolling_sd_day	Rolling window standard deviation of glucose measurements of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_iqr_day	Rolling window interquartile range of glucose measurements of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_LBGI_day	Low Blood Glucose Index of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_HBGI_day	High Blood Glucose Index of the previous 24 hours	CGM	Numeric	Glycemic variability indices
cv_day	Coefficient of Variation (%) of the previous 24 hours	CGM	Numeric	Glycemic variability indices
J_index_day	J-index of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_TIR_day	Time In Range (%) (between 3.9 and 10 mmol/L) of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_TAR_day	Time Above Range (%) (above 10 mmol/L) of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_TAHR_day	Time Above High Range (%) (above 13.9 mmol/L) of the previous 24 hours	CGM	Numeric	Glycemic variability indices

rolling_TBR_day	Time Below Range (%) (below 3.9 mmol/L) of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_TBLR_day	Time Below Low Range (%) (below 3 mmol/L) of the previous 24 hours	CGM	Numeric	Glycemic variability indices
rolling_mage_day	Mean Amplitude of Glucose Excursions (mmol/L) of the previous 24 hours	CGM	Numeric	Glycemic variability indices
sex_female	Whether the sex of the subject is female	Subject	Factor	Baseline subject characteristics
age	Age of subject at time of measurement	Subject	Integer	Baseline subject characteristics
DM2_years	Years of Diabetes Type 2 diagnosis	Subject	Integer	Baseline subject characteristics
smoking_current	Whether subject currently smokes	Subject	Factor	Baseline subject characteristics
smoking_ever	Whether subject ever smoked	Subject	Factor	Baseline subject characteristics
packyears	Amount of pack years	Subject	Integer	Baseline subject characteristics
alcohol_units_month	Amount of alcohol units per month	Subject	Integer	Baseline subject characteristics
length	Length of subject in cm	Subject	Integer	Baseline subject characteristics
weight	Weight of subject in kg	Subject	Integer	Baseline subject characteristics
BMI	Body Mass Index in kg/m ²	Subject	Numeric	Baseline subject characteristics
waist_circumference	Waist Circumference in cm	Subject	Integer	Baseline subject characteristics
hip_circumference	Hip Circumference in cm	Subject	Integer	Baseline subject characteristics
HbA1c	Hemoglobin A1c (mmol/mol) (glycated hemoglobin) from blood test	Subject	Integer	Baseline subject characteristics
fast_insulin	Whether the subject uses fast-acting insulin	Subject	Factor	Medication
fast_insulin_dosage	Dosage of fast-acting insulin per day	Subject	Integer	Medication
mix_insulin	Whether the subject uses a mix (combination of slow- and fast-acting insulin) insulin	Subject	Factor	Medication
mix_insulin_dosage	Dosage of mix insulin	Subject	Integer	Medication
long_insulin	Whether the subject uses slow-acting insulin	Subject	Factor	Medication
long_insulin_dosage	Dosage of slow-acting insulin per day	Subject	Integer	Medication
Metformin	Whether the subject uses metformin	Subject	Factor	Medication
Metformin dosage	Dosage of metformin per day	Subject	Integer	Medication
SU_derivatives	Whether the subject uses sulfonylurea-derivatives, which lower mean blood sugar	Subject	Factor	Medication

SU_derivatives.dosage	Dosage of SU-derivatives per day	Subject	Integer	Medication
steps.day	Sum of steps of the past 24 hours	Steps	Numeric	Activity
steps.hour	Sum of steps of the past hour	Steps	Numeric	Activity



B Appendix: Tuning process

The tuning process per model type is explained in the sections below. The RMSE results per tuning strategy, model type, and dataset are also reported for the first data split training.

B.1 Tuning parameters

B.1.1 Linear Regression

Linear Regression only has one parameter within caret, when using the ‘lm’ method, the *intercept* can be set to TRUE or FALSE. [53] In this case, doing a random search will not make a difference with the grid search. So in this case, the model performance is evaluated for an *intercept* set at default (TRUE) at a grid (TRUE or FALSE).

