Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 349-362

Improving postprandial glucose forecasting using diagnosis-aware stacked learning

Fatma Indriani*[1], Mohammad Reza Faisal[1], Naufal Said[1]
Department of Computer Science, Lambung Mangkurat University, Indonesia[1]

Article Info
Keywords:
Postprandial Glucose Prediction, Stacked Learning, Group-Specific Modeling, Multimodal Health Data, Continuous Glucose Monitoring, Meal Composition Features

Article history:
Received: November 06, 2025
Accepted: March 10, 2026
Published: May 01, 2026

Cite:
F. Indriani, M. R. Faisal, and N. Said, "Improving Postprandial Glucose Forecasting Using Diagnosis-Aware Stacked Learning", *KINETIK*, vol. 11, no. 2, May 2026.
https://doi.org/10.22219/kinetik.v11i2.2566

*Corresponding author.
Fatma Indriani
E-mail address:
f.indriani@ulm.ac.id

Abstract
Predicting glucose levels after a meal (postprandial glucose) can help anticipate abnormal responses and improve diabetes management. Yet such prediction remains difficult because post-meal glucose depends on multiple interacting factors, including prior glucose trends, meal composition, and recent activity. This study develops machine learning models to forecast short-term post-meal glucose levels using the CGMacros dataset, which combines continuous glucose monitoring (CGM) data from Dexcom and Libre sensors with meal macronutrient annotations and activity measurements. Several feature combinations and regression models were evaluated to identify an optimal representation. Results show that combining baseline glucose statistics with meal composition yields the lowest error across all regressors. Building on this feature configuration, a stacked learning framework was implemented in which a global model provides initial predictions refined by diagnosis-specific CatBoost regressors for Healthy, Pre-diabetes, and Type 2 Diabetes groups. Across 18 configurations spanning two sensors and three horizons (30, 60, 120 minutes), stacking reduced normalized RMSE by 3.5 ± 3.7% on average, with the strongest improvements at 120-minute horizons (mean 5.5%) and for linear global models (up to 13.6% reduction). Gains varied by diagnosis group and sensor type, highlighting the importance of device-aware validation. These results demonstrate that diagnosis-aware stacking enhances both accuracy and robustness, offering a practical foundation for personalized glucose forecasting in digital health systems.

1. Introduction

Glucose regulation after meals is a complex physiological process influenced by individual metabolic status, meal composition, and physical activity. Accurate short-term prediction of glucose levels after a meal (often termed *postprandial glucose*) is an important capability for digital health systems, supporting better management of diabetes and pre-diabetes and providing insight into glycemic variability among healthy individuals. Postprandial glucose dynamics arise from the interplay of insulin secretion, carbohydrate absorption, and energy expenditure, all of which vary widely between and within individuals [1], [2]. Advances in wearable technologies, particularly Continuous Glucose Monitoring (CGM), now enable continuous observation of glucose in free-living conditions. The availability of such data has encouraged the use of machine learning for predictive modeling of short-term glucose trends [3], [4], [5].

Traditional machine learning methods have been applied to glucose prediction tasks with encouraging results. Bertachi et al. (2020) [6] compared several algorithms for predicting nocturnal hypoglycemia and reported that Support Vector Machines achieved the best sensitivity and specificity among global models. Alkalifah et al. (2024) [7] evaluated multiple regression methods including decision trees, boosting ensembles, and Gaussian Process Regression, showing that tree-based and ensemble models achieved low error for near-term glucose fluctuations. Bergford et al. (2023) [8] examined hypoglycemia risk prediction during exercise and found Random Forest slightly outperformed logistic regression in balanced accuracy. Similarly, Kladov et al. (2024) [9] demonstrated that combining clustering of glucose patterns with Random Forest or Gradient Boosting improved nocturnal glucose prediction, underscoring the value of structured temporal representations. Collectively, these studies confirm that even relatively simple algorithms, when properly tuned, can yield clinically useful predictions.

Personalization has emerged as another promising direction. Neumann et al. (2025) [10] compared ensemble methods with deep learning for Type 1 diabetes patients and showed that personalized Random Forest and XGBoost models achieved lower RMSE than global deep learning baselines. Their results emphasized the importance of tailoring models to subgroups or individuals rather than relying solely on population-level training. This perspective aligns with broader work in ensemble learning, where stacked frameworks have been used to refine predictions in other medical domains. For instance, Ren et al. (2022) [11] applied a stacked ensemble model for ICU mortality prediction, showing that integrating heterogeneous learners and subgroup-specific features improved accuracy compared with single-model

approaches. Recent reviews on ensemble methods in healthcare point to the ability of hybrids and stacks to capture heterogeneous patterns [12], [13]. Given diagnosis-dependent physiology, this motivates evaluating subgroup-specific models for glucose prediction, as done in our study.

Despite these advances, two challenges remain. First, few studies have systematically examined the contribution of different feature sets for postprandial prediction. Many prior works rely primarily on CGM traces and occasionally include activity data, but the role of meal composition, which is a primary driver of postprandial glucose excursions, is often underexplored [2], [4]. Second, while personalization has been attempted at the individual level, there is limited investigation into subgroup-level refinements based on diagnosis categories such as Healthy, Pre-diabetes, and Type 2 Diabetes. Diagnosis categories capture clinically relevant physiological differences while still providing sufficient data aggregation, potentially balancing personalization with generalization [14], [15], [16]. Without addressing these gaps, existing models offer limited guidance for meal-specific glucose forecasting in heterogeneous clinical populations.

In this study, we address these gaps through a two-stage investigation. We first evaluate feature combinations across several global models to identify the most effective input representation. Experiments combining CGM, activity, and nutritional features demonstrate that baseline glucose values together with meal composition yield the strongest predictive performance across datasets and horizons. This result underscores the value of explicitly modeling dietary macronutrient inputs rather than relying solely on glucose trends or activity data.

Building on this foundation, we present a stacked learning framework that pairs a single global regressor with diagnosis-specific refinements. The global model is trained on all participants with the feature configuration identified in our preliminary study. Its predictions are then appended as inputs to models trained separately for Healthy, Pre-diabetes, and Type 2 Diabetes. This preserves a shared representation while allowing each subgroup model to adjust for diagnosis-related physiology. We consider three global learners: Random Forest, Ridge Regression, and CatBoost. CatBoost serves as the refinement model for each subgroup. Experiments span 30-, 60-, and 120-minute horizons on two CGM datasets.

The contributions are threefold. First, a systematic comparison demonstrates that combining baseline glucose statistics with meal composition provides the most effective feature representation for meal-anchored prediction. Second, a stacked framework that integrates global and diagnosis-specific models is introduced and evaluated, yielding consistent gains on Dexcom and horizon-dependent gains on Libre, with the largest benefits at longer horizons and for linear global baselines. Third, subgroup analyses reveal that the value of diagnosis-aware refinement depends on diagnosis category and forecast distance, offering guidance on when to deploy refinement in practice.

Taken together, structured feature evaluation and diagnosis-aware stacking offer a practical path for predicting glucose levels after meals. Rather than relying only on deep architectures or fully individualized personalization, the combination of targeted features and subgroup refinement delivers meaningful accuracy gains while remaining scalable for real-world digital health systems.

