

ORIGINAL RESEARCH ARTICLE

Few-shot personalized postprandial glucose prediction with model-agnostic neighbor fusion

Fatma Indriani*, **Irwan Budiman**, **Dwi Kartini**, **Radityo Adi Nugroho**,
Muhammad Zainal Muttaqin, and **Mohammad Reza Faisal**

Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat, Banjarmasin, South Kalimantan, Indonesia

Abstract

The postprandial glucose response varies substantially among individuals, so a model trained on a population predicts well on average yet makes consistent, person-specific errors for individual participants. Personalization can close this gap, but most methods require retraining or a personal meal history that a newly enrolled subject has not yet accumulated. This study aimed to personalize an arbitrary pre-trained glucose model at inference time, without retraining, gradient updates, or diagnosis labels. We propose Neighbor Fusion Personalization (NFP), a model-agnostic inference-time layer that combines a population-model prediction with two nearest-neighbor estimates. The global estimate uses similar meals from the training population and is available at cold start, whereas the personal estimate uses the subject's accumulating meal history. An adaptive weight, $w(n) = n/(n + \lambda)$, gradually shifts the prediction from population-level toward personal evidence as the available history increases. Across 44 subjects and 5 base architectures spanning linear, tree-ensemble, and neural models, NFP reduced normalized root mean square error for glucose 60 min after a meal at every tested checkpoint from three meals onward, reaching a mean improvement of 0.032 (about 13%, approximately 5 mg/dL) at 5 meals, significant on all 5 architectures ($p_{\text{fdr}} \leq 0.01$) with no per-model tuning. It outperformed subject-specific calibration and a mixed-effects baseline in the few-shot regime, and added about 0.11 ms and under 25 kB per prediction. On two independent cohorts spanning type 1 and type 2 diabetes, the direction of the effect was reproduced under frozen hyperparameters, with smaller gains that scaled with meal-feature fidelity. Because it is training-free and model-agnostic, NFP offers a practical personalization layer for continuous glucose monitoring, available from the first prediction. Its clinical value for dietary decision support remains to be tested prospectively.

*Corresponding author:

Fatma Indriani
(f.indriani@ulm.ac.id)

Citation: Indriani F, Budiman I, Kartini D, Nugroho RA, Muttaqin MZ, Faisal MR. Few-shot personalized postprandial glucose prediction with model-agnostic neighbor fusion. *Artif Intell Health*. doi: 10.36922/AIH026270074

Received: June 29, 2026

Revised: July 5, 2026

Accepted: July 13, 2026

Published online: September 9, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Continuous glucose monitoring; Postprandial glucose prediction; Personalization; Model-agnostic personalization; K-nearest neighbors; Few-shot learning; Inference-time adaptation

1. Introduction

Blood glucose rises after every meal, and the magnitude of that rise, known as the postprandial response, varies substantially from person to person. Two individuals eating the same food can produce glucose trajectories that differ by more than 50 mg/

dL.¹ This variation is not noise: it reflects stable metabolic differences that persist across repeated meal exposures and across days of monitoring. As the prevalence of type 2 diabetes and prediabetes continues to rise worldwide², accurate prediction of postprandial responses may help guide dietary choices and manage long-term health risk, since the size and duration of glucose excursions are themselves linked to vascular complications.³

Continuous glucose monitors (CGMs) now enable meal-level trajectories to be captured at scale, and machine learning models trained on population data predict postprandial glucose with reasonable average accuracy.^{1,4–10} However, average accuracy can conceal systematic subject-specific bias: a model may consistently overpredict for some participants and underpredict for others. The error is not random scatter around the truth but a person-specific offset that a population model cannot remove, because the model has, by construction, averaged over exactly the individual differences that produce it.

Personalization is the natural remedy for this systematic bias, but it demands individual meal data that a new subject has not yet produced. The way in which existing methods obtain and use such data, and the trade-offs associated with those methods, define the gap addressed in this study.

Most personalization methods adapt to a new subject by changing model parameters. Meta-learning is the dominant route, learning an initialization that can be fine-tuned rapidly using a small number of subject-specific observations^{11–13}, building on model-agnostic meta-learning and metric-based few-shot methods from the broader literature.^{14–17} Transfer and fine-tuning methods reuse a source model for a target subject, whether by matching similar source subjects before transfer¹⁸ or by fine-tuning a general network on a few samples from the new subject¹⁹, and multitask formulations share information across participants to reduce the amount of subject-specific data required.²⁰ These approaches can achieve high personalized predictive accuracy once sufficient subject-specific data are available. The adaptation step, however, requires gradient-based parameter updating for each new participant, and incremental schemes additionally require a minimum history before that update stabilizes.²¹ These requirements restrict such methods to differentiable models and introduce participant-specific computational costs; moreover, stable updating may not be possible immediately after enrollment.

Methods that avoid retraining take two further routes. Case-based retrieval matches an incoming meal to similar historical records and uses their observed outcomes directly^{22,23}, an idea that also underlies case-based decision support for insulin dosing, including temporal variants

that retrieve short event sequences and adaptive variants that retain patient history to track drift over time.^{24,25} Useful as these systems are, the retrieved neighbors are essentially the whole predictor, so the method has nothing to fall back on when a subject has no history yet.

The second route begins from a population prior and calibrates it toward the individual online. Data-assimilation methods update subject-specific states and a few physiological parameters as self-monitoring data arrive, in both offline and online forms^{26,27}, while recursive linear models refresh their parameters every few minutes from the subject's own trace.²⁸ Mixed-effects and adaptive-selection models carry the same partial-pooling logic into clinical time series, placing greater weight on population-level information when subject-specific data are scarce and greater weight on individual information as those data accumulate.^{29,30} These methods initially shrink estimates toward the population and place greater weight on the individual as subject-specific data accumulate. Neighbor Fusion Personalization (NFP) follows a similar general intuition, but the existing approaches are tied to particular model structures and cannot wrap an arbitrary black-box predictor without rederivation.

Retrieval-augmented forecasting has recently emerged as an inference-time alternative to retraining, building on retrieval-augmented generation in language models.^{31,32} In general time-series work, methods such as ReTime, RAE, TimeRAE, and TS-RAG retrieve related series or motifs and fuse them with the query, improving forecasts without updating model weights.^{33–36} GlyRAG carries the idea into CGM forecasting, retrieving similar glucose episodes from an embedding space and fusing them with the query before prediction.³⁷ These methods establish that retrieval can augment inference without modifying parameters, yet they target zero-shot forecasting rather than subject-level adaptation, and they do not separate the retrieval pool into cross-subject and within-subject sources. The combination of cross-subject and within-subject neighbor prediction under a weighting scheme that responds to accumulating personal data remains unaddressed. Separately, recent work has pursued interpretability in personalized glucose modeling, using Shapley-value analysis to explain reinforcement-learning policies for glucose monitoring.³⁸

Across these families, the prior literature establishes three capabilities that have rarely been brought together: a fixed population predictor, cross-subject retrieval that is available during cold start, and a within-subject memory that accumulates without retraining. [Table 1](#) contrasts the dominant method families along five properties that matter at deployment: whether adaptation retrains the model, whether it needs accumulated personal history, whether it

can predict during cold start, whether it wraps an arbitrary base model, and whether it adapts at the subject level. Each family leaves at least one property unmet. As summarized in Table 1, NFP combines in a single inference-time framework the deployment-relevant properties not jointly provided by prior work. Parameter-based methods retrain per subject and cannot provide subject-specific adaptation before personal history is available; case-based methods lean entirely on retrieved cases with little to fall back on at cold start; and mixed-effects and assimilation methods are each tied to a particular model structure. Retrieval-augmented forecasting can support cold-start prediction without retraining, yet it targets generic zero-shot prediction rather than subject-level adaptation and rarely separates cross-subject from within-subject evidence. NFP combines the three capabilities in one inference-time layer, fusing a fixed population predictor, a cross-subject retrieval prior, and an accumulating within-subject memory under a history-dependent weighting rule, with no retraining and no diagnosis labels.

In computational terms, this is a cold-start personalization problem, where a model trained on many subjects must serve a new individual before that person has logged enough meals for conventional personalization. This paper makes four contributions:

- (i) We propose NFP, a training-free, model-agnostic layer that fuses a population model with cross-subject (global k-nearest-neighbor prediction [Global-kNN]) and within-subject (Personal-kNN) neighbors under an adaptive weighting schedule, with no retraining or diagnosis labels.
- (ii) A chronological few-shot protocol measured personalization gain at successive history checkpoints across the cold-start regime.
- (iii) Evaluation and component comparisons across five architectures on CGMacros¹ showed statistically

significant normalized root mean square error (NRMSE) gains at every tested checkpoint from $n = 3$ onward ($n = 3, 5, 10$ and 20), and the gain generalizes across all three prediction horizons and across the three metabolic groups in the dataset.

- (iv) Two independent cohorts, Shanghai T2DM and HUPA-UCM, showed personalization gains under frozen NFP hyperparameters, with a magnitude that scales with meal-feature fidelity.

2. Methods

2.1. Problem formulation

We consider a population training set $\mathcal{D}_\text{tr} = \{(x_i, y_i)\}$ drawn from a group of subjects, where each x_i is a per-meal feature vector and y_i is the observed postprandial glucose at a fixed horizon. A population model f_θ is fitted once on $\mathcal{D}_\text{tr}$, and its parameters θ remain fixed thereafter; we write its prediction for a meal x as $\hat{y}_\text{pop} = f_\theta(x)$. At deployment, an unseen test subject produces meals in chronological order, $(x_1, y_1), \dots, (x_T, y_T)$. To measure personalization at a controlled history size, we adopt a fixed-shot setting. At history size n , the first n chronological meals form the personal history $\mathcal{H}_n = \{(x_i, y_i)\}_{i=1}^n$, and every later meal of the same subject is evaluated as a query against this fixed support set. The history is therefore frozen at n meals within a checkpoint, no query outcomes never enter $\mathcal{H}_n$, and the case $n = 0$ is the cold-start condition, where no personal outcome is available. The objective is to predict glucose at horizons of 30, 60, and 120 min for each query meal.

