ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Knowledge-constrained neural additive modeling for hypoglycemia risk stratification in older adults with type 2 diabetes

Zheng Rong[1†], Hui Si[2†], Wei Bai[3†], Difei Wang[4], Longfeng Sun[2], Hui Tian[5], Shuangtong Yan[5], Xinyu Miao[5], Xin Kang[1*], and Xin Hou[2*]

[1]Department of Computer Science, Tokushima University, Tokushima, Tokushima, Japan

[2]Department of Geriatrics, The First Affiliated Hospital of China Medical University, Shenyang, Liaoning, China

[3]Department of Aircraft Design, College of Aerospace Engineering, Shenyang Aerospace University, Shenyang, Liaoning, China

[4]Department of Geriatrics, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China

[5]Department of Endocrinology, the Second Medical Center and National Clinical Research Center for Geriatric Diseases, Chinese People's Liberation Army General Hospital, Beijing, China

[†]*These authors contributed equally to this work.*

***Corresponding authors:**
Xin Kang
(kang-xin@is.tokushima-u.ac.jp);
Xin Hou
(houxin83@163.com)

Citation: Rong Z, Si H, Bai W, *et al*. Knowledge-constrained neural additive modeling for hypoglycemia risk stratification in older adults with type 2 diabetes. *Artif Intell Health.* 2026;3(4):026130031. doi: 10.36922/AIH026130031

Received: March 28, 2026

Revised: May 9, 2026

Accepted: May 13, 2026

Published online: June 5, 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.

Abstract

Hypoglycemia remains a major barrier to safe glucose-lowering treatment in older adults with type 2 diabetes, yet clinically relevant models should combine nonlinear expressiveness, transparent feature attribution, and well-calibrated probability estimates. We developed a knowledge-constrained neural additive model (CNAM) for clinically interpretable stratification of recorded hypoglycemia history in older adults with type 2 diabetes within a multicenter cross-sectional outpatient cohort. The model was evaluated with a locked 10% test split, nested cross-validation, probability calibration, threshold-based decision analysis, decision-curve analysis, and explanation-stability assessment. It achieved competitive locked-test discrimination (receiver operating characteristic [ROC]–area under the curve [AUC] = 0.7350; precision–recall [PR]-AUC = 0.4127) and, after sigmoid calibration fitted on pooled development out-of-fold predictions, showed a calibrated Brier score of 0.1136 and an expected calibration error of 0.0189. Additional operating-point analyses showed trade-offs between sensitivity and specificity, and decision-curve analysis indicated broadly similar net-benefit profiles for CNAM and logistic regression across the evaluated threshold range. The monotonicity-constrained parameterization structurally enforces directionally plausible feature effects on selected predictors, with ablation results providing context against unconstrained variants that exhibit nonzero violations. These findings support CNAM as a complementary, interpretable stratification framework for recorded hypoglycemia history rather than as a replacement for established clinical risk scores. External validation, severity-specific outcome capture, and local recalibration remain necessary before broader clinical use.

Keywords: Hypoglycemia; Type 2 diabetes; Older adults; Neural additive models; Calibration; Interpretability

1. Introduction

Hypoglycemia remains a major obstacle to safe glucose-lowering therapy in older adults with type 2 diabetes.[1-4] In this population, multimorbidity, impaired counter-regulation, treatment complexity, and long disease duration jointly increase the likelihood of clinically important hypoglycemic events.[5-8] Because hypoglycemia is clinically relevant to vulnerability assessment, medication review, and individualized glycemic targets, tools for stratifying recorded hypoglycemia history should provide outputs that clinicians can interpret and trust.[1,4]

Despite this clinical importance, an important methodological gap remains. Many hypoglycemia studies emphasize association analysis or setting-specific prediction, whereas comparative evaluations of clinically interpretable nonlinear models remain limited. Existing clinical prediction studies often focus solely on discrimination, even though clinical decision-making also depends on calibration, threshold behavior, and the stability of model explanations under reasonable data perturbations.