2. Research Method

Our methodology follows a two-stage experimental design (Figure 1). In the first stage, we conduct feature engineering and base model evaluation to identify the optimal feature combinations. We prepare data from continuous glucose monitoring (CGM) alongside activity, meal, and demographic information, then generate multiple feature sets reflecting different data modalities (base features, demographic features, activity-related features, meal-related features, and various combinations thereof). We train and evaluate several individual regression models, which include Ridge Regression, Support Vector Regression, Random Forest (RF), and CatBoost, to determine which feature combination yields the best predictive performance. The best feature set is carried forward to the second stage.

In the second stage, we evaluate a stacked ensemble architecture designed to capture both population-level patterns and diagnosis-specific variations. We first train a global model on the full dataset, then develop three specialized models tailored to distinct glycemic subgroups: Healthy individuals, Pre-diabetic individuals, and those with Type 2 Diabetes, all using CatBoost as the underlying algorithm. We compare the performance of single models against stacked configurations across three prediction horizons (30, 60, and 120 minutes) to assess whether stratified, diagnosis-aware modeling improves forecast accuracy. The following subsections detail the datasets, feature engineering pipeline, single model and stacking model scheme, and evaluation metrics employed in this investigation.

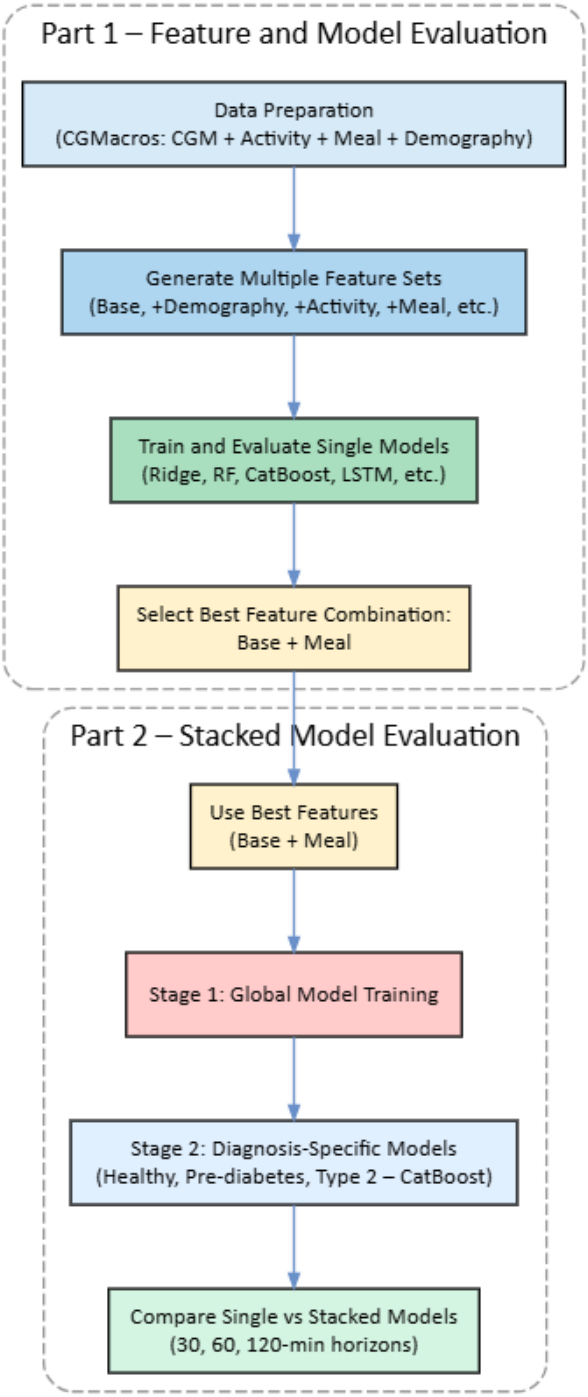

Figure 1. Research Overview

2.1 Datasets

This study uses the CGMacros v1.0.0 dataset [17], which integrates Continuous Glucose Monitoring (CGM) data with dietary and activity annotations under free-living conditions. CGM recordings were obtained from two commercial sensors, Dexcom and Libre, enabling evaluation across heterogeneous device platforms.

Each meal event in the dataset is accompanied by detailed annotations of macronutrient composition (carbohydrates, protein, and fat) as well as activity features including heart rate, estimated calories, and metabolic equivalents (METs). These annotations are time-aligned with CGM records, making it possible to track glucose excursions after eating. For this study, we focused on predicting postprandial glucose at **30, 60, and 120 minutes** following meals.

To incorporate clinical heterogeneity, participants were categorized into three diagnostic groups: **Healthy, Pre-diabetes, and Type 2 Diabetes**. After preprocessing and alignment, the Dexcom and Libre subsets each contained more than 1600 meal events, distributed relatively evenly across the three groups. Table 1 summarizes the final dataset used in our experiments.

Table 1. Dataset Summary After Preprocessing

Dataset	Total meal events	Healthy	Pre-diabetes	Type 2 Diabetes
Dexcom	1684	626	609	449
Libre	1705	633	617	455

The dataset thus provides a balanced and multimodal foundation for modeling postprandial glucose responses. The presence of both sensor types allows us to assess the robustness of predictive models across devices, while diagnosis-based grouping supports the development of subgroup-aware refinements.

2.2 Feature Engineering

Feature engineering was designed to represent both the physiological state preceding each meal and the static characteristics of participants. Continuous glucose monitoring (CGM), activity records, and meal logs were time-aligned to the meal timestamp (t_0). Each observation thus summarized the context immediately before food intake. Missing or irregular samples were linearly interpolated after resampling to one-minute intervals to maintain synchronization between signals. For every meal, the feature vector included glucose statistics, recent activity summaries, macronutrient composition, and demographic information. Continuous variables were standardized using z-score normalization, and categorical ones were label-encoded.

From the CGM signal, short-term statistics were extracted to describe pre-meal glucose trends and baselines. Activity features were derived from heart-rate, metabolic-equivalent (METs), and activity-energy (ActCalories) data streams, summarizing both instantaneous exertion and its lingering effects over several windows (15, 30, 60, 120 minutes preceding the meal). Nutritional features quantified the major dietary components of each logged meal (carbohydrates, protein, fat, fiber, and total calories) which determine the magnitude and duration of post-meal glucose excursions. Participant-specific attributes, including age, body-mass index (BMI), gender, and diagnosis category (Healthy = 0, Pre-diabetes = 1, Type 2 Diabetes = 2), provided additional contextual variability relevant to glucose metabolism.

To evaluate how these modalities contribute to predictive performance, features were grouped into several combinations tested in Section 3.1. Table 2 presents all engineered variables, their data source, derivation method, and physiological purpose. Together they form a multimodal representation capturing glucose dynamics, physical activity, dietary intake, and individual differences.