Personalization here is framed as an inference-time adaptation operator $A(f_\theta, x_q, \mathcal{H}_n)$ that adjusts the prediction for a query meal x_q using the personal history, without modifying θ . This separates the present setting from conventional few-shot learning, where the few examples are used to estimate new model parameters. NFP estimates

Table 1. Comparison of method families for postprandial glucose personalization according to five deployment-relevant characteristics

Method family	Retraining	Personal history	Cold-start support	Model-agnostic personalization	Subject-level adaptation
Fine-tuning, meta-learning ^{11–13,18,19}	Required	Usually required	Limited	No	Yes
Case-based retrieval ^{22–25}	No	Usually required	Weak	Usually yes	Yes
Mixed-effects and data-assimilation methods ^{26–30}	Online updating	Required	Population prior	No	Yes
Time-series RAG ^{33–37}	No	Not necessarily personal	Yes	Usually model-specific	Usually no
NFP (this work)	No	Optional initially	Yes	Yes	Yes

Abbreviations: NFP: Neighbor Fusion Personalization; RAG: Retrieval-augmented generation.

no parameters from $\mathcal{H}_n$; the personal meals act as an inference-time memory, retrieved at prediction time and never folded into the model itself. The operator uses only same-subject meals drawn from $\mathcal{H}_n$, which enforces subject-level and chronological isolation between support and query.

2.2. Dataset and preprocessing

The CGMacros dataset¹ contains meal records from 45 subjects across 3 metabolic groups (healthy controls, individuals with prediabetes, and those with type 2 diabetes). Each subject wore a Dexcom G6 Pro CGM (Dexcom, United States of America) for an average of 9.4 ± 1.5 days, with readings recorded at five-minute intervals. A total of 1,301 meal events are annotated with macronutrient content (carbohydrates, protein, fat, and energy content in kilocalories), meal start time, and participant characteristics and clinical variables including age, body mass index, glycated hemoglobin (HbA1c), and gender. Meal counts range from 1 to 54 per subject (median 29). One subject with only one recorded meal was excluded from evaluation, leaving 44 evaluable subjects. Table 2 summarizes the dataset composition.

Table 2. CGMacros dataset composition by metabolic group¹

Group	Subjects	Meals	Glucose@60, mean $\pm$ SD (mg/dL)
Healthy	15	469	138 $\pm$ 24
Prediabetic	16	448	154 $\pm$ 29
Type 2 diabetic	14	384	192 $\pm$ 40
Total	45	1,301	–

Note: One participant with only one recorded meal was excluded from the evaluation.

For each meal event, a 60-min pre-meal CGM window comprising 12 readings at 5-min intervals and ending immediately before meal onset was extracted, and the last reading before onset was also retained as a scalar (pre-meal glucose). Combined with four meal-composition features, two cyclic meal-time features defined as $\sin(2\pi h/24)$ and $\cos(2\pi h/24)$ where h denotes the hour of day, and participant characteristics and clinical variables (age, body mass index, HbA1c, and gender), this forms a 23-dimensional feature vector per meal (Table 3). NFP draws its neighbors from two subsets of these features. The global subset (macronutrient variables and pre-meal

glucose, five dimensions) keeps only meal-level descriptors expected to generalize across participants, which is what makes cross-subject matching meaningful. The personal subset (CGM window, macros, pre-meal glucose, and meal time, 19 dimensions) adds the subject's metabolic state at onset but drops demographics, which do not vary within a participant and therefore cannot distinguish among that participants' meals. All features are z-normalized per fold using training-set statistics only, so no test-fold information enters the normalization.

Table 3. Per-meal feature subsets for population models and NFP neighbor prediction

Feature	Dimension	Full	Global	Personal
Pre-meal CGM window	12	✓		✓
Carbohydrates (g)	1	✓	✓	✓
Protein (g)	1	✓	✓	✓
Fat (g)	1	✓	✓	✓
Energy (kcal)	1	✓	✓	✓
Pre-meal glucose	1	✓	✓	✓
Meal time, sine/cosine encoded	2	✓		✓
Age	1	✓		
BMI	1	✓		
HbA1c	1	✓		
Gender (binary)	1	✓		
Total		23	5	19

Abbreviations: BMI: Body mass index; HbA1c: Glycated hemoglobin; NFP: Neighbor Fusion Personalization.

Whether personalization can improve prediction depends on how much of the variation in postprandial glucose reflects stable, subject-specific differences rather than patterns shared across the population. A between-versus within-subject variance decomposition (Table 4) shows that 45–49% of the variance in the raw glucose targets is between-subject across the three horizons, which is the share a personalized model can exploit and which sets a ceiling on achievable gains. For the meal-to-meal excursion, however, the variance distribution is reversed: 78–86% of the variance lies within subjects, indicating that this response component is largely meal-specific and therefore less amenable to personalization. The mean within-subject standard deviation of the 60-min excursion ranges from about 18 to 38 mg/dL across diagnosis groups, which sets a floor on the error any method can reach.

Table 4. Between- and within-subject shares of variance in raw glucose targets and glucose excursions in the CGMacros dataset at each prediction horizon, estimated using a one-way subject-level variance decomposition

Horizon (min)	Raw target		Excursion	
	Between	Within	Between	Within
30	45	55	14	86
60	49	51	22	78
120	49	51	15	85

2.3. Population baseline models

Five models serve as population baselines: ridge regression (B1), extreme gradient boosting (XGBoost; B2), a bidirectional long short-term memory (BiLSTM; B3), a multilayer perceptron (MLP; B4), and random forest (RF; B5). They span linear, tree, and neural architectures to ensure personalization gains are model-agnostic. All five use the same 23 input variables, represented according to each architecture, and produce a separate output for blood glucose at 30, 60, and 120 min after meal onset. All five are trained on training-fold subjects with no subject-specific adaptation. Their predictions for the test subjects serve as the reference against which personalization gains are measured. NFP does not modify or retrain these models; instead, it adjusts their predictions at inference time using nearest-neighbor prediction (Figure 1). For reference, a naive persistence baseline that carries the pre-meal glucose forward yields NRMSE@60 of 0.342, which all five population models substantially outperformed. Table 5 summarizes the architectures and key hyperparameters.

Table 5. Population baseline models and key hyperparameters

ID	Model	Key hyperparameters
B1	Ridge regression	RidgeCV; $\alpha \in \{0.01, 0.1, 1, 10, 100\}$, selected by leave-one-out CV; fit separately per horizon.
B2	XGBoost	200 trees, max depth 6, learning rate 0.1; early stopping after 20 rounds on a held-out validation set (20% of training subjects, stratified by diagnosis).
B3	BiLSTM	BiLSTM(32 units/direction) on CGM sequence; static branch MLP(11 $\rightarrow$ 32 $\rightarrow$ 16); fusion MLP(80 $\rightarrow$ 64 $\rightarrow$ 32) with Dropout(0.3); three linear heads.
B4	MLP	Linear(23 $\rightarrow$ 64 $\rightarrow$ 32 $\rightarrow$ 16) with ReLU and Dropout(0.3); three linear heads.
B5	RF	500 trees, min_samples_leaf = 5, OOB score; fit separately per horizon.

Abbreviations: BiLSTM: Bidirectional long short-term memory; CGM: Continuous glucose monitor; CV: Cross-validation; MLP: Multilayer perceptron; OOB: Out-of-bag; ReLU: Rectified linear unit; RF: Random forest; XGBoost: Extreme gradient boosting.

2.3.1. Neural model architectures

The BiLSTM (B3) reads the CGM window as a length-12 sequence through a BiLSTM (32 units per direction) and processes the eleven static features (macros, pre-meal glucose, meal time, and demographics) in a parallel two-layer branch. The two branches are concatenated and passed through a fusion network before three independent linear heads predict the 30, 60, and 120-min targets. The MLP (B4) instead flattens all 23 features into a single vector and applies a three-layer network feeding the same heads. Layer dimensions are listed in Table 5.

2.3.2. Training setup

Both neural models were trained with AdamW ($\eta = 10^{-3}$, weight decay 10^{-4}) under cosine annealing ($T_{max} = 200$, $\eta_{min} = 10^{-5}$) and a batch size of 64. Early stopping monitors validation root mean square error (RMSE) at 60 min with a patience of 20 and a maximum of 200 epochs. Output head biases are initialized to the per-horizon training mean. Each horizon has a separate output, while the neural models share their internal representations across the three horizons. Ridge regression (B1) and RF (B5) were fitted on all training-fold subjects; XGBoost (B2) and the neural models use an 80/20 within-fold split (stratified by diagnosis) as a validation set for early stopping.

2.4. Neighbor fusion personalization

At deployment, a test subject begins with no personal meal history. For each incoming meal, the population model produces a prediction $\hat{y}_{pop}$ based on population-level patterns. The goal of NFP is to improve this prediction using the subject's own records as they accumulate. At prediction time for a given meal, the available personal history is $\mathcal{H}_n = \{(x_i, y_i)\}_{i=1}^n$, where x_i is the feature vector and y_i the observed glucose at 60 min for the i -th preceding meal from the same subject, ordered chronologically.