Methodologically, this gap reflects two underlying limitations within current artificial intelligence (AI) applications. Conventional regression models are transparent but may miss clinically meaningful nonlinear effects, while flexible black-box models can improve prediction at the cost of interpretability and confidence in clinical use.[9,10]

To address these limitations, we develop a knowledge-constrained neural additive model (CNAM) for clinically interpretable stratification of recorded hypoglycemia history in older adults with type 2 diabetes. Because the present study is cross-sectional, the goal is not to make temporal predictions about future incident events, but to identify clinically meaningful vulnerability patterns and clinically interpretable risk strata within the observed cohort. The proposed model combines step-based additive shape functions with smoothness regularization and monotonicity constraints on selected predictors, while preserving exact additive explanations at both the global and patient levels. We further evaluate the model using calibration analysis, threshold-based risk stratification, decision curve analysis, and explanation-stability assessment, ensuring the evaluation protocol aligns with the intended decision-support setting.

This study has two main contributions. First, it formulates a clinically oriented constrained additive neural model with step-based shape functions, smoothness regularization, monotonicity constraints on selected predictors, and exact additive explanations. Second, it evaluates the model with a decision-support protocol that considers discrimination, calibration, threshold behavior, decision curve analysis, explanation stability, and robustness checks. Thus, the contribution is the combination of clinically motivated constraints, exact additive explanation, and decision-support evaluation for hypoglycemia stratification, not the proposal of a completely new additive-model family.

2. Data and methods

2.1. Related work

2.1.1. Hypoglycemia risk stratification and prediction in diabetes

Shi *et al.*[6] developed a territory-wide electronic health record-based machine-learning model to predict severe hypoglycemia leading to hospitalization in older adults with diabetes. Their study demonstrated the value of large-scale clinical data for pragmatic hypoglycemia modeling, but the resulting framework did not focus on exact additive interpretability or structurally constrained feature-wise explanation. Ling *et al.* reported that glucose control, sulfonylurea exposure, and insulin therapy are closely associated with severe hypoglycemia and mortality in older adults with type 2 diabetes.[11] Their analysis clarified important clinical correlates, yet it primarily addressed association rather than interpretable individualized risk modeling. Shabestari *et al.*[12] further examined predictive factors of hypoglycemia in type 2 diabetes using a prospective machine-learning design and interpretable analysis, but their framework did not emphasize stability-oriented constrained additive modeling.

Existing hypoglycemia risk scores and glycemic-risk indices already offer compact ways to flag patients who may need clinical review or closer monitoring. CNAM is not proposed as a replacement for these tools. Its role here is complementary: to examine whether constrained additive learning can yield calibrated probabilities and exact feature-wise explanations in this cohort. Direct head-to-head benchmarking would require external datasets that contain the relevant score inputs, monitoring information, and compatible outcome definitions.

Collectively, prior research demonstrates the clinical importance of hypoglycemia prediction while also exposing a persistent gap. Existing work often prioritizes association analysis, setting-specific prediction, or discrimination-only reporting, whereas fewer studies evaluate whether the learned model is simultaneously interpretable, well-calibrated, threshold-aware, and robust enough for clinical decision support.

2.1.2. Interpretable additive models and knowledge-constrained learning

Hastie and Tibshirani[13] established the statistical foundation for additive modeling by showing that smooth, feature-wise functions can balance flexibility and interpretability. Agarwal *et al.*[14] introduced neural additive models, showing that exact feature-wise decomposability can be retained while remaining competitive on tabular benchmarks. Nori *et al.*[15] subsequently presented explainable boosting machines as a strong additive baseline that combines boosting with interpretable shape functions. You *et al.*[16] showed that monotonic function classes can be incorporated into deep lattice networks, demonstrating that structural constraints can encode domain knowledge in high-stakes settings. Mariotti *et al.*[17] further explored constrainable neural additive models, demonstrating that carefully designed constraints can improve the balance between interpretability and predictive performance.

These additive and constrained-learning studies provide the methodological basis for our work, but, by themselves, they do not resolve the specific demands of clinically interpretable hypoglycemia stratification. In particular, prior studies seldom combine step-based shape control, training-time knowledge constraints, exact additive explanations, post hoc probability calibration, and formal explanation-stability analysis within one clinically oriented pipeline. Our contribution is therefore an integrated decision-support-oriented formulation for this problem setting rather than a claim of a wholly new model class.