Table 2. Engineered Features with Derivation and Purpose

Feature name	Source	Computation / Description	Physiological purpose
GL_mean_T15	CGM	Mean glucose over 15 min preceding t_0	Captures short-term trend before eating
GL_mean_T30	CGM	Mean glucose over 30 min preceding t_0	Reflects pre-meal glycemic stability
GL_preMeal	CGM	Glucose value at t_0	Baseline glucose immediately before intake
HR_current	Activity	Heart rate at t_0	Indicates momentary exertion level
METs_current	Activity	METs at t_0	Proxy for physical intensity
ActCalories_current	Activity	Activity-energy expenditure at t_0	Estimates metabolic output at meal time
HR_mean_T15	Activity	Mean HR during 15 min before t_0	Recent cardiovascular load
METs_mean_T15	Activity	Mean METs during 15 min before t_0	Short-term activity level
ActCalories_mean_T15	Activity	Mean energy expenditure during 15 min before t_0	Short-term energy output
HR_mean_T30	Activity	Mean HR during 30 min before t_0	Moderate-term activity intensity
METs_mean_T30	Activity	Mean METs during 30 min before t_0	Captures sustained effort

ActCalories_mean_T30	Activity	Mean calories burned during 30 min before t_0	Energy trend before meal
HR_mean_T60	Activity	Mean HR during 60 min before t_0	Longer pre-meal activity effect
METs_mean_T60	Activity	Mean METs during 60 min before t_0	Longer activity window
ActCalories_mean_T60	Activity	Mean calories burned during 60 min before t_0	Prolonged energy expenditure
HR_mean_T120	Activity	Mean HR during 120 min before t_0	Captures residual exertion over two hours
METs_mean_T120	Activity	Mean METs during 120 min before t_0	Long-term activity influence
ActCalories_mean_T120	Activity	Mean calories burned during 120 min before t_0	Long-term metabolic load
Carbs	Meal log	Carbohydrate grams per meal	Primary driver of post-meal glucose rise
Protein	Meal log	Protein grams per meal	Slows glucose absorption and prolongs response
Fat	Meal log	Fat grams per meal	Modulates gastric emptying and insulin need
Fiber	Meal log	Fiber grams per meal	Dampens glucose spike amplitude
Calories	Meal log	Total meal energy (kcal)	Overall metabolic load
Age	Demography	Participant age (years)	Reflects insulin sensitivity decline with age
BMI	Demography	Body-mass index (kg/m²)	Indicates metabolic status
Gender	Demography	Encoded 0/1	Captures sex-related hormonal differences
Diagnosis	Demography	Healthy = 0, Pre-DM = 1, T2D = 2	Represents metabolic classification influencing response

2.3 Single-Model Setup

Single-model experiments were designed to evaluate the predictive capability of different feature combinations and regression algorithms for short-term glucose forecasting. Each feature set defined in this stage extended the Base configuration with additional contextual information derived from activity, meal, and demographic data. To maintain comparability with the subsequent stacked-learning framework, all feature sets included the categorical variable Diagnosis (Healthy = 0, Pre-diabetes = 1, Type 2 Diabetes = 2). The content of each feature combination is summarized in Table 3.

Table 3. Feature Set Definition

Feature set name	Features included
Base	GL_mean_T15, GL_mean_T30, GL_preMeal, HR_current, METs_current, ActCalories_current,Diagnosis
Base + Extra Activity	**Base** + HR/METs/ActCalories means at T15, T30, T60, T120
Base + Meal	**Base** + Carbs, Protein, Fat, Fiber, Calories
Base + Demography	**Base** + Age, BMI, Gender
Base + Extra + Demography + Meal	All features combined

We tested six machine learning models to establish baselines and to identify suitable learners for stacked learning. The selection spanned linear (Ridge Regression[18]), kernel-based (SVR with RBF kernel [19]), ensemble bagging Random Forest [20]), and gradient boosting methods (CatBoost Regressor [21] and LightGBM Regressor [22]), ensuring coverage of different modeling paradigms. A mean-value predictor (Dummy Regressor) [23] was also included as a reference baseline.

All models were implemented in Python using scikit-learn [23], CatBoost [24], and LightGBM [25] libraries. Training and validation followed a five-fold grouped cross-validation procedure, where meals from the same participant

were kept within a single fold to prevent leakage. This design ensures that performance reflects generalization to unseen individuals rather than repeated exposure to the same participant's data. Hyperparameters were set to default values or lightly tuned based on validation performance within each fold. Table 4 summarizes key hyperparameters for each model. Evaluation metrics included MAE, RMSE, and NRMSE [26], calculated per fold and averaged across folds.

Table 4. Models and Key Hyperparameters

Model	Library	Key Hyperparameters
Ridge Regression	scikit-learn	*alpha=1.0, random_state=42*
Support Vector Regression (SVR)	scikit-learn	*kernel="rbf", C=1.0, epsilon=0.1*
Random Forest	scikit-learn	*n_estimators=100, random_state=42, n_jobs=-1*
CatBoost Regressor	CatBoost	*iterations=500, learning_rate=0.01, depth=6, l2_leaf_reg=3, random_state=42*
LightGBM Regressor	LightGBM	*default boosting params, random_state=42*
Dummy Regressor	scikit-learn	*strategy="mean"*

Although six models were evaluated in the global feature-set experiments, only three (i.e Ridge, Random Forest, and CatBoost) were carried forward into the stacked learning framework. These represent linear, bagging, and boosting strategies, respectively. CatBoost was additionally chosen as the Stage 2 learner (subgroup-specific) in stacked models because of its robustness in handling non-linear tabular features and relatively small subgroup sizes.

2.4 Stacking Framework and Evaluation Protocol

The stacked-learning framework was designed to improve predictive accuracy by combining a global regressor with diagnosis-specific refinements. This implementation follows a two-stage sequential design applied within each cross-validation split. Stage 1 learns a global representation of glucose dynamics across all participants, while Stage 2 fine-tunes the predictions within each diagnosis category using additional contextual information from the Stage 1 output.

Training procedure (within each CV fold):

1. **Stage 1 (Global Model):** Train a global model on training fold (1-4) using the best feature configuration from Section 2.3 (Base + Meal) to predict glucose at 30-, 60-, and 120-minute horizons. Generate predictions on the training data and concatenate them with the original features to form extended inputs for Stage 2.
2. **Stage 2 (Diagnosis-Specific Models):** Partition the training data by diagnosis group (Healthy, Pre-diabetes, Type 2 Diabetes). For each group, train a separate CatBoost model on the extended features (Base + Meal + Stage 1 prediction). The *Diagnosis* variable is excluded at this stage because data partitioning already defines each subgroup explicitly.
3. **Evaluation:** Apply the trained Stage 1 model to the held-out fold (fold 5) to generate predictions. Route these predictions to the corresponding Stage 2 subgroup model based on each participant's diagnosis to produce final outputs.

This design preserves training-testing separation while enabling diagnosis-aware refinements within each split. CatBoost was chosen for Stage 2 because of its ability to capture nonlinear feature interactions with minimal tuning and its resilience to small sample sizes. The complete workflow is summarized in the Figure 2 below.

Performance was assessed using Normalized RMSE (NRMSE), which is computed by dividing RMSE by the mean observed glucose value, enabling comparisons across horizons and sensors with different glucose distributions.