2.4.1. Global-kNN prediction

Global-kNN searches the training population for meals most similar to the meal being predicted. Its k_g nearest neighbors are found using Euclidean distance on the global feature subset, and the cross-subject prediction is the inverse-distance-weighted mean of those neighbors' true outcomes, with larger weights assigned to neighbors at shorter Euclidean distances, following the classical nearest-neighbor rule³⁹:

$$\hat{y}_g = \frac{\sum_{j \in \mathcal{N}_g} w_j y_j}{\sum_{j \in \mathcal{N}_g} w_j}, w_j = \frac{1}{d_j + \epsilon} \quad (1)$$

where $\mathcal{N}_g$ is the set of k_g nearest neighbors, d_j is the

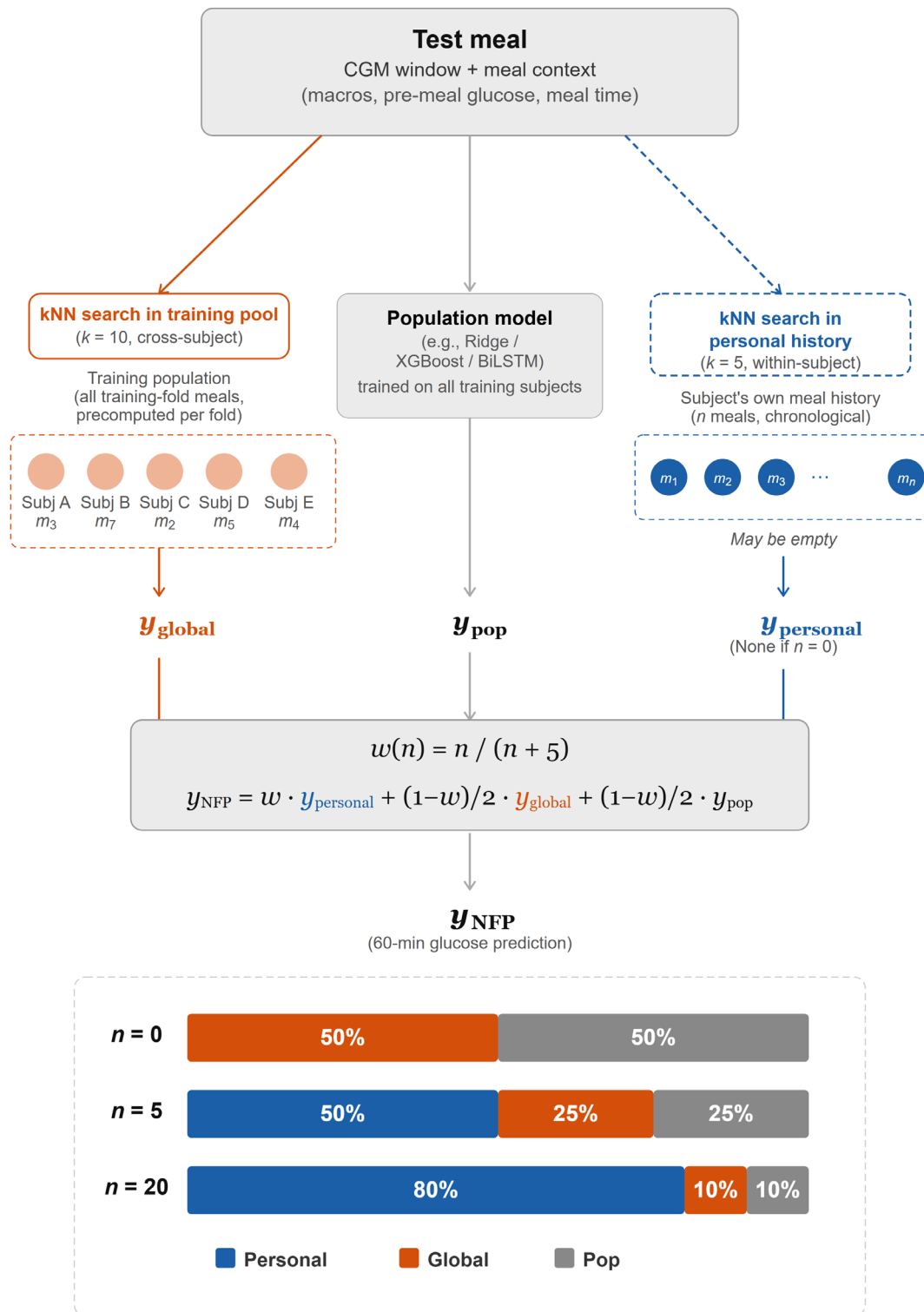


Figure 1. NFP framework combining the population-model, Global-kNN, and Personal-kNN predictions using a history-dependent adaptive weighting scheme, $w(n)$

Abbreviations: BiLSTM: Bidirectional long short-term memory; CGM: Continuous glucose monitor; kNN: k-nearest neighbors; NFP: Neighbor Fusion Personalization; Pop: Population; Ridge: Ridge regression; Subj: Subject; XGBoost: Extreme gradient boosting.

Euclidean distance to neighbor j , and $\epsilon = 10^{-6}$ prevents division by zero. Because these neighbors come only from training data, the index is built once per fold. The standalone full-feature Global-kNN comparator uses the full 23-dimensional feature vector with $k = 20$, so its weak performance is not an artifact of feature starvation. This comparator differs in feature set and neighbor count from the global component embedded in NFP, which uses only the five-dimensional subset of Table 3 as a cold-start anchor ($k_g = 10$), so it bounds the value of cross-subject retrieval rather than isolating the embedded component exactly.

2.4.2. Personal-kNN prediction

Personal-kNN applies the same neighbor search to the subject's own history instead of the training population. The personal prediction $\hat{y}_p$ uses Equation 1, with neighbors drawn from $\mathcal{H}_n$ on the personal feature subset. When $n < k_p$, all available personal meals are used, and when $n = 0$ no personal prediction exists.

2.4.3. Adaptive weighting

The three signals are combined under a schedule that depends only on how many personal meals are available so far:

$$w(n) = \frac{n}{n + \lambda} \quad (2)$$

$$\hat{y}_{\text{NFP}} = w\hat{y}_p + \frac{1-w}{2}\hat{y}_g + \frac{1-w}{2}\hat{y}_{\text{pop}} \quad (3)$$

Algorithm 1. NFP inference for one meal

```

1: Input: query  $x$ , population prediction  $\hat{y}_{\text{pop}}$ , training index  $(X^{\text{tr}}_g, y^{\text{tr}})$ ,
   personal history  $H_n$ , parameters  $k_g, k_p, \lambda$ 
2:  $w \leftarrow n/(n+\lambda)$ 
3:  $\hat{y}_g \leftarrow \text{InvDistKNN}(x, X^{\text{tr}}_g, y^{\text{tr}}, k_g)$  // global subset
4: if  $n > 0$  then
5:    $\hat{y}_p \leftarrow \text{InvDistKNN}(x, H_n, k_p)$  // personal subset
6: else
7:    $\hat{y}_p \leftarrow 0$ 
8: end if
9:  $\hat{y}_{\text{NFP}} \leftarrow w \cdot \hat{y}_p + ((1-w)/2) \cdot \hat{y}_g + ((1-w)/2) \cdot \hat{y}_{\text{pop}}$ 
10: return  $\hat{y}_{\text{NFP}}$ 

```

The two non-personal signals, Global-kNN and the population model, split the remaining weight $1 - w$ equally, acting as a joint prior that fades as personal history grows. At $n = 0$ this prior carries all the weight and the output is an equal blend of Global-kNN and the population model, the only evidence available before the subject has logged a meal. As meals accumulate, $w \rightarrow 1$ and Personal-kNN takes over, with λ setting how quickly the handover occurs. Hyperparameters were selected by grid search on the training folds; the final values are $k_g = 10$, $k_p = 5$, and $\lambda = 5$, with results stable for $\lambda \in \{3, 5, 10\}$. At $\lambda = 5$ the personal component carries 50% weight at $n = 5$ meals and 80% at $n = 20$. Algorithm 1 states the full inference path for a single meal.

2.4.4. Computation and deployment

Neighbor Fusion Personalization adds two nearest-neighbor searches to each prediction and introduces no trainable parameters. The global index is built once per fold from the training meals and reused for every test subject, so its cost is amortized across the cohort. Each query then ranks the 5-dimensional global subset over the training pool and the at most 20 meals of personal history on the 19-dimensional personal subset, both negligible against a single forward pass of the population model. Because the population model is queried only through its output, NFP stores nothing beyond the training feature matrix and the subject's own meals, so its added memory and computation remain modest relative to the base model.

2.5. Design rationale

Three design choices distinguish NFP from a single nearest-neighbor predictor, and each follows from the structure of the problem. The first is the use of separate feature spaces for the two retrieval sources. Global-kNN matches meals across subjects, so it keeps only meal-level descriptors that transfer between people, namely macronutrients and pre-meal glucose. Participant-level variables such as age and HbA1c help a population model place a subject within the cohort, but they are constant within a subject and cannot separate one of that subject's meals from another, so the personal subset omits them. Matching on a compact meal-level representation may also be more reliable than matching in the full feature space, where between-subject distances are dominated by demographic spread rather than meal content.

The second choice is the shared weight on the two non-personal signals. The population model and Global-kNN both describe the average response rather than the individual, so they act as a joint prior that fades together as personal evidence accumulates. The third choice is the gradual handover governed by λ . Rather than switching abruptly to Personal-kNN once any history exists, the weight $w(n) = n / (n + \lambda)$ raises personal trust smoothly with the number of observed meals, so λ behaves as a shrinkage parameter that sets how much personal evidence is trusted at each history size. This can be read as a bias-variance compromise. Personal-kNN may represent the individual more faithfully, but its estimate has high variance when only one or two meals are available, whereas the population-level estimates are stable yet retain subject-

specific bias. Adaptive fusion leans on the stable prior while personal data are scarce and on the personal estimate once it becomes reliable, echoing the partial-pooling behavior of mixed-effects and assimilation models^{26,29} without committing to their model structure.

2.6. Evaluation protocol

The evaluation protocol comprised subject-level cross-validation, chronological few-shot simulation, assessment of temporal drift, and subject-level statistical analysis.

2.6.1. Cross-validation

Subjects are assigned to folds as units, so no subject's meals appear in both training and test partitions. Each of the five folds holds approximately nine test subjects, with the remaining 35 forming the training set. Population models and the Global-kNN index are built from training-fold data exclusively.

2.6.2. Few-shot simulation

In sequential online deployment, the personal history grows after each observed outcome. For controlled evaluation, however, we freeze the history at predefined checkpoints to isolate the effect of history size, as illustrated in Figure 2. For each test subject, meals are sorted by time and indexed from 1 to T_s . At checkpoint n , the first n meals form the personal history $\mathcal{H}_n$ and all later meals are evaluated. The experiment repeats at $n \in \{0, 1, 2, 3, 5, 7, 10, 15, 20\}$, tracing how personalization gain changes as personal data accumulate. A subject is evaluated at checkpoint n only if at least five meals remain after the first n (i.e., $\geq n + 5$).