2.1.3. Calibration, decision utility, and reporting quality

Van Calster *et al.*[18] described calibration as the Achilles' heel of predictive analytics and argued that decision-oriented evaluation should report calibration alongside discrimination.[18]

Together, these studies indicate that discrimination alone is insufficient when a model is meant to support clinical action. Recent reporting and risk-of-bias guidance, including Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + AI and Prediction Model Risk Of Bias Assessment Tool + AI, have further emphasized that machine-learning prediction studies should report calibration, validation design, and applicability boundaries alongside discrimination.[19,20] A clinically interpretable hypoglycemia model intended for decision support must also provide reliable probabilities, transparent decision thresholds, and reporting practices that make its limitations visible to readers and end users.

2.2. Methodology

2.2.1. Problem formulation and notation

For each patient, the available clinical predictors are transformed into a structured input representation that combines standardized continuous variables with one-hot encoded categorical indicators. The objective is to learn an intrinsically interpretable additive model that maps these transformed predictors to a patient-level logit score and a corresponding probability of hypoglycemia history. We define the model on the logit scale as **Equation 1**:

$$\eta\left(x_i\right) = b + \sum_{j=1}^{p} w_j f_j\left(z_{ij}\right) \tag{1}$$

In this formulation, the baseline term captures the overall log-odds, while each transformed feature contributes through its own one-dimensional response function and an associated scaling weight. The corresponding raw predicted probability is obtained through the logistic transformation shown in **Equation 2**:

$$p_{\text{raw}}\left(x_i\right) = \sigma\left(\eta\left(x_i\right)\right) \tag{2}$$

Here, the sigmoid operator maps the additive logit to the raw probability of hypoglycemia history. This formulation preserves exact additive decomposition while allowing each feature to contribute nonlinearly to the final prediction.

2.2.2. Step-based additive shape parameterization

The proposed model is a CNAM, in which each predictor is assigned a controllable one-dimensional shape function, and the final prediction is obtained by summing these feature-wise effects. Unlike black-box neural networks, CNAM does not rely on high-order latent interactions to form its prediction. Instead, each transformed feature is assigned a dedicated shape function, and the model output is the weighted sum of these feature-specific effects.

For the transformed feature j, we parameterize the shape function as a sum of differentiable step components (**Equation 3**):

$$f_j\left(z\right) = c_j + \sum_{k=1}^{K_j} a_{jk}\sigma\left(s_{jk}\left(z - q_{jk}\right)\right) \tag{3}$$

where K_j is the number of step components for the feature j, q_{jk} denotes the location of the k-th component, s_{jk} controls its slope, a_{jk} controls its amplitude, c_j is the feature-specific offset, and $\sigma(\cdot)$ is the sigmoid function. Each feature-specific response function is therefore governed by a set of knot locations, step heights, step

sharpness parameters, and a feature-specific bias term. Compared with generic per-feature subnetworks, this step-based parameterization makes it easier to control directional constraints and shape smoothness. In the implementation, knot locations are initialized from feature quantiles within each training fold. This choice keeps the learned shape functions directly controlled by the CNAM parameterization, without relying on a separate additive initialization model.

Because the model remains additive throughout training and inference, the contribution of each transformed feature to the final logit can be directly inferred from the learned decomposition.

2.2.3. Knowledge-constrained objective and monotonic parameterization

A central methodological feature of the proposed framework is that clinical knowledge enters the model before explanation, through both regularization and function-class constraints. We incorporate domain knowledge in two complementary ways. First, we regularize the learned shape functions to discourage unstable or overly complex feature effects. Second, for selected predictors with clinically motivated directional assumptions, we impose monotone increasing or monotone decreasing behavior directly through the shape-function parameterization.