B.1.2 Decision Tree

Decision Tree has one parameter when using ‘rpart’, which is *cp*. [53] *cp* is a complexity parameter, where if it is higher, more pruning will be done on the tree and the complexity of the model decreases, meaning fewer nodes are involved. The default values are based on a default grid determined by caret.

B.1.3 Random Forest

Random Forest has 3 parameters when using ‘ranger’: *mtry*, *min.node.size*, and *splitrule*. [53] *mtry* refers to the number of features that need to be considered when making a split. [54, 55] These are randomly trawn. Lower values of *mtry* give less correlated and more stable trees but also lead to lower accuracy. [55] *min.node.size* refers to the minimum number of observations in a terminal leaf node. This implies that if a split in a node results in one split with a lower number than this minimum node size, this node will become a terminal node. This is a way to manage tree depth. [55] *splitrule* can be either ‘variance’, ‘extratrees’, or ‘maxstat’. Variance is the original splitting rule, and means that a selection is made out of all splits of the *mtry* amount of variables based on the weighted variance. Extratrees is extremely randomized trees and this *splitrule* randomizes the cut-off values for node splitting. This rule adds more randomness to the trees. Maxstat refers to maximally selected rank statistics, this rule selects variables based on a p-value approximation. The variable with the lowest p-value is chosen for splitting, and then an adjusted p-value is calculated for testing of the *mtry* variables. If this adjusted p-value is smaller than a specified error, the split is made. This way the optimal split points are determined. [55, 56]

B.1.4 XGBoost

XGBoost has 7 parameters when ‘xgbTree’ is used. [42, 53]

nrounds refers to the amount of boosting iterations. Default values are 50, 100, and 150.

max.depth refers to the maximum tree depth, increasing this value gives a more complex model which is more likely to overfit. Possible values are 0 to infinity. Default values are 1, 2, and 3.

eta refers to shrinkage and values are between 0 and 1. This means the boosting process is made more conservative with higher values, due to the weights being shrunk at each boosting step. The default values are 0.3 and 0.4.

gamma refers to the minimum loss reduction requirement to make more nodes on the tree, which means a larger *gamma* causes the algorithm to be more conservative. Possible values are 0 to infinity, and the default value is 0.

colsample.bytree refers to the subsample ratio of columns when constructing each tree. This refers to the fraction of randomly selected features, which will then be used to train each tree. Possible values are between 0 and 1, and default values are 0.6 and 0.8.

min_child_weight refers to the minimum sum of instance weight needed in a child. If a leaf node is made with a sum of instance weight less than this minimum, the tree should stop making more nodes after this point. This implies that a higher value will cause the algorithm to be more conservative. Possible values are 0 to infinity, and the default value is 1.

subsample refers to the subsample ratio of the training instances. This ratio of the training data is randomly sampled prior to growing trees, and this should prevent overfitting. Possible values are between 0 and 1; the default values are 0.5, 0.75, and 1.

B.2 Tuning evaluation

The RMSE of each model and strategy is reported in table 7.

B.2.1 Linear Regression

For all LR models, there was no difference between setting *intercept* as TRUE or FALSE for the outcome of the model. So, for LR it does not matter whether the default or grid tuning strategy is used.

B.2.2 Decision Tree

For the personal DT models, the default strategy resulted in a higher RMSE compared to the random search strategy. A grid was made for each personal DT model. This was re-evaluated multiple times by adjusting the grid slightly. This was done based on plots of the hyperparameter vs the RMSE, to evaluate what range would be appropriate. One grid (for the Moderate-Risk patient) caused improvements in RMSE, and the other grids did not improve RMSE compared to the random search tuning strategy.

For the population DT model, the default strategy also underperformed. The grid strategy performed slightly better compared to the random strategy.