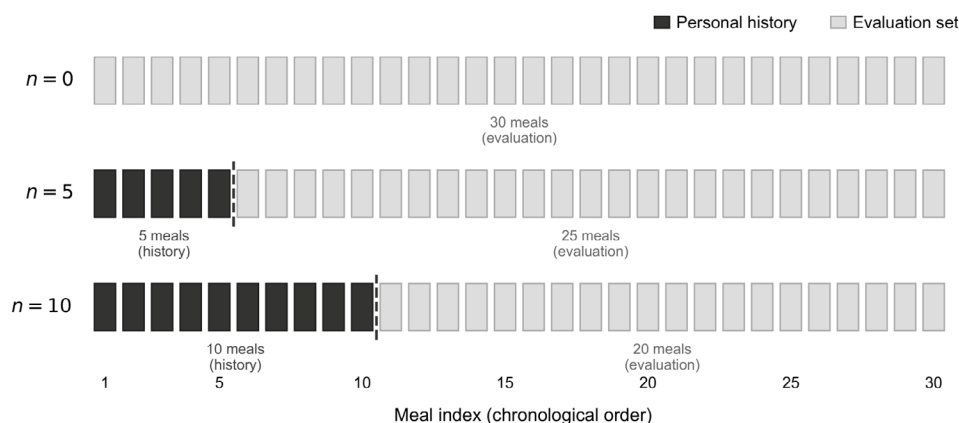


Figure 2. Few-shot evaluation protocol for a single test subject

recorded meals). At $n = 20$, five subjects fall below this threshold and are excluded (Table 6).

Table 6. Evaluable subjects and meals per history checkpoint

Personal-history checkpoint, n (meals)	Evaluable subjects	Evaluation meals
0	44	1,300
5	44	1,080
10	44	860
20	39	413

2.6.3. Temporal drift

Because the protocol uses each subject's earliest meals as history and evaluates later ones, the gain could in principle reflect later meals being easier to predict rather than personalization. This does not appear to be the case in CGMacros. The mean per-subject change in target@60 from early to late meals is small (+3.5 mg/dL), and only 7 of 44 subjects show an absolute drift above 20 mg/dL. The mean target@60 and its deviation from the subject mean are also stable across meal positions, so early meals are neither systematically higher nor noisier than later ones, which supports their use as history.

2.6.4. Metrics and significance

The primary metric is NRMSE improvement at the 60-min horizon (Δ NRMSE@60): for each test subject, the NRMSE of the personalizing method subtracted from that of the population model, averaged over evaluation meals at checkpoint n . Positive values indicate a gain over the population model. NRMSE is RMSE divided by the subject's mean glucose, which normalizes for the different baseline glucose levels across subjects, so the same absolute error counts for more in a subject who runs low than in one who runs high. Each subject belongs to one test fold, so subject-level scores are averaged over the held-out subjects pooled across the five folds, using all evaluable subjects at each checkpoint. Significance is assessed at $n = 3, 5, 10$, and 20 with a one-sided Wilcoxon signed-rank test, a non-parametric paired test suited to the small and skewed per-subject sample, with Benjamini-Hochberg correction controlling the false discovery rate across the five models at each checkpoint. We also report 95% bootstrap confidence intervals from 2,000 subject-level resamples.

2.6.5. Association between population-model bias and personalization gain

To evaluate whether personalization gain is associated with population-model error, we summarize each test subject

by two quantities at the 60-min horizon. The population-model bias is the absolute value of the mean signed residual, $|\overline{\hat{y}_{\text{pop}} - y}|$, over the subject's evaluation meals, which captures the systematic offset the population model leaves on that subject. The personalization gain is the subject's Δ NRMSE@60 at $n = 5$. Both quantities are averaged across the five base models so that each subject contributes a single bias-gain pair, computed from the same canonical NFP configuration used elsewhere ($k_g = 10$, $k_p = 5$, $\lambda = 5$). The association between bias and gain is measured by the Pearson correlation, with the Spearman rank correlation as a robustness check.

2.7. External-cohort validation

To test whether the NFP mechanism transfers beyond CGMacros, we repeated the evaluation on two independent public cohorts. The hyperparameters selected on CGMacros were kept fixed ($k_g = 10$, $k_p = 5$, $\lambda = 5$); within each external dataset, only the population models were retrained, under the same subject-wise five-fold cross-validation and few-shot protocol. No NFP parameter was retuned, so any gain reflects transfer rather than dataset-specific tuning, and the datasets were not pooled because device, sampling rate, clinical population, and meal annotation differ.

Shanghai T2DM⁴⁰ is the primary external cohort because it matches the type 2 diabetes motivation. It comprises 100 patients with glucose recorded every 15 min and free-text dietary records, for 3,013 evaluable meals across 109 monitoring periods. Meals were represented by coarse proxies parsed from the food descriptions (total recorded weight, number of items, and a staple-carbohydrate weight from keyword matching), together with pre-meal glucose and meal time, rather than measured macronutrients, and the pre-meal window used the four readings spanning the 60 min before onset. Each monitoring period was treated as a separate few-shot sequence, while all periods from one patient shared a cross-validation fold.

HUPA-UCM⁴¹ is a deliberately harder transfer setting: subjects with type 1 diabetes on insulin therapy, FreeStyle Libre glucose at five-minute resolution, and carbohydrate-only meal information with no demographics. Each subject was truncated to the first 14 days so that no single long recording dominated the retrieval pool, leaving 845 evaluable meals across 21 subjects, and the pre-meal window matched CGMacros (twelve 5-min readings). In both cohorts, the Global-kNN and Personal-kNN feature subsets followed the construction of Section 2.2 using the variables available, and population NRMSE was comparable across the five base models, so none was degenerate.

3. Results

3.1. Personalization across history size

Table 7 reports absolute NRMSE@60 for all four methods at seven history checkpoints. Global-kNN is model-independent, so its value repeats across the five baselines within a checkpoint, and Personal-kNN is undefined at $n = 0$. NFP achieves the lowest average NRMSE from $n = 0$ through $n = 15$. Personal-kNN matches NFP at $n = 15$ and slightly outperforms it at $n = 20$, as the contribution of personal history to the blended prediction increases. The analysis below centers on $n = 15$, where population RMSE ranges 38–44 mg/dL across architectures.

At five meals of history, NFP improves all five architectures, from +0.027 (RF) to +0.044 (XGBoost), for a mean of +0.032. This 13% relative error reduction (approximately 5 mg/dL) is significant across all five models ($p_{\text{fdr}} \leq 0.010$). Per-subject gains appear in Figure 3 (error bars: 95% bootstrap confidence interval).

Personal-kNN reaches a mean gain of +0.027 at $n = 5$ and is significant on all five models as well, but its confidence intervals are wider than those of NFP. The difference comes from the global anchor, which steadies the blend while personal history is still small.

The standalone full-feature Global-kNN comparator is near-zero or negative on four of five models, ranging from -0.006 to -0.002 . The exception is XGBoost (+0.014), where the larger residuals of the gradient-boosted model make neighbor-based correction marginally useful. These effect sizes indicate that cross-subject retrieval alone contributes little once a competent population model is in place.

At $n = 10$, NFP maintains its advantage, with a mean gain of +0.034 versus +0.032 for Personal-kNN (Figure 4).

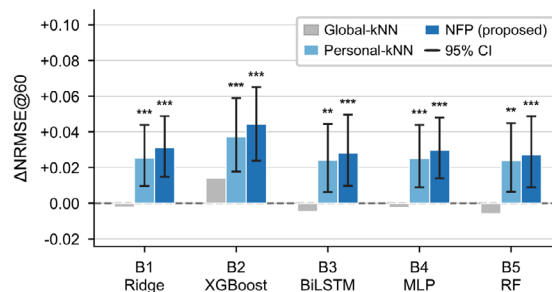
3.2. Learning curves

Figure 4 traces the mean gain over the population model as personal history grows from zero to twenty meals. Shaded bands show 95% bootstrap confidence intervals. Population and Global-kNN remain near zero at all n . Personal-kNN produces consistent gains from $n = 1$ onward, rising from +0.006 at $n = 1$ to +0.045 at $n = 15$. NFP leads at every checkpoint. At $n = 3$, NFP is statistically significant on all five architectures ($p_{\text{fdr}} \leq 0.01$ on three, $p_{\text{fdr}} \leq 0.05$ on two).

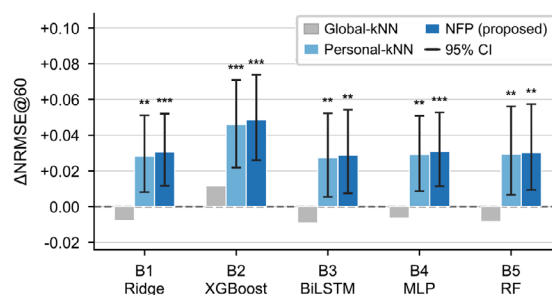
The advantage of NFP over Personal-kNN is concentrated in early history ($n \leq 5$). At $n = 0$, Personal-kNN is undefined and reverts to the population model, while NFP returns +0.005 via the Global-kNN blend. The gap is +0.006 at $n = 1$ (NFP +0.012 versus Personal-kNN +0.006) and holds near +0.005 through $n = 5$, before falling

to +0.002 at $n = 10$. From $n = 15$ onward, the two methods are within +0.001 of each other.

A



B



C

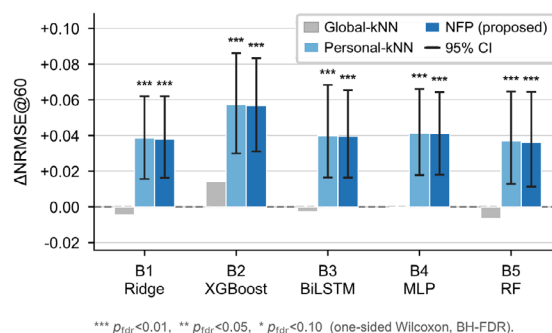


Figure 3. Δ NRMSE improvement per architecture: (A) $n = 5$ meals (44 subjects), (B) $n = 10$ meals (44 subjects), (C) $n = 20$ meals (39 subjects). Abbreviations: BiLSTM: Bidirectional long short-term memory; MLP: Multilayer perceptron; NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error; RF: Random forest; XGBoost: Extreme gradient boosting.