Smoothness and amplitude are controlled through regularization on the internal parameters of the shape functions. The optimization objective combines the task loss with penalties on the additive scaling weights, the step-speed parameters, the step heights, and the feature-specific bias terms (**Equation 4**):

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{lin}} R_{\text{lin}}(w) + \lambda_s R_{\text{smooth}}(s) + \lambda_h R_{\text{height}}(a) + \lambda_b R_{\text{bias}}(c) \quad (4)$$

where L_{task} is the binary cross-entropy loss with logits, w denotes the linear aggregation weights, s the step slopes, a the step amplitudes, and c the bias terms. The regularizer $R_{\text{lin}}(w)$ encourages a sparse additive combination, $R_{\text{smooth}}(s)$ discourages abrupt local transitions, $R_{\text{height}}(a)$ limits excessively large component amplitudes, and $R_{\text{bias}}(c)$ suppresses unnecessarily large offsets. The coefficients λ_{lin}, λ_s, λ_h, and λ_b are nonnegative hyperparameters that control the relative strengths of these penalties.

For monotone features, the model does not rely on a post-hoc correction. Instead, monotonicity is induced structurally in the step representation. For a monotone-increasing feature, the step heights and step speeds are constrained to be nonnegative within the differentiable step function, yielding a nondecreasing learned shape. For a monotone-decreasing feature, the same mechanism is applied with sign reversal, producing a nonincreasing shape. This construction preserves differentiability while enforcing domain-consistent directionality at the level of the learned function class.

2.2.4. Structural global and local explanations

An important property and intended design goal of the proposed model is that interpretability is intrinsic rather than post hoc. A global explanation is provided by the learned collection of feature-wise response functions, which can be visualized directly to show how each predictor influences the log-odds of hypoglycemia history across its observed range. This is particularly useful for continuous predictors such as age, diabetes duration, body mass index (BMI), and glycated hemoglobin (HbA1c), where nonlinear effects are clinically plausible.

Local explanation follows from the same additive structure. For patient i, the logit can be decomposed exactly as **Equation 5**:

$$\eta(x_i) = b + \sum_{j=1}^{p} \alpha_{ij} \quad (5)$$

where b is the global intercept, p is the number of transformed predictors, and α_{ij} is the contribution of the transformed feature j for patient i to the final logit. Each contribution is defined as **Equation 6**:

$$\alpha_{ij} = w_j f_j(z_{ij}) \quad (6)$$

where z_{ij} denotes the transformed input of the feature j for patient i, and w_j scales the corresponding feature-wise shape function. Thus, each individual prediction is accompanied by a feature-wise rationale that sums exactly to the model score, without requiring surrogate explainers or approximate attribution methods.

2.2.5. Explanation stability as a model property

Because explanation quality depends not only on readability but also on reproducibility, we treat shape stability as a methodological target rather than as a secondary visualization. We therefore evaluate the stability of learned shape functions across cross-validation folds and random seeds. For each key continuous feature, fold-wise shape curves are exported on a common quantile grid, linearly interpolated when needed, and their agreement is summarized by the mean pairwise Pearson correlation. This provides a direct measure of whether clinically interpreted nonlinear patterns remain reproducible under resampling.

This stability analysis complements standard predictive metrics. In high-stakes clinical settings, a model whose explanations are highly unstable is difficult to trust even if its discrimination is acceptable. We therefore treat explanation stability as part of the methodological evaluation of the proposed CNAM framework.

2.2.6. Framework overview

Figure 1 provides a framework-level overview of the proposed CNAM. Figure 1A summarizes the training-time architecture. Preprocessed clinical predictors are mapped to feature-wise shape functions, whose additive aggregation defines the raw probability. Smoothness and monotonicity constraints act on shape-function learning rather than on the output layer. Figure 1B summarizes the decision-support-oriented outputs of the same framework. Raw CNAM outputs support a probability branch, in which development out-of-fold calibration yields calibrated probabilities for threshold-based risk stratification and clinical utility analysis, and an explanation branch, in which the additive structure supports global shape curves, patient-level decomposition, and stability analysis.

Figure 2 provides a more detailed schematic view of the proposed framework. Figure 2A formalizes the step-based additive representation from clinical predictors to the additive logit and raw probability. Figure 2B shows how knowledge-guided learning is implemented through a constrained objective that combines linear sparsity, shape smoothness, amplitude control, and bias control. Figure 2C summarizes, in a compact form, the decision-support-oriented view of calibrated decision support and structural explanation.