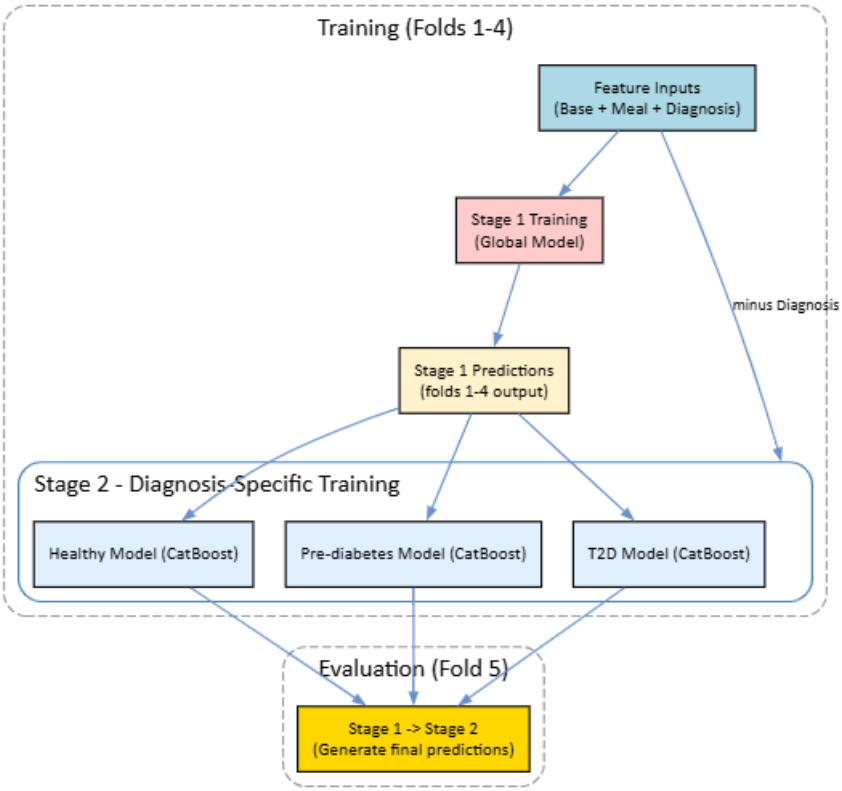

Figure 2. Stacking Framework

3. Results and Discussion

3.1 Global Model Feature Evaluation and Benchmarks

The first set of experiments jointly examined two dimensions: the effect of different feature sets and the relative performance of six machine learning models. Each model was trained under five-fold grouped cross-validation on multiple feature configurations, producing a grid of results. For clarity, we present the findings from two complementary perspectives: (i) the contribution of feature sets, averaged across models, and (ii) the relative accuracy of different models when using the best feature set.

Table 5 summarizes the mean NRMSE values across sensors and horizons, aggregated across all models. **Base + Meal** consistently achieved the lowest error in both Dexcom and Libre subsets. In contrast, adding demographic variables or extended activity features provided no advantage and in some cases worsened performance. The superiority of Base + Meal is consistent across sensors and horizons, underscoring the predictive value of meal macronutrients. This finding aligns with clinical understanding: carbohydrate intake is the primary driver of glucose excursions, while protein and fat modulate the timing and shape of the response.

Table 5. Mean NRMSE of Global Models with Different Feature Sets

Feature set	Dexcom T30	Dexcom T60	Dexcom T120	Libre T30	Libre T60	Libre T120
Base	0.210 ± 0.022	0.286 ± 0.029	0.303 ± 0.033	0.241 ± 0.015	0.331 ± 0.021	0.329 ± 0.033
Base + Demography	0.216 ± 0.019	0.299 ± 0.023	0.311 ± 0.031	0.245 ± 0.016	0.340 ± 0.028	0.332 ± 0.035
Base + Extra Activity	0.210 ± 0.021	0.287 ± 0.026	0.305 ± 0.030	0.243 ± 0.015	0.329 ± 0.019	0.327 ± 0.033
Base + Extra + Demography + Meal	0.218 ± 0.012	0.294 ± 0.024	0.302 ± 0.025	0.241 ± 0.016	0.310 ± 0.022	0.308 ± 0.039
Base + Meal	**0.209 ± 0.014**	**0.275 ± 0.021**	**0.288 ± 0.024**	**0.236 ± 0.017**	**0.308 ± 0.026**	**0.309 ± 0.040**

Next, we focused on the relative performance of models using the best-performing feature set (Base + Meal). Table 6 summarizes the NRMSE for six learners across prediction horizons, averaged across both Dexcom and Libre datasets. The Dummy model (mean predictor) performed worst, as expected. Ridge produced stable results, but was outperformed by tree-based methods at all horizons. Random Forest and CatBoost consistently achieved the lowest errors across short and long term horizons. LightGBM performance was intermediate between Ridge and the top-performing tree models. SVR underperformed substantially, showing higher errors and less stability compared to tree-based learners, likely due to its sensitivity to hyperparameter tuning and feature scaling. The results confirm that ensemble tree-based methods are well-suited for tabular glucose prediction tasks, capturing nonlinear relationships between pre-meal glucose, meal composition, and postprandial outcomes without requiring extensive feature engineering or deep architectures.

Table 6. Global Model Performance (NRMSE) with Base + Meal Composition Features

Regression Method	T30	T60	T120
Random Forest	**0.186 ± 0.008**	**0.265 ± 0.020**	**0.272 ± 0.029**
CatBoost	0.189 ± 0.010	0.264 ± 0.016	0.274 ± 0.025
LightGBM	0.196 ± 0.011	0.275 ± 0.019	0.284 ± 0.032
Ridge	0.199 ± 0.010	0.279 ± 0.017	0.286 ± 0.020
SVR	0.273 ± 0.025	0.328 ± 0.035	0.333 ± 0.043
Dummy (mean)	0.292 ± 0.027	0.340 ± 0.033	0.340 ± 0.043

3.2 Stacked vs Global Models

We compared diagnosis-aware stacking against single-stage global models across both datasets and all horizons. Each configuration is written as **Global→Stage 2**, where Stage 2 is always a CatBoost regressor trained within the corresponding diagnosis group. Both global and stacked configurations use the same feature basis (Base +Meal), ensuring that gains reflect the two-level design rather than input difference.

Averaged across all 18 configurations (3 global learners × 3 horizons × 2 sensors), stacking improved NRMSE by 3.47% ± 3.67 (mean ± SD). Gains were larger and more consistent on Dexcom than on Libre, and increased with forecast horizon. On Dexcom, mean improvements were 5.27% (T30), 6.15% (T60), and 7.94% (T120). On Libre, corresponding values were 0.65% (T30), 2.32% (T60), and 3.11% (T120). These patterns suggest that subgroup-specific refinement becomes more valuable as the forecast target moves further from the meal, where physiology and behavior yield more complex responses not fully captured by a single global fit. Table 7 presents detailed results for all configurations, and Figure 3 illustrates the improvement trend for each stacking configuration across prediction horizons.