3.3. Consistency across base architectures

The personalization gain is not confined to a particular kind of base model. At $n = 5$, NFP lowers NRMSE@60 for all five architectures, and it does so for both the weakest and the strongest population model. XGBoost has the highest population error (0.248) and receives the largest absolute correction, while RF has the lowest population error (0.228) and the smallest, yet both improve. Ridge

Table 7. Absolute NRMSE@60 for the population and personalization methods across selected personal-history checkpoints

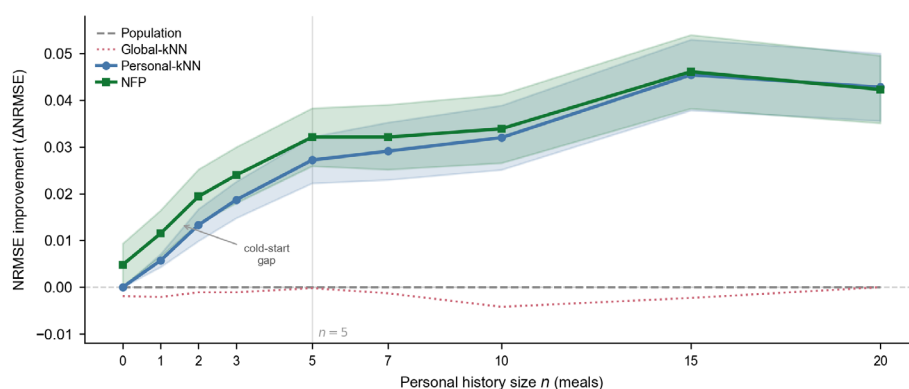
Model	Population	Global-kNN	Personal-kNN	NFP
<i>n</i> = 0 meals: cold start (44 subjects)				
B1 Ridge	0.237	0.241	–	0.231 ^a
B2 XGBoost	0.255	0.241 ^a	–	0.242
B3 BiLSTM	0.236	0.241	–	0.233 ^a
B4 MLP	0.237	0.241	–	0.234 ^a
B5 RF	0.232 ^a	0.241	–	0.233
Mean	0.240	0.241	–	0.235 ^a
<i>n</i> = 1 meal (44 subjects)				
B1 Ridge	0.238	0.242	0.232	0.226 ^a
B2 XGBoost	0.256	0.242	0.247	0.235 ^a
B3 BiLSTM	0.236	0.242	0.232	0.228 ^a
B4 MLP	0.238	0.242	0.234	0.229 ^a
B5 RF	0.233	0.242	0.227	0.226 ^a
Mean	0.240	0.242	0.234	0.229 ^a
<i>n</i> = 3 meals (44 subjects)				
B1 Ridge	0.234	0.236	0.217	0.210 ^a
B2 XGBoost	0.250	0.236	0.223	0.214 ^a
B3 BiLSTM	0.230	0.236	0.214	0.210 ^a
B4 MLP	0.232	0.236	0.215	0.210 ^a
B5 RF	0.228	0.236	0.212	0.209 ^a
Mean	0.235	0.236	0.216	0.211 ^a
<i>n</i> = 5 meals (44 subjects)				
B1 Ridge	0.232	0.234	0.206	0.201 ^a
B2 XGBoost	0.248	0.234	0.211	0.203 ^a
B3 BiLSTM	0.229	0.234	0.205	0.201 ^a
B4 MLP	0.231	0.234	0.206	0.202 ^a
B5 RF	0.228	0.234	0.204	0.201 ^a
Mean	0.234	0.234	0.206	0.201 ^a
<i>n</i> = 10 meals (44 subjects)				
B1 Ridge	0.224	0.232	0.196	0.194 ^a
B2 XGBoost	0.243	0.232	0.198	0.195 ^a
B3 BiLSTM	0.223	0.232	0.196	0.194 ^a
B4 MLP	0.225	0.232	0.196	0.194 ^a
B5 RF	0.223	0.232	0.194	0.193 ^a
Mean	0.228	0.232	0.196	0.194 ^a
<i>n</i> = 15 meals (44 subjects)				
B1 Ridge	0.223	0.229	0.181 ^a	0.181 ^a
B2 XGBoost	0.243	0.229	0.182	0.181 ^a
B3 BiLSTM	0.222	0.229	0.182	0.181 ^a
B4 MLP	0.227	0.229	0.183	0.182 ^a

(Cont'd...)

Table 7. (Continued)

Model	Population	Global-kNN	Personal-kNN	NFP
B5 RF	0.220	0.229	0.180 ^a	0.180 ^a
Mean	0.227	0.229	0.182	0.181 ^a
<i>n</i> = 20 meals (39 subjects)				
B1 Ridge	0.211	0.216	0.173 ^a	0.173 ^a
B2 XGBoost	0.230	0.216	0.173 ^a	0.173 ^a
B3 BiLSTM	0.213	0.216	0.173 ^a	0.174
B4 MLP	0.215	0.216	0.174 ^a	0.174 ^a
B5 RF	0.209	0.216	0.172 ^a	0.173
Mean	0.216	0.216	0.173 ^a	0.174

Note: Lower values indicate better performance. Superscript “a” denotes the lowest value in each row; tied values are marked equally. Abbreviations: BiLSTM: Bidirectional long short-term memory; MLP: Multilayer perceptron; NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error; RF: Random forest; XGBoost: Extreme gradient boosting.


Figure 4. Mean Δ NRMSE improvement vs. personal history size *n*

Abbreviations: NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error.

regression, BiLSTM, and MLP fall between these extremes. Because the five baselines span linear, tree-ensemble, and neural families, consistent gains across them support model-agnosticity more directly than the breadth of the list alone.

A further pattern is the convergence of the personalized errors. Population NRMSE@60 ranges from 0.228 to 0.248 across architectures at *n* = 5, a spread of 0.020, whereas the corresponding NFP values fall within 0.201 to 0.203, a spread of 0.002. By *n* = 20 the NFP values sit within 0.173 to 0.174 regardless of base model. This suggests that accumulated personal evidence supplies much of the signal at later checkpoints, narrowing the differences between base learners. The same reading explains why Personal-kNN approaches NFP at *n* = 15, which reflects the intended weighting rather than a weakness of the fusion. The modest lead NFP holds at *n* = 0 is best described as cold-start stabilization through the global anchor, not full

personalization, since no personal history is yet available.

3.4. Generalization across prediction horizons

The analysis so far fixes the 60-min horizon, where the postprandial peak typically falls. Table 8 extends the evaluation to the 30 and 120-min targets under the same protocol and weights. NFP helps at every horizon, but the size of the gain differs across them. At 30 min, the population model is already accurate (NRMSE 0.175), and NFP delivers a smaller +0.013 at *n* = 5, whereas at 60 and 120 min it secures +0.032 and +0.030, a reduction near 13% that persists two hours after the meal. Raising history to *n* = 10 lifts every horizon further while preserving this ordering. The gain is therefore not specific to the 60-min target, and it is smallest at the shortest horizon.

3.5. Personalization by metabolic group

Neighbor Fusion Personalization carries no diagnosis

label, yet its benefit is not spread evenly across metabolic groups. Table 9 breaks the 60-min improvement at $n = 5$ by diagnosis. The group counts match Table 2 except that the single excluded one-meal subject was prediabetic, leaving 15 prediabetic subjects here. Prediabetic subjects gain most (+0.045), type 2 diabetic subjects next (+0.030), and healthy controls least (+0.021), and the same ordering holds at $n = 10$. The two groups of greatest clinical interest thus receive the largest gains without any group-specific tuning, which supports using NFP as a single layer rather than a per-cohort model.

3.6. Relationship between population bias and personalization gain

Figure 5 relates each subject's personalization gain to the systematic error the population model leaves on that subject, both measured at the 60-min horizon and averaged across the five base models as defined in the evaluation protocol. Each point is one subject, positioned by population-model bias against NFP gain at $n = 5$. The two are strongly and positively associated (Pearson $r = 0.77$, $p < 10^{-9}$). Subjects with low bias, clustered near the origin, show small gains, whereas the subjects the population model misreads most show the largest gains. The diagnosis groups fall along the same line rather than forming separate clusters, with healthy subjects at the low-bias end and prediabetic and type 2 diabetic subjects spread toward the high-bias end.

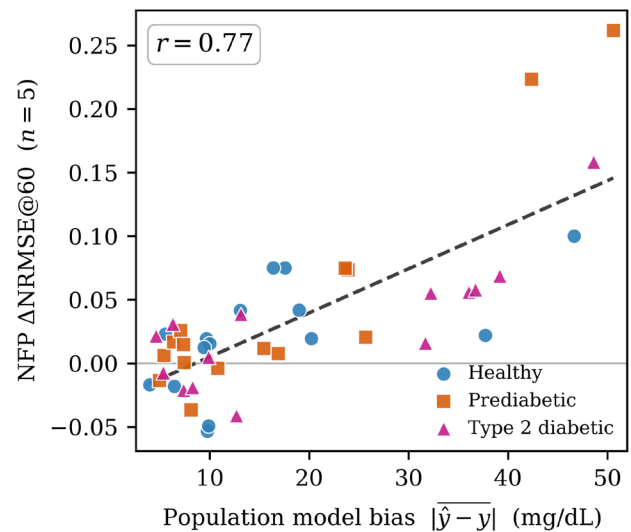


Figure 5. Per-subject population-model bias versus NFP gain at $n = 5$, 60-min horizon. Each point is one subject; the dashed line represents the least-squares regression fit. Abbreviations: NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error.

3.7. External-cohort validation

Table 10 reports the personalization gain on the two external cohorts, averaged across the five retrained base architectures. In Shanghai T2DM, the average

Table 8. NRMSE by prediction horizon, averaged across the five architectures

Personal-history meals (n)	Horizon (min)	Population	Personal-kNN	NFP	Δ NFP
5	30	0.175	0.168	0.162	+0.013
	60	0.234	0.206	0.201	+0.032
	120	0.236	0.215	0.206	+0.030
10	30	0.172	0.157	0.154	+0.018
	60	0.228	0.196	0.194	+0.034
	120	0.229	0.204	0.200	+0.029

Note: Δ NFP is the gain over the population model. Abbreviations: NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error.