2.3. Experimental setup

2.3.1. Dataset, outcome, and predictors

We conducted a multicenter cross-sectional study using outpatient data collected between November 2015 and May 2017 from 150 hospitals across 21 provinces in China. Eligible participants were adults aged 60 years or older with type 2 diabetes. Baseline predictors were derived from standardized questionnaires and examinations and included demographics, socioeconomic indicators, diabetes history, treatment patterns, anthropometric variables, laboratory measurements, and documented chronic complications. To improve clinical feasibility and practical usability in routine care, we restricted the candidate predictor set to variables routinely available in the dataset.

The target outcome was recorded as hypoglycemia history, identified retrospectively from patient-reported history and clinical records collected in this cross-sectional outpatient survey. We used a blood glucose threshold of 3.9 mmol/L (70 mg/dL) as the clinical alert value for hypoglycemia, consistent with consensus definitions. The available dataset did not systematically distinguish Level 1, Level 2, and Level 3 hypoglycemia. We therefore analyzed a single composite recorded-history endpoint rather than severity-specific hypoglycemia outcomes. All participants provided informed consent, and the study was approved by the institutional ethics committee. The analytic cohort comprised 2,603 older adults with type 2 diabetes, and the prevalence of documented hypoglycemia history was 15.1%.

For descriptive statistical analysis, continuous variables were summarized as mean ± standard deviation, and categorical variables were summarized as counts and percentages. Group differences between participants with and without hypoglycemia were assessed using the Student's t-test for continuous variables and the chi-square test for categorical variables, as appropriate. All statistical tests were two-sided, and $p < 0.05$ was considered statistically significant.

2.3.2. Experimental protocol

To obtain an unbiased estimate of model performance, we reserved a stratified locked test set comprising 10% of the cohort and used the remaining participants as the development set. All model fitting, hyperparameter selection, calibration fitting, and threshold selection were performed exclusively within the development data, and the locked test set was accessed only once for final evaluation. The full experiment reused the same locked-test split, outer-fold partition, and CNAM hyperparameters established in the initial development stage, thereby ensuring that all reported comparisons were made on fixed splits.

Within the development set, model development followed a nested cross-validation design with five outer folds and five inner folds. For CNAM, early stopping was based on an internal split from the outer training fold, so the outer validation fold was not used for optimization control. Continuous predictors were median-imputed within training folds and then standardized. Categorical predictors were assigned a dedicated missing category when needed and encoded using one-hot encoding. Final locked-test predictions were generated by cross-validated model averaging, that is, by averaging predictions from outer-fold models rather than relying only on a single model refitted on the full development set. This choice reduced dependence on one particular training realization and aligned the final test predictions with the nested validation design. In parallel with the primary evaluation

pipeline, we conducted a review-driven missing-data audit on the aligned raw analysis layer, which preserves the original variable-level missingness before the paper-facing train/test clustered cross-validation preprocessing. Missingness was summarized for every modeled predictor. Supplementary sensitivity analyses then examined the robustness of the main conclusions under two additional checks: replacing the primary fold-wise median strategy for numerical variables with multiple imputation by chained equations (MICE)-style for the baseline models, and excluding participants with more than 20% missing predictors.

2.3.3. Baselines and implementation details

The candidate models were penalized logistic regression, a histogram-based gradient-boosted tree model, an explainable boosting machine (EBM), and the proposed CNAM. Logistic regression served as a transparent clinical reference, the tree model as a higher-capacity nonadditive comparator, and EBM as a strong interpretable additive baseline. Logistic regression, the tree-based model, and EBM were tuned in the inner cross-validation loop. CNAM was trained with the same fold structure as the baseline models and used the constrained objective defined in Section 2.2.3.