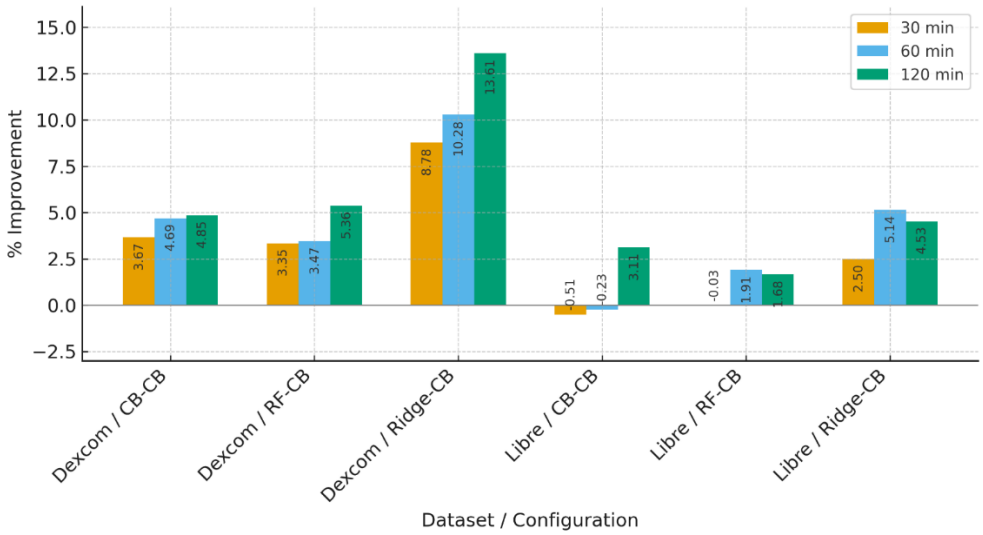

Figure 3. Improvement of Each Stacking Model by Prediction Horizon

Table 7. Comparison of Stacking Model Performance to Global Models (NRMSE)

Dataset	T	Model	NRMSE (Global)	NRMSE (Stage 2)	% Improvement
Dexcom	30	CB-CB	0.1756	0.1692	3.67%
Dexcom	30	RF-CB	0.1729	0.1671	3.35%
Dexcom	30	Ridge-CB	0.1956	0.1785	8.78%
Dexcom	60	CB-CB	0.2501	0.2384	4.69%
Dexcom	60	RF-CB	0.2508	0.2421	3.47%
Dexcom	60	Ridge-CB	0.2711	0.2432	10.28%
Dexcom	120	CB-CB	0.2647	0.2519	4.85%
Dexcom	120	RF-CB	0.2638	0.2496	5.36%
Dexcom	120	Ridge-CB	0.2860	0.2471	13.61%
Libre	30	CB-CB	0.2032	0.2042	-0.51%
Libre	30	RF-CB	0.1990	0.1991	-0.03%
Libre	30	Ridge-CB	0.2028	0.1978	2.50%
Libre	60	CB-CB	0.2777	0.2784	-0.23%
Libre	60	RF-CB	0.2796	0.2743	1.91%
Libre	60	Ridge-CB	0.2867	0.2720	5.14%
Libre	120	CB-CB	0.2836	0.2748	3.11%
Libre	120	RF-CB	0.2808	0.2761	1.68%
Libre	120	Ridge-CB	0.2852	0.2723	4.53%

On Dexcom, stacking improves performance for all three global learners at every horizon reported in Table 7. At 30 minutes, mean improvement across the three global models is about 5.3 percent, rising to about 6.2 percent at 60 minutes and about 7.9 percent at 120 minutes. The largest single gain occurs at 120 minutes for Ridge→CatBoost, where NRMSE drops from 0.2860 to 0.2471, a 13.61 percent reduction. Random Forest and CatBoost global models also benefit, although more modestly, with typical improvements of 3 to 5 percent. This horizon-dependent widening suggests that subgroup-specific refinement becomes more valuable as the forecast target moves further from the meal, where physiology and behavior yield more complex responses that are not fully captured by a single global fit.

Libre shows a more tempered pattern. At 30 minutes, CB→CB and RF→CB exhibit small negative changes of −0.51 percent and −0.03 percent, while Ridge→CB improves by 2.50 percent. By 60 minutes, the picture becomes positive overall, with gains of 1.91 percent for RF→CB and 5.14 percent for Ridge→CB, and a negligible −0.23 percent for CB→CB. At 120 minutes, all three configurations improve, ranging from 1.68 percent to 4.53 percent. Averaged across all models, Libre improvements rise from about 0.7 percent at 30 minutes to about 2.3 percent at 60 minutes and about 3.1 percent at 120 minutes. The weaker and sometimes negative changes at 30 minutes are consistent with higher short-horizon variability and potential timestamp noise in food logs and sensor alignment, which can dilute the value of a second-stage correction when the global estimate is already near the measurement limit.

The stacked design helps most when the Stage 1 model is simpler. Ridge→CatBoost yields the largest average gains on both datasets and all horizons, roughly 10.9 percent on Dexcom and about 4.1 percent on Libre when averaged per dataset and horizon. This pattern indicates that Stage 2 CatBoost is capturing nonlinear residual structure that a linear global model leaves unexplained. When Stage 1 is already a strong nonlinear learner, as with CatBoost or Random Forest, stacking still helps but the marginal improvements are smaller, typically in the 1 to 5 percent range depending on dataset and horizon. These observations align with the conceptual goal of stacking, which is to provide a corrective layer that focuses on the remaining bias rather than duplicating capacity already present in the base learner.

Two practical implications follow from the results in Table 6 and the trend lines in Figure 4. First, diagnosis-aware refinement is most attractive for horizons at or beyond 60 minutes, where the average gain is larger and more stable across datasets. This is a relevant window for meal planning and safety warnings, so even single-digit percentage improvements can be meaningful in real use. Second, if computational budget or latency constraints require a single global model, CatBoost remains a strong choice; however, when a linear global model is preferred for simplicity or interpretability, adding the Stage 2 CatBoost per diagnosis group can recover much of the lost accuracy at longer horizons.

Finally, the few negative or flat outcomes on Libre at 30 minutes highlight conditions where stacking may not add value. In such cases, model selection could adapt to the data regime by enabling the Stage 2 pass only when the forecast horizon is long enough or when validation metrics indicate sufficient residual structure to exploit. Overall, averaged over all eighteen configurations, stacking improves NRMSE by about 4.2 percent, with the largest and most consistent benefits appearing on Dexcom and at longer horizons, as illustrated in Figure 4.

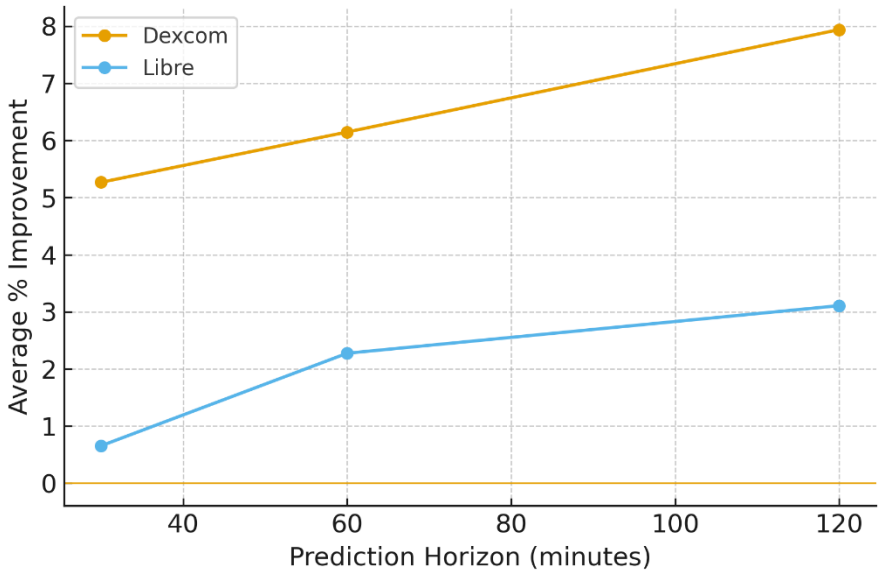

Figure 4. Improvement of Sensor Dataset for Each Prediction Horizon

3.3 Sub-group level analysis

To assess how stacked learning affects different populations, we analyzed performance separately for Healthy, Pre-diabetes, and Type 2 Diabetes groups. Figures 5a and 5b visualize the percentage improvement of stacked models over global baselines across prediction horizons, revealing several consistent patterns alongside notable exceptions.