Table 9. NFP Δ NRMSE@60 by metabolic group, averaged across the five architectures

Group	Subjects	Δ NFP ($n = 5$)	Δ NFP ($n = 10$)
Healthy controls	15	+0.021	+0.020
Prediabetes	15	+0.045	+0.050
Type 2 diabetes	14	+0.030	+0.031

Note: Δ NFP is the gain over the population model. Abbreviations: NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error.

Table 10. External-cohort validation of NRMSE@60 and NFP gains at 60 and 120 min, averaged across five base-model architectures

Personal-history meals (n)	Population	NFP	$\Delta\text{NFP@60}$	$\Delta\text{NFP@120}$
Shanghai T2DM (type 2 diabetes, 100 patients, 3,013 meals)				
0	0.238	0.234	+0.004	–
1	0.240	0.232	+0.008	–
3	0.240	0.230	+0.010	–
5	0.240	0.233	+0.007	+0.024
10	0.236	0.223	+0.013	+0.028
HUPA-UCM (type 1 diabetes + insulin, 21 subjects, 845 meals)				
0	0.325	0.337	–0.012	–
1	0.321	0.323	–0.002	–
3	0.320	0.308	+0.012	–
5	0.317	0.305	+0.012	+0.024
10	0.313	0.286	+0.027	+0.036

Note: Population models were retrained separately for each cohort, whereas the NFP hyperparameters were kept fixed. Abbreviations: NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error.

$\Delta\text{NRMSE@60}$ was positive at every few-shot checkpoint, rising from +0.004 at cold start to +0.013 at $n = 10$, and the gain was larger at the 120-min horizon (+0.024 at $n = 5$ and +0.028 at $n = 10$). The improvement was concentrated in the higher-error nonlinear and neural baselines: NFP gains were statistically significant for XGBoost and the BiLSTM across checkpoints ($p_{\text{fdr}} < 0.05$), weaker for RF, and slightly negative for ridge regression, so the effect held on average rather than for every model.

HUPA-UCM behaved as the harder setting its design implies. NFP did not improve cold-start prediction ($\Delta\text{NRMSE@60} = -0.012$ at $n = 0$), which reflects a weak two-dimensional global anchor without demographics, but it became positive once personal history accumulated, reaching +0.012 at $n = 5$ and +0.027 at $n = 10$, with the largest gains again at 120 min (+0.036 at $n = 10$). These gains were led by XGBoost. In both cohorts, the 30-min horizon showed no benefit, consistent with the short-horizon response being well captured by the pre-meal trajectory.

A consistent finding across both cohorts concerns the fusion design. Pure Personal-kNN was highly unstable when history was short, with NRMSE@60 of 0.364 in Shanghai and 0.491 in HUPA-UCM at $n = 1$, far worse than the population model, whereas the fused estimate remained stable and improved on it. The standalone Global-kNN comparator never beat the population model, matching the CGMacros result. Together these external cohorts support the direction and mechanism of NFP under frozen

hyperparameters, while showing that the magnitude of benefit is smaller when the meal representation is coarser and treatment effects are unobserved.

3.8. Comparison with personalization baselines

To place NFP against conventional per-subject adaptation, we evaluated three inference-time baselines that wrap the same frozen population predictions under the identical few-shot protocol. The first, subject-specific bias calibration, subtracts the mean residual over the personal history. Linear recalibration instead fits a slope and intercept of the true target on the population prediction across the available history. A third baseline, an empirical-Bayes random-intercept (mixed-effect) model, shrinks each per-subject bias toward the population by a factor estimated on the training folds. Table 11 reports the resulting $\Delta\text{NRMSE@60}$.

Naive calibration is unstable when history is short, giving a negative mean gain at $n \leq 3$, and linear recalibration diverges (mean -0.648 at $n = 3$) because a two-parameter fit on three to five points is ill-conditioned. The mixed-effects model, which pools the per-subject offset toward the population, is the strongest baseline and is positive throughout. NFP still exceeds it at every checkpoint, significantly so in the few-shot regime (one-sided Wilcoxon on per-subject gains, $p = 0.032$, 0.004 , and 0.006 at $n = 1$, 3 , and 5), and the two converge only at $n = 10$ ($p = 0.13$). The advantage is largest where history is scarcest, consistent with NFP conditioning on the specific meal rather than applying one per-subject offset.

Table 11. NFP versus per-subject calibration and mixed-effects baselines, mean $\Delta\text{NRMSE}@60$ across five architectures (higher is better)

Personal-history meals (n)	Calibration	Linear	Mixed-effect	NFP
1	−0.027	−0.027	+0.004	+0.012
3	−0.005	−0.648	+0.018	+0.024
5	+0.017	−0.023	+0.026	+0.032
10	+0.025	+0.012	+0.029	+0.034

Abbreviations: NFP: Neighbor Fusion Personalization; NRMSE: Normalized root mean square error.

3.9. Sensitivity analyses

Three further analyses probe design choices raised in review: the form of the weighting schedule, robustness to noisy meal entries, and recency weighting of personal neighbors.

3.9.1. Weighting schedule

Table 12 compares the adaptive schedule $w(n) = n/(n + \lambda)$ with alternatives at matched checkpoints. Changing λ within {3, 5, 10} moves the mean gain by at most 0.004, confirming the stability noted earlier. A constant weight of 0.5 matches the schedule except at the smallest history ($n = 1$: +0.006 versus +0.012), where the schedule's early caution is exactly what helps. Switching abruptly to Personal-kNN as soon as any history exists is far worse (−0.050 at $n = 1$), which is the direct justification for a gradual handover. An uncertainty-gated variant, which down-weights the personal estimate when its neighbors disagree, does not improve on the fixed schedule and is worse from $n = 3$ onward. The simple parametric schedule is therefore at least as good as the alternatives without added tuning.

3.9.2. Robustness to noisy meals

Because macronutrient entries are often estimated imprecisely in practice, we added Gaussian noise to the carbohydrate, protein, and fat values seen by the personalization layer at $n = 5$. NFP retained 96 to 100% of its gain as the noise grew from 0.1 to a full standard deviation of each feature ($\Delta\text{NRMSE}@60$ of +0.032, +0.032,

+0.032, +0.032, and +0.031 at noise levels 0, 0.1, 0.25, 0.5, and 1.0), because the pre-meal CGM window dominates the personal match and the population and global anchors are unaffected.

3.9.3. Recency weighting

Weighting personal neighbors toward more recent meals gave no material benefit on CGMacros and reduced the gain at larger history (for example, +0.026 versus +0.043 at $n = 20$ under the strongest recency setting), consistent with the low temporal drift reported in Section 2.6. Equal weighting of past meals is therefore retained, and recency weighting is left for deployments where behavioral drift is expected.

4. Discussion

4.1. Personalization corrects systematic individual bias

The weak performance of the standalone Global-kNN comparator is informative. The population models already receive the pre-meal CGM window and HbA1c, both of which carry strong between-subject metabolic information, with individual pre-meal CGM bins correlating up to about 0.58 and HbA1c about 0.53 with the 60-min target across all meals (Pearson), so similar training meals surface little that the model has not already used. This is also why classical case-based prediction from matched cases^{22,23} adds no gain once a competent parametric model is in place. Personal-kNN succeeds for the opposite reason. It conditions on the specific meal rather than applying a fixed per-subject correction. The per-subject regression slope of the 60-min glucose response on carbohydrate amount varies widely across individuals, from slightly negative (about −0.3) to about 0.72 mg/dL per gram, yet the population model applies one subject-agnostic mapping to everyone. Drawing outcomes from the subject's own past meals that most resemble the current one captures this per-person, per-meal-type response, with no gradient step or retraining, unlike meta-learning approaches that fine-tune for every new subject.^{11,12}

Table 12. Weighting-scheme ablation, mean $\Delta\text{NRMSE}@60$ across the five architectures. Schedule is $w(n) = n/(n + \lambda)$ with $\lambda = 5$

Personal-history meals (n)	Schedule	$\lambda = 3$	$\lambda = 10$	Constant 0.5	Hard switch	Uncertainty
1	+0.012	+0.012	+0.009	+0.006	−0.050	+0.012
3	+0.024	+0.022	+0.021	+0.022	−0.026	+0.016
5	+0.032	+0.031	+0.028	+0.032	+0.005	+0.023
10	+0.034	+0.031	+0.033	+0.033	+0.014	+0.026

Note: Schedule: $w(n) = n/(n + \lambda)$ with $\lambda = 5$.

This subject-level pattern, shown in [Figure 5](#), suggests that personalization acts as a targeted correction rather than a uniform lift. Subjects the population model already fits well have little systematic bias to remove and gain little, whereas the subjects it misreads are the ones NFP helps most. That the diagnosis groups fall along a single line rather than forming separate regimes suggests the effect is governed by the size of the per-subject bias rather than by metabolic label, which is consistent with NFP carrying no diagnosis information.

4.2. From cold start to individual evidence

The global component earns its place only in the cold-start regime ($n \leq 3$), after which the schedule in [Equation 2](#) shifts weight to Personal-kNN. The learning curves make this concrete. NFP leads Personal-kNN by about +0.005 through the first five meals, and the two converge once personal history dominates. The pattern echoes the partial-pooling intuition of Bayesian personalization methods^{26,29}, which lean on a population prior when subject data are scarce and on individual evidence once they are plentiful. NFP reaches the same behavior without committing to any mechanistic model of glucose dynamics, and without the per-model derivation those methods require, because the population prediction enters only as one of three numbers being averaged.