Implementation details were as follows: The baseline models were implemented with Python-based scikit-learn and interpret pipelines, whereas CNAM was trained with PyTorch Lightning. Logistic regression used the lbfgs solver with C = 1.0, max_iter = 100, and tolerance = 1×10^{-4}. The tree baseline was a histogram-based gradient boosting model with learning_rate = 0.1, max_iter = 100, and the same fixed outer-fold structure. CNAM used 128-step components for each numeric feature, batch_size = 256, learning rate = 0.001, weight decay = 1×10^{-4}, smoothness incentive = 0.01, linear-layer constraint coefficient = 1×10^{-4}, smallness incentive = 1×10^{-4}, standard neural-network initialization, and balanced positive-class weighting. Training used stochastic gradient descent, at most 200 epochs, and early stopping with patience 20 monitored on an internal validation split drawn from the outer-training fold; the internal early-stopping split used 10% of the outer-training fold with minimum positive and negative counts enforced, and the outer validation fold remained reserved for performance estimation only. The primary CNAM and ablation variants shared this implementation configuration, except for the intended constraint components being ablated. For the monotonicity-related ablation, increasing constraints were applied to age and diabetes duration, and a decreasing constraint was applied to age at diagnosis. Calibration and threshold selection were handled within the development data, as described in Section 2.3.4 and Algorithm S1. All experiments used fixed random seeds to improve reproducibility. Final model comparison was performed on the locked test set after nested development was completed, ensuring that all reported results reflected a decision-support-oriented evaluation rather than an optimistic single-split estimate.

2.3.4. Evaluation metrics

We used receiver operating characteristic (ROC)–area under the curve (AUC) and precision–recall (PR)-AUC to summarize discrimination. Threshold-dependent performance was reported using sensitivity, specificity, positive predictive value, and negative predictive value. Primary operating thresholds were selected on calibrated development probabilities and then applied unchanged to calibrated locked-test probabilities. Probability reliability was evaluated with the Brier score, expected calibration error, and calibration intercept and slope. The primary expected calibration error used 10 equal-width bins; equal-width and equal-mass binning sensitivity analyses are reported in Table S11. Clinical utility was assessed with threshold-based net-benefit analysis. Where applicable, 95% confidence intervals (CIs) for calibration metrics, threshold-dependent metrics, and selected net-benefit differences were estimated with 1,000 nonparametric bootstrap resamples.

Calibration was kept as a post hoc probability-mapping step and was not treated as the core model contribution. For each model, raw held-out predictions from the five development outer folds were pooled into one development out-of-fold prediction stack. One sigmoid calibration model was fitted on this pooled development stack. For the locked test set, raw predictions from the five outer-fold models were first averaged to obtain the cross-validation-ensemble raw probability, and the same sigmoid calibrator was then applied unchanged. The locked test set was not used for model fitting, hyperparameter selection, calibration fitting, threshold selection, or model selection. Algorithm S1 gives the full procedure. Because the Youden criterion is a statistical operating point rather than a cost-based clinical threshold, we also evaluated sensitivity-oriented and fixed-probability operating points for CNAM.

3. Results

3.1. Cohort characteristics and factors associated with hypoglycemia

The baseline characteristics of the study cohort are summarized in Table 1. Among the 2,603 enrolled participants, 393 (15.1%) were classified as having hypoglycemia. Age, sex, BMI, and HbA1c were broadly comparable between the two groups. Multivariable logistic

regression identified several clinically meaningful factors associated with hypoglycemia. Earlier diabetes onset was associated with higher hypoglycemia prevalence: participants diagnosed before 60 years of age had a higher prevalence than those diagnosed later (18.2% vs. 11.1%; odds ratio [OR] = 1.78; $p < 0.001$). Diabetes duration showed a graded association with hypoglycemia. Compared to duration ≤ 10 years, the odds increased for durations of 11–20 years (16.9%; OR = 1.51; $p < 0.001$), 21–30 years (22.4%; OR = 2.14; $p < 0.001$), and 31–40 years (40.5%; OR = 5.04; $p < 0.001$), indicating a dose-response pattern consistent with cumulative disease burden.

Medication exposure was also strongly associated with a history of hypoglycemia. Use of insulin and/or insulin secretagogues was associated with substantially higher hypoglycemia prevalence than use of neither therapy (18.2% vs. 7.4%; OR = 2.79; $p < 0.001$). Secretagogue use alone remained associated with a history of hypoglycemia (15.2% vs. 7.4%; OR = 2.26; $p < 0.001$), whereas the interaction between insulin and secretagogue use was not significant. Beyond treatment exposure, a family history of diabetes was associated with hypoglycemia history (OR = 1.84; 95% CI = 1.47–2.32; $p < 0.001$). Chronic diabetic complications were likewise associated with higher odds (OR = 1.72; $p < 0.001$), including retinopathy (OR = 1.48; $p = 0.003$), peripheral neuropathy (OR = 1.52; $p < 0.001$), autonomic neuropathy (OR = 2.04; $p < 0.001$), and peripheral atherosclerosis (OR = 1.25; $p = 0.048$).