Several consistent patterns emerge. Healthy participants exhibit steady gains that generally increase with horizon length, though with important sensor-specific differences. For Dexcom, improvements rise cleanly from 30 to 120 minutes across most configurations, with Ridge-CB showing strong lifts (13.47% → 9.92% → 12.31%). Libre data follow a similar upward trend for RF-CB and CB-CB pairings (e.g., CB-CB: 6.10% → 11.40% → 6.49%). However, Libre–Healthy–Ridge-CB presents a clear outlier: performance degrades consistently across all horizons (−6.03% at 30 min, −5.44% at 60 min, −6.87% at 120 min), indicating a systematic failure mode where the linear global baseline leaves residuals that are either too small, biased, or noisy for CatBoost to model effectively. Rather than correcting prediction errors, the refiner appears to amplify variance or introduce miscalibration. This sensor-by-diagnosis-by-model interaction suggests that Ridge's regularization, combined with Libre's noise characteristics in healthy glucose ranges, produces residuals that lack the structured signal required for effective second-stage refinement.

Type 2 Diabetes groups show the clearest and most consistent benefits. Improvements peak sharply at 60 minutes, particularly for Ridge-CB on both Dexcom (15.61%) and Libre (15.03%), before moderating slightly at 120 minutes (13.20% and 11.30%, respectively). This mid-horizon advantage likely reflects the refiner's ability to capture diagnosis-specific postprandial dynamics, where delayed glucose peaks and prolonged absorption create structured residuals that a stratified CatBoost model can exploit. By 120 minutes, increased heterogeneity from meal composition variability, physical activity, and sensor drift broadens the residual distribution, softening but not eliminating the marginal gains.

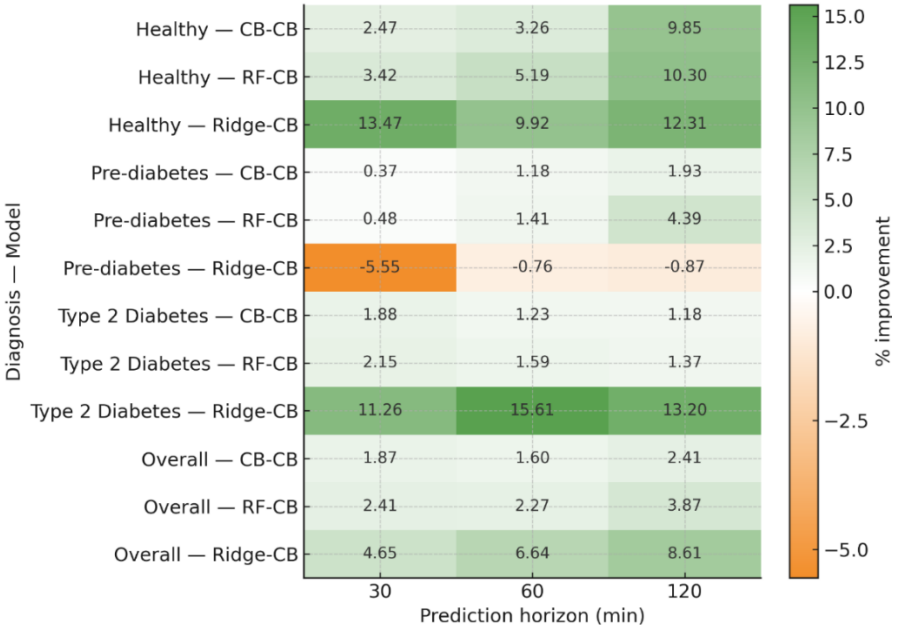

Figure 5a. Dexcom Dataset: Subgroup-level Improvements in NRMSE (%)

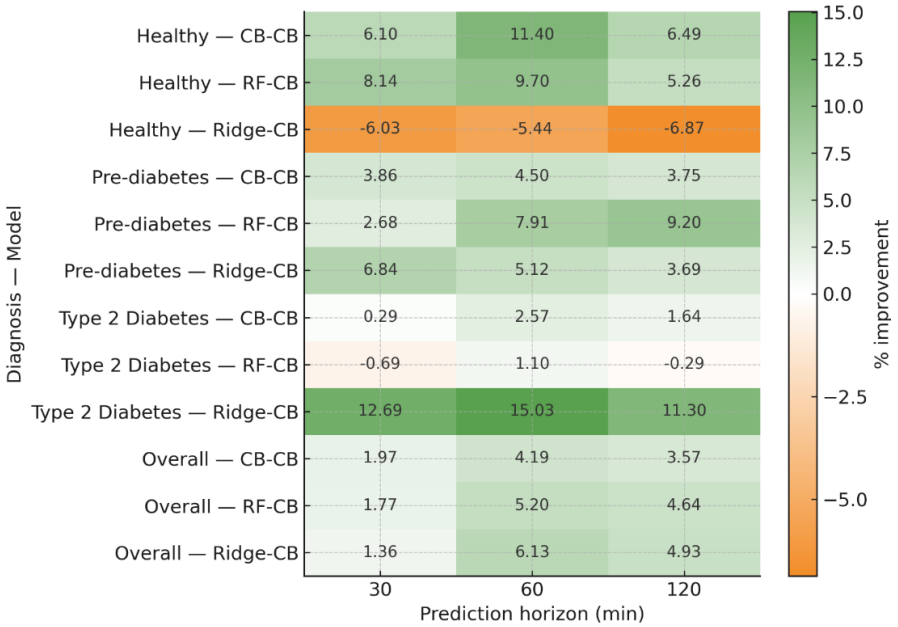

Figure 5b. Libre Dataset: Subgroup-level Improvements in NRMSE (%)

The Pre-diabetes group exhibits the weakest and most variable improvements. Short-horizon performance on Dexcom hovers near zero or dips negative (e.g., Ridge-CB: −5.53% at 30 minutes), consistent with a well-calibrated baseline where refinement risks overfitting transient noise. As the horizon extends to 60–120 minutes, gains become more visible, especially on Libre (e.g., RF-CB: 7.91% at 60 minutes, 9.20% at 120 minutes), indicating that residual structure emerges with time even in this physiologically transitional category. The high variability suggests that Pre-diabetes may encompass heterogeneous metabolic states that resist coarse stratification.