B.2.3 Random Forest

The grids for the personal models were made by changing and readjusting the grids many times, and resulted in choosing one splitrule, namely 'extratrees'. *mtry* and *splitrule* seem to be related, higher *mtry* in 'variance' splitrule causes higher RMSE, while higher *mtry* in 'extratrees' gives much lower RMSE. See figure 7, which shows that choosing 'extratrees' with a higher *mtry* will provide a lower RMSE. In the personal models, all strategies seem to have a similar performance, though the grid performed best overall in this run. For the population model, 'extratrees' with a high *mtry* also performed best in this run of testing and tuning, and the grid strategy had the lowest RMSE.

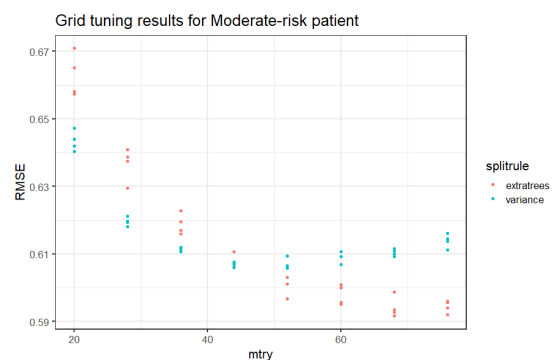


Figure 7: Example of Moderate-risk patient: grid with 2 splitrules, *mtry* vs RMSE.

B.2.4 XGBoost

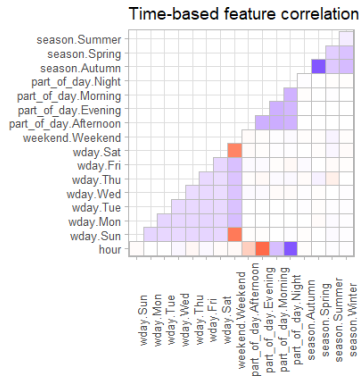
In XGBoost tuning, the random search did outperform the default strategy, but due to very many possible parameter combinations with low RMSE, the grid required for a grid strategy would become very large. For example, since there

Table 7: Result of different tuning strategies

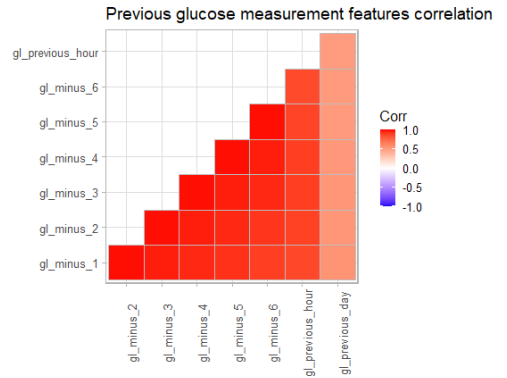
Model	Strategy	Low Risk	Moderate Risk	High Risk	Population
		RMSE	RMSE	RMSE	RMSE
LR	Default	0.4745	0.5514	0.5263	0.2900
	Grid	0.4745	0.5514	0.5263	0.2900
DT	Default	0.8476	2.2473	2.2555	1.4883
	Random	0.5838	0.7247	0.6790	0.3201
	Grid	0.5874	0.6954	0.6942	0.3167
RF	Default	0.5040	0.5498	0.5531	0.2826
	Random	0.5088	0.5600	0.5589	0.2831
	Grid	0.5020	0.5484	0.5467	0.2825
XGBoost	Default	0.5179	0.6164	0.5936	0.2914
	Random	0.4855	0.5338	0.5159	0.2783
	Grid	0.5259	0.5799	0.5907	0.2883
Baseline	None	0.6293	0.6468	0.6656	0.3868

are 7 parameters to tune, using a grid with 3 values for each parameter would result in $3^7 = 2187$ possible permutations, which would take too long to run. Two parameters (*colsample_bytree* and *subsample*) were used for a grid, the rest of the parameters were kept at default values. The grid did outperform the default strategy, however, it did not outperform a random search in this run.