The horizon and diagnosis results extend the same mechanism along two further axes. In each, the gain scales with the fraction of the target that is individual rather than population-level. The 30-min target and the healthy group sit at the low end, where the response is small and well captured by the pre-meal trajectory, so little personal signal remains to exploit. The 60- and 120-min targets and the prediabetic and diabetic groups sit at the high end, where absorption and clearance differ enough between people that conditioning on a subject's own history pays off. NFP helps precisely where the individual share of the response is largest, which is also where accurate postprandial prediction matters most clinically. In absolute terms, the mean gain at $n = 5$ is about 5 mg/dL at the 60-min horizon; this is a modest reduction in predictive error rather than a demonstrated change in clinical outcome, and whether it improves dietary decision support would require a prospective study. Across five architectures, three horizons, and three metabolic groups, under a single set of hyperparameters, the effect holds, which points to a property of the personalization problem rather than of any one model or cohort.

4.3. Practical deployment

Three properties make NFP straightforward to deploy. It is training-free, so a new subject is served from the first

prediction, including the zero-history condition, with no gradient step and no waiting period. It is model-agnostic, since the population model enters only through its output, so the same layer wraps a linear model, a tree ensemble, or a neural network without modification. It requires no diagnosis label, which removes a barrier in settings where metabolic status is unknown or unconfirmed. Each prediction adds two nearest-neighbor searches over a cached training index and a short personal history. We benchmarked these costs directly ([Table 13](#)). On a CPU, the NFP layer adds about 0.11 ms per prediction and stores under 25 kB per subject, comprising a 20 kB global index shared across the cohort and a personal history under 2 kB. This overhead is a small fraction of a single forward pass of the neural and tree-ensemble base models (0.2 to 15%), comparable to a bare linear model in absolute terms, and negligible against the five-minute interval between CGM readings. The absence of gradient updates and the small retrieval pools therefore make NFP compatible with resource-constrained deployment.

Table 13. Computational cost of the NFP layer (CPU, brute-force nearest-neighbor search over a fold with 1,010 training meals)

Quantity	Value
Global index (shared across cohort)	20 kB, built in 0.19 ms
Per-query Global-kNN search	71 μ s
Per-query Personal-kNN search	39 μ s
Total NFP overhead per prediction	0.11 ms
Personal history stored per subject	1.5 kB
Overhead vs. base-model inference	0.2–15% (RF to BiLSTM)
Overhead vs. 5-min CGM interval	0.00004%

Abbreviations: BiLSTM: Bidirectional long short-term memory; CGM: Continuous glucose monitor; NFP: Neighbor Fusion Personalization; RF: Random forest.

In practice, the layer follows a simple online cycle. The pretrained population model and the global retrieval index are installed once and shared across users. A new subject first receives the blend of the population model and Global-kNN, since no personal meals exist yet. After each meal, once its postprandial outcome becomes observable, the meal is appended to the subject's local history, and later predictions retrieve only from meals that precede the query. The architecture permits an implementation in which personal records remain on the device while the global index stays fixed, keeping the personal data local. Several operating conditions shape this cycle. Meal times and nutrient values must be available at prediction time, and the personal component cannot update until the

relevant postprandial outcome has been recorded. Unusual meals with no close personal analog may still depend mostly on the population prior, and because the current rule treats all earlier meals as equally eligible, long-term behavioral or metabolic drift could make older neighbors less representative.

4.4. External validity across cohorts

The two external cohorts test transfer of the mechanism rather than reproduction of the CGMacros effect size, and they support three claims. The clearest is that the effect direction transfers. Once some personal history is available, $\Delta\text{NRMSE}@60$ turns positive in both cohorts, immediately in Shanghai T2DM and from $n = 3$ in the harder HUPA-UCM setting. The fusion design also holds out of sample. Pure Personal-kNN is unstable when history is short, performing far worse than the population model at $n = 1$ in both cohorts, and it is the shrinkage toward the population and global anchors that keeps early personalization safe. What differs across cohorts is the magnitude of benefit, which tracks feature fidelity. The CGMacros gain of +0.032 at $n = 5$ falls to +0.007 in Shanghai and +0.012 in HUPA-UCM, consistent with those cohorts carrying only coarse food-weight proxies or carbohydrate alone rather than the measured macronutrients available in CGMacros. This marks a boundary on where the gain is large rather than a failure of the method. NFP continues to help most at the 60 and 120-min horizons and for the higher-error base models, the same pattern seen within CGMacros, and in HUPA-UCM it does so despite insulin-mediated dynamics that the meal features do not observe.

4.5. Limitations and validity considerations

Several considerations bound the scope of these results. For internal validity, the protocol assigns subjects to folds as units, normalizes using training-fold statistics only, orders each subject's meals chronologically, and excludes future meals from every prediction, which together guard against information leaking from test data into training. Hyperparameters were selected on the training folds, yet selection within a single cohort of this size may still favor dataset-specific settings, and Personal-kNN remains sensitive to k_p when history is short, with $k_p = 1$ falling well below $k_p = 5$. The weighting schedule was tuned on the 60-min target and reused unchanged at 30 and 120 min, so the multi-horizon gains are, if anything, conservative.

For construct validity, glucose at 30, 60, and 120 min captures clinically interpretable post-meal states but does not represent the full excursion. Alternative targets such as peak rise, incremental area under the curve, or time to peak describe other aspects of the response and may show different personalization behavior, which the present

design does not measure.

For external validity, a substantial limitation arises from the data themselves, where most of the postprandial excursion variance lies within subjects rather than between them (Table 4), so much of the meal-to-meal spread is meal-specific. The high within-subject variability indicates a difficult prediction regime and may reflect both unobserved influences, such as physical activity, sleep, medication, or stress, and limitations of the recorded representation. The plateau beyond $n = 15$ in Figure 4 may therefore arise from limited observable information, from saturation of neighbor-based adaptation, or from both. At $n = 20$, five subjects with too few meals drop out, lowering statistical power at the largest history sizes. The primary results come from a single 45-subject cohort recorded under one CGM setting; the two external cohorts extend the evidence to type 1 and type 2 diabetes under different devices and sampling rates, but they rely on coarser meal features and, in HUPA-UCM, on insulin-treated dynamics that the features do not capture, so the personalization gains there are correspondingly smaller. None of the evaluations involves prospective deployment, so the findings speak to predictive error rather than to dietary decisions or glycemic outcomes, and larger prospective studies remain to be done.

5. Conclusion

This study proposed NFP, a training-free, model-agnostic method that personalizes any pre-trained population glucose model at inference time by fusing two k-NN predictors, one drawn from the training population and one from the subject's own accumulating meal history. Evaluated on CGMacros across five architectures and 44 subjects, NFP achieved statistically significant NRMSE improvement at every tested checkpoint from three meals onward ($n = 3, 5, 10, 20$), reaching a mean ΔNRMSE of +0.032 at $n = 5$ with no retraining or diagnosis labels.

The component comparisons and breakdown analyses indicate where the gain comes from. A standalone full-feature Global-kNN comparator adds no measurable benefit over a strong population model, and the gap between NFP and Personal-kNN closes beyond the cold-start window ($n \leq 5$), which indicates that the global anchor matters mainly before personal history accumulates, while Personal-kNN drives the improvement as that history grows. The gain holds across the 30, 60, and 120-min horizons and is largest for prediabetic and type 2 diabetic subjects (+0.045 and +0.030 at $n = 5$), the groups whose responses carry the most individual structure. NFP also exceeded subject-specific calibration and a mixed-effects baseline in the few-shot regime, added about 0.11 ms per

prediction, and reproduced the direction of the effect on two independent type 1 and type 2 diabetes cohorts under frozen hyperparameters, with a magnitude that scaled with meal-feature fidelity. Its consistency across five architectures, three horizons, three metabolic groups, and two external cohorts, under a single set of hyperparameters, points to a property of the personalization problem rather than of any one model or cohort.

Because it is training-free and model-agnostic, NFP is a practical personalization layer for continuous glucose monitoring, available from the first prediction and modest in added computation. Several directions remain open. Richer history representations, such as embeddings from CGM foundation models or pretrained time-series models^{42,43}, may improve neighbor matching beyond the engineered features used here. Weighting neighbors by recency could address the long-term behavioral or metabolic drift that the current equal-eligibility rule ignores. Pairing NFP with explainability, as in recent Shapley-based analyses of glucose-monitoring models³⁸, could clarify which past meals drive each personalized prediction. Prospective evaluation across larger and more diverse cohorts would test whether the gains observed here translate into improved dietary decisions and glycemic outcomes.

Acknowledgments

None.

Funding

This work was funded by the DRTPM Research Program, Ministry of Higher Education, Science, and Technology of Indonesia (Main Contract No. 082/C3/DT.05.00/PL-MULTITAHUN LANJUTAN/2026; Derivative Contract No. 215/UN8.2/PG/2026).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Fatma Indriani

Data curation: Irwan Budiman, Radityo Adi Nugroho

Formal analysis: Fatma Indriani, Irwan Budiman, Dwi Kartini, Muhammad Zainal Muttaqin

Investigation: Fatma Indriani

Methodology: Fatma Indriani

Software: Fatma Indriani

Validation: Fatma Indriani

Visualization: Muhammad Zainal Muttaqin

Writing—original draft: Fatma Indriani

Writing—review & editing: Fatma Indriani, Irwan

Budiman, Dwi Kartini, Muhammad Zainal Muttaqin, Mohammad Reza Faisal

Ethics approval and consent to participate

Not applicable. This study uses the publicly available, de-identified CGMacros dataset.

Consent for publication

Not applicable.

Availability of data

This study uses three publicly available datasets. Model development and the primary analyses use the CGMacros dataset (Das A, Kerr D, Glantz N, *et al. Sci Data*. 2025;12(1):1557. doi:10.1038/s41597-025-05851-7). External validation uses two independent cohorts: the Shanghai T2DM dataset (Zhao Q, Zhu J, Shen X, *et al. Chinese diabetes datasets for data-driven machine learning. Sci Data*. 2023;10:35. doi:10.1038/s41597-023-01940-7) and the HUPA-UCM diabetes dataset (Hidalgo JI, Alvarado J, Botella M, Aramendi A, Velasco JM, Garnica O. HUPA-UCM diabetes dataset. *Data Brief*. 2024;55:110559. doi:10.1016/j.dib.2024.110559). The processed data and the analysis code supporting the findings of this study are available at Zenodo (<https://doi.org/10.5281/zenodo.21392945>).