Overall, stacked models deliver broadly positive results, with the strongest and most stable improvements observed in Healthy (excluding the Libre–Ridge-CB anomaly) and Type 2 Diabetes cohorts, and with Dexcom exhibiting more consistent lifts. The identified failure mode (i.e. Libre–Healthy–Ridge-CB at short horizons) highlights a sensor-by-diagnosis interaction where insufficient or biased residuals lead the refiner to overcorrect. In practice, we recommend stacked models for horizons ≥60 minutes, validating configurations per sensor, and substituting RF-CB or CB-CB for

Ridge-CB when deploying on Libre–Healthy populations. Modest regularization (e.g., early stopping, learning rate shrinkage) in the refiner stage may further guard against the observed negative deltas.

3.4 Interpretations and Practical Implications

The results demonstrate that both feature selection and model architecture critically influence postprandial glucose prediction accuracy. The identification of **Base + Meal** as the optimal feature set aligns with clinical physiology: carbohydrate intake drives glucose excursions, while protein and fat modulate response timing and magnitude. Similar conclusions have emerged in prior nutritional and CGM-based studies [6], [7]. Notably, demographic features such as age or BMI provided no measurable benefit once diagnosis status was incorporated, suggesting that coarse physiological groupings (Healthy, Pre-diabetes, Type 2 Diabetes) capture the most relevant long-term metabolic differences without requiring granular individual descriptors.

The analysis of stacked learning highlights the value of diagnosis-aware stratification. Improvements were most pronounced at 60–120-minute horizons, where prediction difficulty increases and group-level physiology diverges more sharply. These findings resonate with Neumann et al. (2025) [10], who demonstrated that personalized models outperform population-based approaches for Type 1 Diabetes management. Our work extends this principle by showing that diagnosis-level stratification offers a practical middle ground: more specific than a single global model, yet less data-intensive than fully individualized personalization.

To further contextualise our results, Table 8 places the best-performing configuration in this study against representative CGM prediction studies from prior literature. Because no published study to our knowledge directly targets postprandial glucose prediction using meal macronutrient features on a comparable dataset, the comparison is necessarily indirect. Prior works address fasting, nocturnal, or continuous glucose forecasting under different input modalities and patient populations. This gap is itself meaningful: it reflects the novelty of the postprandial prediction task and motivates the methodological choices made in this study. Whereas prior CGM prediction studies largely rely on single global models and underexplore nutritional context, our findings not only quantify the critical contribution of meal macronutrients, but also demonstrate how a diagnosis-aware stacked learning framework can further refine these predictions, offering a practical and scalable pathway for personalized glucose management systems.

Table 8. Indirect Comparison with Representative CGM-based Glucose Prediction Studies

Study	Task	Input Features	Model	Horizon	Metric
Bertachi et al. (2020) [6]	Nocturnal hypoglycemia event prediction	CGM only	SVM	6 hour sleep period	Accuracy 80.77%
Alkalifah et al. (2024) [7]	CGM prediction, and glycemic classification	CGM, body temperature, heart rate, blood pressure	CatBoost / GPR	continuous	RMSE 40.6 mg/dL
Kladov et al. (2024) [9]	Nocturnal glucose	Nocturnal CGM	RF / GBM	15 and 30 mins	RMSE ~0.68-1.20
Neumann et al. (2025) [10]	T1D glucose (personalized)	CGM + meal + activity	XGBoost/LSTM	5–30 min	RMSE ~7.46-17.74
This study (Ridge→CB, Dexcom)	**Postprandial glucose**	**CGM + meal macronutrients**	**Stacked (Ridge→CatBoost)**	**30, 60, 120 min**	**NRMSE ~0.21–0.25**

Subgroup results reveal that benefits are not uniform across populations. Healthy participants showed strong gains at longer horizons (10–13% for Dexcom Ridge-CB at 120 minutes), where subgroup adjustments correct persistent deviations. Type 2 Diabetes groups benefited most at 60 minutes (up to 15–16% improvement), reflecting their distinct postprandial kinetics. The mixed and occasionally negative results in Pre-diabetes suggest this category may be too heterogeneous for binary stratification, which is an observation echoed in studies exploring clustering-based patient segmentation [9], [27]. For this intermediate group, finer clinical stratification (e.g., HbA1c subcategories, insulin resistance indices) or adaptive individual-level modeling may be necessary.

Several limitations warrant acknowledgment. The dataset, while multimodal, remains moderate in size, and subgroup analyses occasionally exhibited instability, particularly for Pre-diabetes. The stacking framework relies on diagnosis categories as a proxy for physiology, which may oversimplify inter-individual metabolic variation. Additionally, only tree-based ensemble learners (Ridge, Random Forest, CatBoost) were explored; incorporating recurrent neural networks (LSTMs, GRUs) or attention-based architectures (Transformers) could more explicitly model temporal

dependencies and meal-glucose coupling dynamics. Future work should refine subgroup definitions using continuous metabolic markers, investigate adaptive or hierarchical stratification strategies, and test hybrid architectures that integrate global, group-level, and individual layers.

In summary, these results establish that careful feature engineering combined with diagnosis-aware subgroup refinement provides a reproducible pathway to more accurate postprandial glucose forecasting. They demonstrate that meaningful improvements can be achieved without requiring deep personalization or large-scale data collection, bridging the gap between research prototypes and deployable digital health tools.

4. Conclusion

This study demonstrates that both feature representation and model architecture critically determine postprandial glucose prediction accuracy. A systematic evaluation established that combining baseline glucose measures with meal composition provides the most informative feature set, outperforming alternatives that incorporated demographic attributes or extended activity summaries. Building on this configuration, we implemented a stacked learning framework coupling a global regressor with diagnosis-specific refinement models. Across datasets and horizons, stacking consistently reduced NRMSE relative to single global models, with the most substantial gains observed on Dexcom data and at 60- to 120-minute horizons. Subgroup analyses revealed differential benefits: Healthy participants showed gains that generally increased or remained strong at longer horizons (e.g., Dexcom Ridge-CB: 13.47% at 30 min, 12.31% at 120 min), while Type 2 Diabetes participants exhibited peak improvements at 60 minutes (up to 15.61% on Dexcom Ridge-CB) before moderating at 120 minutes. Pre-diabetes outcomes were highly variable, reflecting metabolic heterogeneity within this transitional category. These findings confirm that diagnosis-aware adaptation captures residual physiological structure that single global models leave unexplained, particularly when the base learner is linear.

From a practical perspective, diagnosis-aware stacking offers a scalable pathway to improved accuracy without requiring fully individualized models, retaining efficiency by training one global model and a small set of subgroup refiners while delivering meaningful error reductions where decision support is most critical. Future work should validate these findings in larger cohorts, explore adaptive subgrouping or continuous phenotype embeddings, evaluate recurrent or attention-based architectures that integrate meal timing and composition into richer temporal representations, and establish decision rules to enable or bypass Stage 2 refinement based on horizon length and validation residuals. In summary, this work establishes that careful feature engineering combined with diagnosis-stratified refinement provides a reproducible framework for more accurate postprandial glucose forecasting, bridging the gap between research prototypes and deployable clinical applications.

Acknowledgement

This work was funded by Universitas Lambung Mangkurat (Contract No. 1868/UN8/LT/2025).