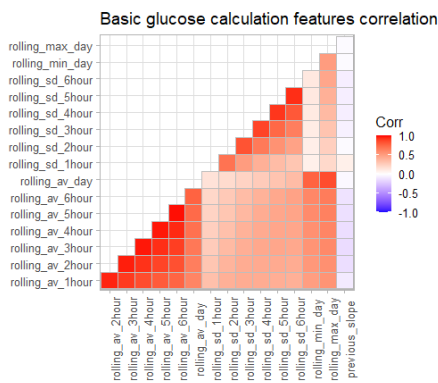
C Appendix: Correlation matrices



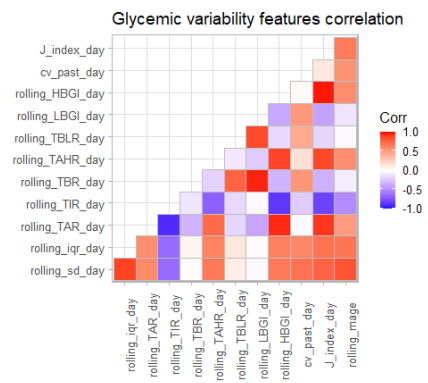
(a) Correlation matrix of time-based features



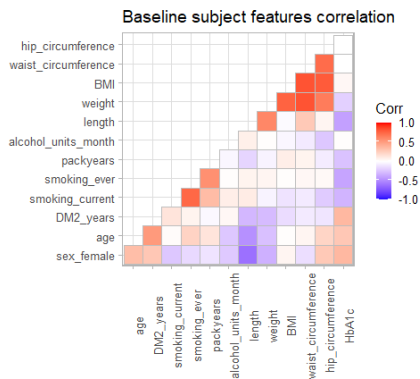
(b) Correlation matrix of previous glucose measurement features



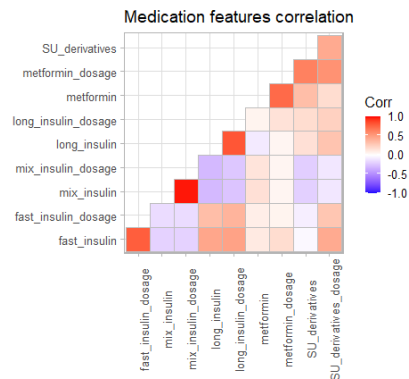
(c) Correlation matrix of basic glucose calculation features



(d) Correlation matrix of glycemic variability features



(e) Correlation matrix of baseline subject features



(f) Correlation matrix of medication features

Figure 8: Correlation matrices of each feature group, except activity-based features