Further disclosure

During the preparation of this manuscript, the authors used generative AI and AI-assisted technologies (a large language model) to improve the readability and language of the text and to assist in drafting portions of the related-work text that summarize existing, published literature. These tools were not used to produce scientific insights, to design the study, to generate, analyze, or interpret data, or to draw scientific conclusions; all such intellectual work, including all reported methods, experiments, results, and conclusions, was performed by the authors. All AI-assisted text was produced under full human supervision, and the authors reviewed, verified, and edited it, including confirming that every cited reference is accurate and correctly characterized. The authors take full responsibility for the integrity and accuracy of the published work. No AI or AI-assisted technology is listed as an author or co-author.

References

1. Das A, Kerr D, Glantz N, *et al.* CGMacros: a pilot scientific dataset for personalized nutrition and diet monitoring. *Sci Data*. 2025;12(1):1557. doi: 10.1038/s41597-025-05851-7

2. Sun H, Saeedi P, Karuranga S, *et al.* IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract.* 2022;183:109119.
doi: 10.1016/j.diabres.2021.109119
3. Battelino T, Danne T, Bergenstal RM, *et al.* Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes Care.* 2019;42(8):1593-1603.
doi: 10.2337/dci19-0028
4. Oviedo S, Vehí J, Calm R, Armengol J. A review of personalized blood glucose prediction strategies for T1DM patients. *Int J Numer Method Biomed Eng.* 2017;33(6):e2833.
doi: 10.1002/cnm.2833
5. Woldaregay AZ, Årsand E, Walderhaug S, *et al.* Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes. *Artif Intell Med.* 2019;98:109-134.
doi: 10.1016/j.artmed.2019.07.007
6. Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: a deep learning framework for accurate glucose forecasting. *IEEE J Biomed Health Inform.* 2020;24(2):414-423.
doi: 10.1109/JBHI.2019.2931842
7. Martinsson J, Schliep A, Eliasson B, Mogren O. Blood glucose prediction with variance estimation using recurrent neural networks. *J Healthc Inform Res.* 2020;4(1):1-18.
doi: 10.1007/s41666-019-00059-y
8. Sun Q, Jankovic MV, Bally L, Mougiakakou SG. Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network. In: *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE; 2018:1-5.
doi: 10.1109/NEUREL.2018.8586990
9. Pérez-Gandía C, Facchinetti A, Sparacino G, *et al.* Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technol Ther.* 2010;12(1):81-88.
doi: 10.1089/dia.2009.0076
10. Mhaskar HN, Pereverzyev SV, van der Walt MD. A deep learning approach to diabetic blood glucose prediction. *Front Appl Math Stat.* 2017;3:14.
doi: 10.3389/fams.2017.00014
11. Zhu T, Li K, Herrero P, Georgiou P. Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning. *IEEE Trans Biomed Eng.* 2023;70(1):193-204.
doi: 10.1109/TBME.2022.3187703
12. Langerica S, Rodriguez-Fernandez M, Núñez F, Doyle FJ. A meta-learning approach to personalized blood glucose prediction in type 1 diabetes. *Control Eng Pract.* 2023;135:105498.
doi: 10.1016/j.conengprac.2023.105498
13. Chen J, Cai D, Shi D, Shi L, Liu W, Ji L. Physiology-informed Gaussian process meta-learning for postprandial glucose trajectory prediction. *IEEE Sens J.* 2025;25(18):34980-34991.
doi: 10.1109/JSEN.2025.3594498
14. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: *Proc 34th Int Conf Mach Learn (ICML)*. 2017;70:1126-1135. <https://proceedings.mlr.press/v70/finn17a.html>
15. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: *Adv Neural Inf Process Syst (NeurIPS)*. 2017;30:4077-4087. https://www.cs.toronto.edu/~zemel/documents/prototypical_networks_nips_2017.pdf
16. Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: *Adv Neural Inf Process Syst (NeurIPS)*. 2016;29:3630-3638. <https://proceedings.neurips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf>
17. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput Surv.* 2020;53(3):63.
doi: 10.1145/3386252
18. Yu X, Yang T, Lu J, *et al.* Deep transfer learning: a novel glucose prediction framework for new subjects with type 2 diabetes. *Complex Intell Syst.* 2022;8:1875-1887.
doi: 10.1007/s40747-021-00360-7
19. Seo W, Park SW, Kim N, Jin SM, Park S. A personalized blood glucose level prediction model with a fine-tuning strategy: a proof-of-concept study. *Comput Methods Programs Biomed.* 2021;211:106424.
doi: 10.1016/j.cmpb.2021.106424
20. Daniels J, Herrero P, Georgiou P. A multitask learning approach to personalized blood glucose prediction. *IEEE J Biomed Health Inform.* 2022;26(1):436-445.
doi: 10.1109/JBHI.2021.3100558
21. Shen Y, Kleinberg S. Personalized blood glucose forecasting from limited CGM data using incrementally retrained LSTM. *IEEE Trans Biomed Eng.* 2025;72(4):1266-1277.
doi: 10.1109/TBME.2024.3494732
22. Zulj S, Carvalho P, Ribeiro RT, Andrade R, Magjarević R. Data size considerations and hyperparameter choices in case-based reasoning approach to glucose prediction. *Biocybern Biomed Eng.* 2021;41(2):733-745.
doi: 10.1016/j.bbe.2021.04.013
23. Aiello EM, Toffanin C, Messori M, Cobelli C, Magni L. Postprandial glucose regulation via KNN meal classification

- in type 1 diabetes. *IEEE Control Syst Lett.* 2019;3(2):230-235.
doi: 10.1109/LCSYS.2018.2844179
24. Brown D, Aldea A, Harrison R, Martin CE, Bayley I. Temporal case-based reasoning for type 1 diabetes mellitus bolus insulin decision support. *Artif Intell Med.* 2018;85:28-42.
doi: 10.1016/j.artmed.2017.09.007
25. Torrent-Fontbona F, López B. Personalized adaptive CBR bolus recommender system for type 1 diabetes. *IEEE J Biomed Health Inform.* 2019;23(1):387-394.
doi: 10.1109/JBHI.2018.2813424
26. Albers D, Levine M, Gluckman B, Ginsberg H, Hripcsak G, Mamykina L. Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS Comput Biol.* 2017;13(4):e1005232.
doi: 10.1371/journal.pcbi.1005232
27. Levine ME, Hripcsak G, Mamykina L, Stuart A, Albers DJ. Offline and online data assimilation for real-time blood glucose forecasting in type 2 diabetes. *arXiv. Preprint posted online September 1, 2017.*
doi: 10.48550/arXiv.1709.00163
28. Eren-Oruklu M, Çinar A, Quinn L, Smith D. Estimation of future glucose concentrations with subject-specific recursive linear models. *Diabetes Technol Ther.* 2009;11(4):243-253.
doi: 10.1089/dia.2008.0065
29. Chung I, Kim S, Lee J, Kim KJ, Hwang SJ, Yang E. Deep mixed effect model using Gaussian processes: a personalized and reliable prediction for healthcare. *Proc AAAI Conf Artif Intell.* 2020;34(4):3649-3657.
doi: 10.1609/aaai.v34i04.5773
30. Liu Z, Hauskrecht M. A Personalized Predictive Framework for Multivariate Clinical Time Series via Adaptive Model Selection. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* ACM; 2017:1169-1177.
doi: 10.1145/3132847.3132859
31. Lewis P, Perez E, Piktus A, *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems.* Vol 33. Curran Associates, Inc.; 2020:9459-9474. Accessed July 13, 2026. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
32. Gao Y, Xiong Y, Gao X, *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv. Preprint posted online 2023.*
doi: 10.48550/arXiv.2312.10997
33. Jing B, Zhang S, Zhu Y, *et al.* Retrieval Based Time Series Forecasting. *arXiv. Preprint posted online 2022.*
doi: 10.48550/arXiv.2209.13525
34. Tire K, Taga EO, Ildiz ME, Oymak S. Retrieval Augmented Time Series Forecasting. *arXiv. Preprint posted online 2024.*
doi: 10.48550/arXiv.2411.08249
35. Zhang H, Xu C, Zhang YF, Zhang Z, Wang L, Bian J. TimeRAF: retrieval-augmented foundation model for zero-shot time series forecasting. *IEEE Trans Knowl Data Eng.* 2025;37(9):5654-5665.
doi: 10.1109/TKDE.2025.3579137
36. Ning K, Pan Z, Liu Y, *et al.* TS-RAG: Retrieval-Augmented Generation based Time Series Foundation Models are Stronger Zero-Shot Forecaster. *arXiv. Preprint posted online 2025.*
doi: 10.48550/arXiv.2503.07649
37. Soumma SB, Ghasemzadeh H. GlyRAG: Context-Aware Retrieval-Augmented Framework for Blood Glucose Forecasting. *arXiv. Preprint posted online 2026.*
doi: 10.48550/arXiv.2601.05353
38. Adjevi A, Abdirashid AM, Aktaş F, Ucar MHB, Solak S. Explainable reinforcement learning for glucose monitoring based on Shapley value analysis. *Comput Methods Programs Biomed.* 2026;278:109266.
doi: 10.1016/j.cmpb.2026.109266
39. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory.* 1967;13(1):21-27.
doi: 10.1109/TIT.1967.1053964
40. Zhao Q, Zhu J, Shen X, *et al.* Chinese diabetes datasets for data-driven machine learning. *Sci Data.* 2023;10(1):35.
doi: 10.1038/s41597-023-01940-7
41. Hidalgo JI, Alvarado J, Botella M, Aramendi A, Velasco JM, Garnica O. HUPA-UCM diabetes dataset. *Data Brief.* 2024;55:110559.
doi: 10.1016/j.dib.2024.110559
42. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *arXiv. Preprint posted online 2022.*
doi: 10.48550/arXiv.2211.14730
43. Ansari AF, Stella L, Turkmen C, *et al.* Chronos: Learning the Language of Time Series. *arXiv. Preprint posted online 2024.*
doi: 10.48550/arXiv.2403.07815