References

[1] D. Zeevi *et al.*, “Personalized Nutrition by Prediction of Glycemic Responses,” *Cell*, vol. 163, no. 5, pp. 1079–1094, Nov. 2015. https://doi.org/10.1016/j.cell.2015.11.001

[2] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, “Continuous Glucose Monitoring Sensors for Diabetes Management: A Review of Technologies and Applications.,” *Diabetes Metab J*, vol. 43, no. 4, pp. 383–397, Aug. 2019. https://doi.org/10.4093/dmj.2019.0121

[3] M. Vettoretti, G. Cappon, A. Facchinetti, and G. Sparacino, “Advanced Diabetes Management Using Artificial Intelligence and Continuous Glucose Monitoring Sensors,” *Sensors*, vol. 20, no. 14, p. 3870, Jul. 2020. https://doi.org/10.3390/s20143870

[4] T. Zhu, L. Kuang, K. Li, J. Zeng, P. Herrero, and P. Georgiou, “Blood Glucose Prediction in Type 1 Diabetes Using Deep Learning on the Edge,” in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May 2021, pp. 1–5. https://doi.org/10.1109/ISCAS51556.2021.9401083

[5] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, “Data-driven blood glucose level prediction in type 1 diabetes: a comprehensive comparative analysis,” *Sci Rep*, vol. 14, no. 1, p. 21863, Sep. 2024. https://doi.org/10.1038/s41598-024-70277-x

[6] A. Bertachi *et al.*, “Prediction of Nocturnal Hypoglycemia in Adults with Type 1 Diabetes under Multiple Daily Injections Using Continuous Glucose Monitoring and Physical Activity Monitor,” *Sensors*, vol. 20, no. 6, p. 1705, Mar. 2020. https://doi.org/10.3390/s20061705

[7] B. Alkalifah, M. T. Shaheen, J. Alotibi, T. Alsubait, and H. Alhakami, “Evaluation of machine learning-based regression techniques for prediction of diabetes levels fluctuations,” *Heliyon*, vol. 11, no. 1, p. e41199, Jan. 2025. https://doi.org/10.1016/j.heliyon.2024.e41199

[8] S. Bergford *et al.*, “The Type 1 Diabetes and EXercise Initiative: Predicting Hypoglycemia Risk During Exercise for Participants with Type 1 Diabetes Using Repeated Measures Random Forest,” *Diabetes Technol Ther*, vol. 25, no. 9, pp. 602–611, Sep. 2023. https://doi.org/10.1089/dia.2023.0140

[9] D. E. Kladov, V. B. Berikov, J. F. Semenova, and V. V. Klimontov, “Machine Learning Algorithms Based on Time Series Pre-Clustering for Nocturnal Glucose Prediction in People with Type 1 Diabetes,” *Diagnostics*, vol. 14, no. 21, p. 2427, Oct. 2024. https://doi.org/10.3390/diagnostics14212427

[10] A. Neumann, Y. Zghal, M. A. Cremona, A. Hajji, M. Morin, and M. Rekik, “A Data-Driven Personalized Approach to Predict Blood Glucose Levels in Type-1 Diabetes Patients Exercising in Free-Living Conditions,” 2024. https://doi.org/10.2139/ssrn.4777350

[11] N. Ren, X. Zhao, and X. Zhang, “Mortality prediction in ICU Using a Stacked Ensemble Model,” *Comput Math Methods Med*, vol. 2022, pp. 1–12, Nov. 2022. https://doi.org/10.1155/2022/3938492

[12] M. Z. Wadghiri, A. Idri, T. El Idrissi, and H. Hakkoum, “Ensemble blood glucose prediction in diabetes mellitus: A review,” *Comput Biol Med*, vol. 147, p. 105674, Aug. 2022. https://doi.org/10.1016/j.compbiomed.2022.105674

[13] A. Alotaibi, “Ensemble Deep Learning Approaches in Health Care: A Review,” *Computers, Materials & Continua*, vol. 82, no. 3, pp. 3741–3771, 2025. https://doi.org/10.32604/cmc.2025.061998

[14] J. Song, T. J. Oh, and Y. Song, “Individual Postprandial Glycemic Responses to Meal Types by Different Carbohydrate Levels and Their Associations with Glycemic Variability Using Continuous Glucose Monitoring,” *Nutrients*, vol. 15, no. 16, p. 3571, Aug. 2023. https://doi.org/10.3390/nu15163571
[15] B. M. Ahmed, M. E. Ali, M. M. Masud, M. R. Azad, and M. Naznin, “After-meal blood glucose level prediction for type-2 diabetic patients,” *Heliyon*, vol. 10, no. 7, p. e28855, Apr. 2024. https://doi.org/10.1016/j.heliyon.2024.e28855
[16] S. Hotta, M. Kytö, S. Koivusalo, S. Heinonen, and P. Marttinen, “Optimizing postprandial glucose prediction through integration of diet and exercise: Leveraging transfer learning with imbalanced patient data,” *PLoS One*, vol. 19, no. 8, p. e0298506, Aug. 2024. https://doi.org/10.1371/journal.pone.0298506
[17] R. Gutierrez-Osuna, D. Kerr, B. Mortazavi, and A. Das, “CGMacros: a scientific dataset for personalized nutrition and diet monitoring,” *Scientific Data (under review)*, 2025.
[18] H. Šinkovec, G. Heinze, R. Blagus, and A. Geroldinger, “To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets,” *BMC Med Res Methodol*, vol. 21, no. 1, p. 199, Dec. 2021. https://doi.org/10.1186/s12874-021-01374-y
[19] Y. Hu *et al.*, “Support Vector Regression Model for Determining Optimal Parameters of HfAlO-Based Charge Trapping Memory Devices,” *Electronics (Basel)*, vol. 12, no. 14, p. 3139, Jul. 2023. https://doi.org/10.3390/electronics12143139
[20] G. N., P. Jain, A. Choudhury, P. Dutta, K. Kalita, and P. Barsocchi, “Random Forest Regression-Based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes,” *Processes*, vol. 9, no. 11, p. 2095, Nov. 2021. https://doi.org/10.3390/pr9112095
[21] B. Chen, Y. Chen, and H. Chen, “An Interpretable CatBoost Model Guided by Spectral Morphological Features for the Inversion of Coastal Water Quality Parameters,” *Water (Basel)*, vol. 16, no. 24, p. 3615, Dec. 2024. https://doi.org/10.3390/w16243615
[22] A. D. Hartanto, Y. Nur Kholik, and Y. Pristyanto, “Stock Price Time Series Data Forecasting Using the Light Gradient Boosting Machine (LightGBM) Model,” *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 4, p. 2270, Dec. 2023. https://doi.org/10.62527/joiv.7.4.1740
[23] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jun. 2018.
[24] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, in NIPS’18. Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 6639–6649. https://doi.org/10.48550/arXiv.1706.09516
[25] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.”
[26] G. Liu, L. Brooks, J. Canty, D. Lu, J. Y. Jin, and J. Lu, “Deep-NCA: A deep learning methodology for performing noncompartmental analysis of pharmacokinetic data,” *CPT Pharmacometrics Syst Pharmacol*, vol. 13, no. 5, pp. 870–879, May 2024. https://doi.org/10.1002/psp4.13124
[27] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, “Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM),” *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021. https://doi.org/10.3390/diagnostics11091714