The model-based explanation outputs were broadly consistent with these regression findings while adding patient-level and nonlinear detail. In the shape-stability analysis, the most reproducible continuous feature effects were observed for HbA1c (mean pairwise correlation 0.975), age at diagnosis (0.945), and diabetes duration (0.900), whereas age (0.770) and BMI (0.680) showed lower cross-fold stability. Representative additive decompositions further showed that long diabetes duration, insulin use, secretagogue exposure, and positive family history could appear among the dominant positive contributors in high-risk cases. These results suggest that hypoglycemia history in this cohort reflects both treatment-related exposure and accumulated disease complexity, and that the clinically central predictors remain interpretable at both the cohort and individual levels.

3.2. Main predictive performance

Locked-test performance for all candidate models is summarized in Table 2. CNAM achieved competitive predictive performance while preserving exact additive explanations at both the population and patient levels. Additional analyses of calibration, threshold behavior, decision curves, and stability also showed that discrimination alone was insufficient for evaluating practical use in routine care. Because the ROC-AUC and PR-AUC confidence intervals overlapped across models, the locked-test results should be interpreted as numerically favorable rather than as definitive evidence of statistical superiority.

The ROC and PR curves on the locked test set are shown in Figure 3, confirming that CNAM achieved competitive discrimination relative to the baseline models.

3.3. Calibration and clinical utility

Calibration curves on the locked test set are shown in Figure 4. Post hoc sigmoid calibration improved probability reliability, especially for CNAM. Table S10 reports calibration uncertainty. For calibrated CNAM probabilities, the locked test Brier score was 0.1136 (95% CI, 0.0858–0.1385), and the expected calibration error was 0.0189 (95% CI, 0.0151–0.0677) using 10 equal-width bins. The calibrated CNAM expected calibration error varied with the binning strategy, from 0.0189 with 10 equal-width bins to 0.0566 with 15 equal-mass bins (Table S11). Figure S5 provides a smoothed calibration curve with bootstrap confidence bands.

Because missing vulnerable patients may be more clinically concerning than unnecessary review, we also examined sensitivity-oriented CNAM operating points. A development out-of-fold threshold targeting approximately 90% sensitivity was 0.0888. Applied unchanged to the locked test set, this threshold achieved 92.3% sensitivity, 29.7% specificity, 18.8% positive predictive value, and 95.7% negative predictive value (Table S12). These thresholds are operating-point sensitivity analyses, not clinical deployment recommendations.

The corresponding decision curves are shown in Figure 5. CNAM and logistic regression had broadly similar net benefit profiles across the evaluated threshold range, consistent with their overlapping discrimination. At a threshold probability of 0.15, the net benefit difference between CNAM and logistic regression was 0.0198 (95% CI, 0.0032–0.0390). At 0.10, 0.20, and 0.30, the differences were small, and their confidence intervals crossed 0 (Table S13). In this analysis, CNAM's practical value therefore rests mainly on combining calibrated probabilities with exact additive explanations and constrained shape plausibility, rather than on a large decision-curve gain over logistic regression. Figure S6 shows the calibration-perturbation sensitivity analysis.