D Appendix: Personal model Clarke Error Grids

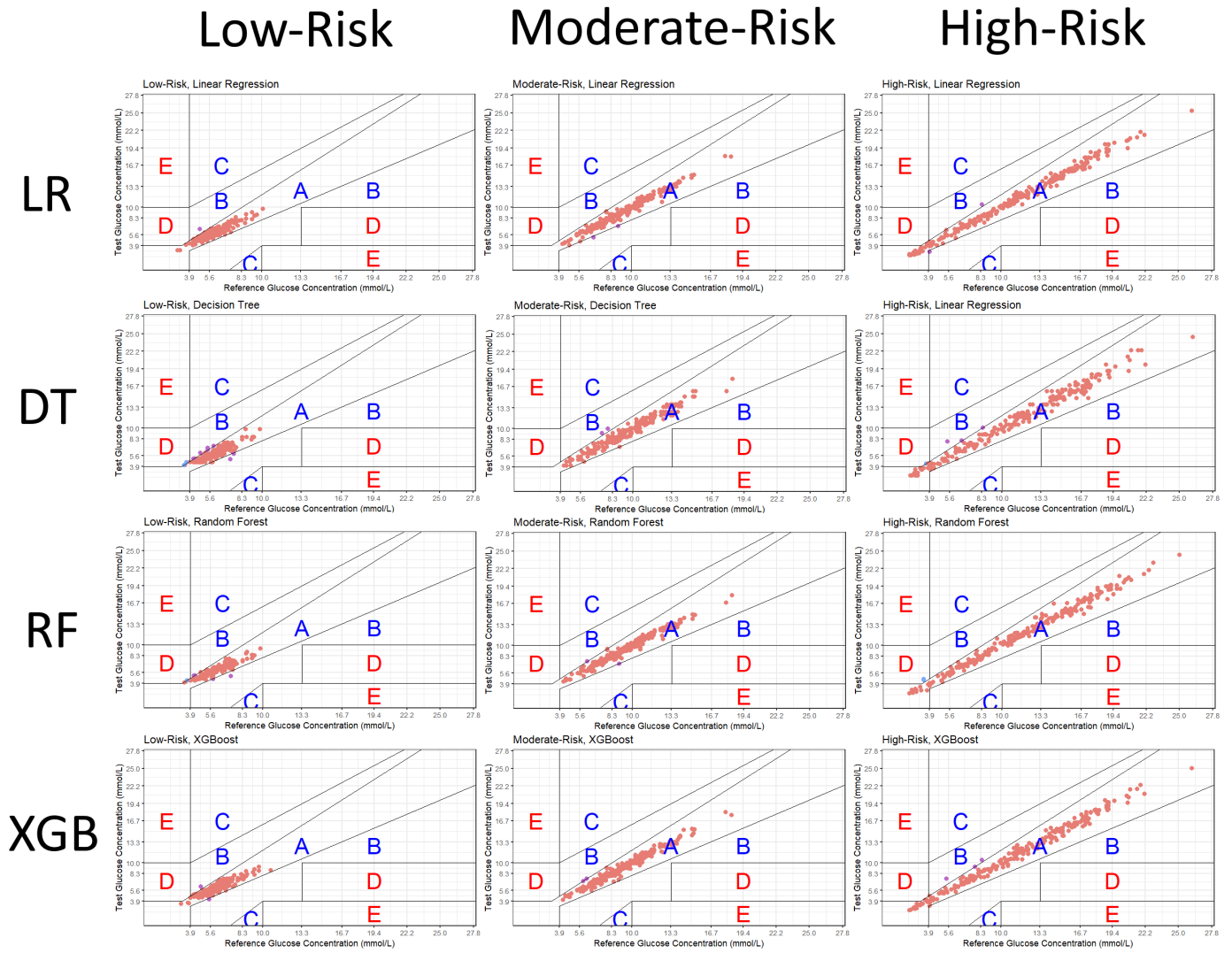


Figure 9: Personal model Clarke Error Grids of the best models. The columns indicate the different patients and the rows indicate the different model types

E Appendix: Hyperparameter values for the best models

Table 8: Hyperparameter values for LR

Dataset	Strategy	intercept
Low-Risk patient	Default	TRUE
Moderate-Risk patient	Default	TRUE
High-Risk patient	Default	TRUE
Population	Grid	FALSE

Table 9: Hyperparameter values for DT

Dataset	Strategy	cp
Low-Risk patient	Grid	0.002
Moderate-Risk patient	Grid	2.00E-04
High-Risk patient	Grid	2.00E-05
Population	Grid	4.00E-06

Table 10: Hyperparameter values for RF

Dataset	Strategy	mtry	splitrule	min.node.size
Low-Risk patient	Default	40	extratrees	5
Moderate-Risk patient	Grid	62	extratrees	3
High-Risk patient	Random	69	extratrees	3
Population	Default	40	variance	5

Table 11: Hyperparameter values for XGBoost

Dataset	Strategy	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
Low-Risk patient	Grid	50	2	0.3	0	1	1	0.9
Moderate-Risk patient	Random	526	8	0.12	2.2	0.38	2	0.42
High-Risk patient	Grid	50	3	0.3	0	1	1	0.9
Population	Random	195	10	0.0773	3.798	0.613	13	0.429