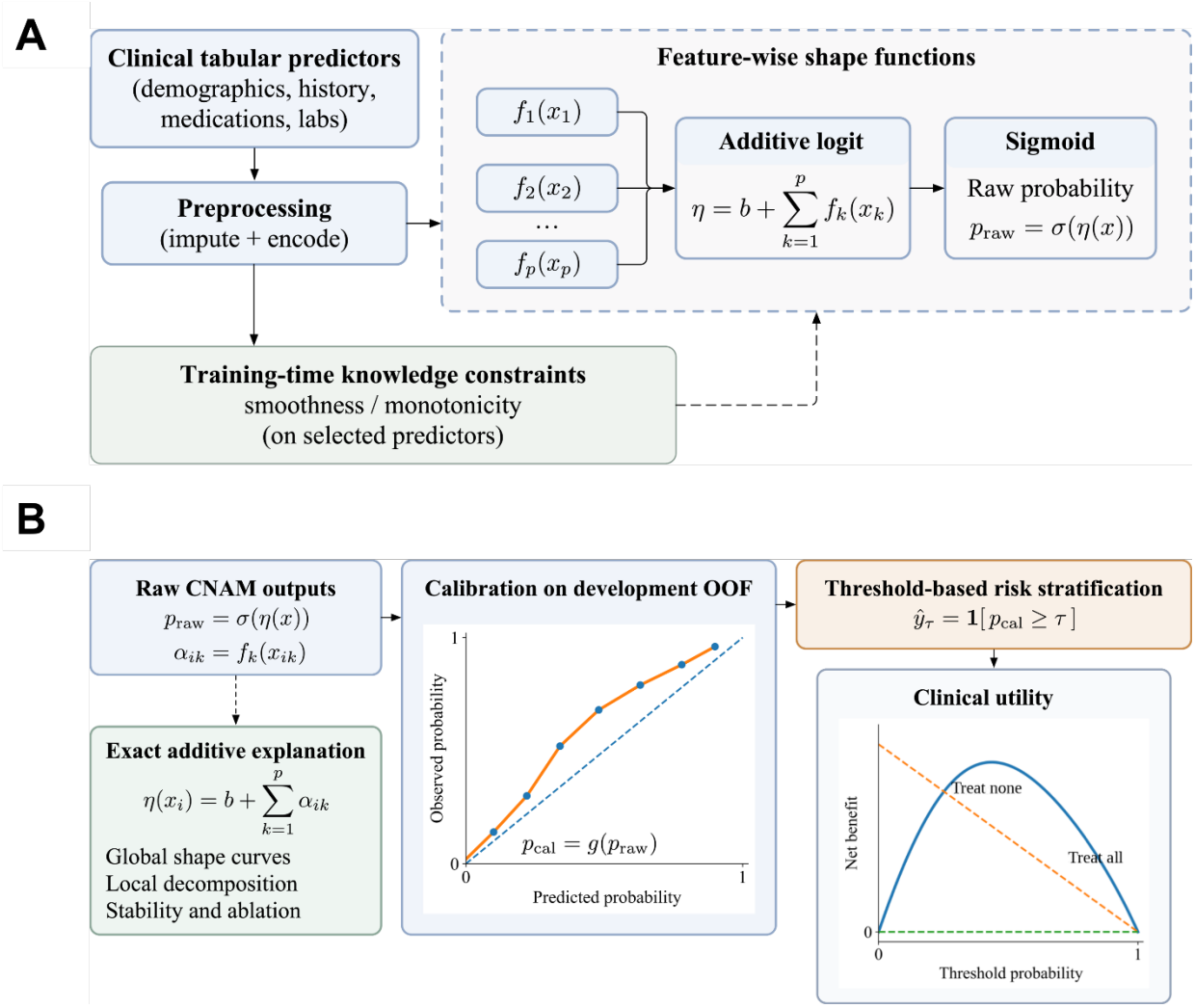


Figure 1. Framework-level overview of the proposed knowledge-constrained neural additive model (CNAM). (A) Training-time additive modeling from clinical tabular predictors to the raw probability, with smoothness and monotonicity constraints applied to feature-wise shape-function learning. (B) Decision-support-oriented outputs of the framework, including probability calibration, threshold-based risk stratification, clinical utility analysis, and exact additive explanations.
Abbreviation: OOF: Out-of-fold.

Table 1. Baseline characteristics of older adults with type 2 diabetes stratified by hypoglycemia status

Characteristics	Hypoglycemia history (*n* = 393)	No hypoglycemia history (*n* = 2,210)	*p* value
Age, years	69.74 ± 7.55	70.17 ± 7.83	0.31
Female, *n* (%)	183 (46.6)	1055 (47.7)	0.67
Male, *n* (%)	210 (53.4)	1155 (52.3)	0.67
BMI, kg/m^2	24.93 ± 3.75	25.17 ± 3.75	0.24
HbA1c, %	8.55 ± 4.86	8.74 ± 3.78	0.40

Abbreviations: BMI: Body mass index; HbA1c: Glycated hemoglobin.

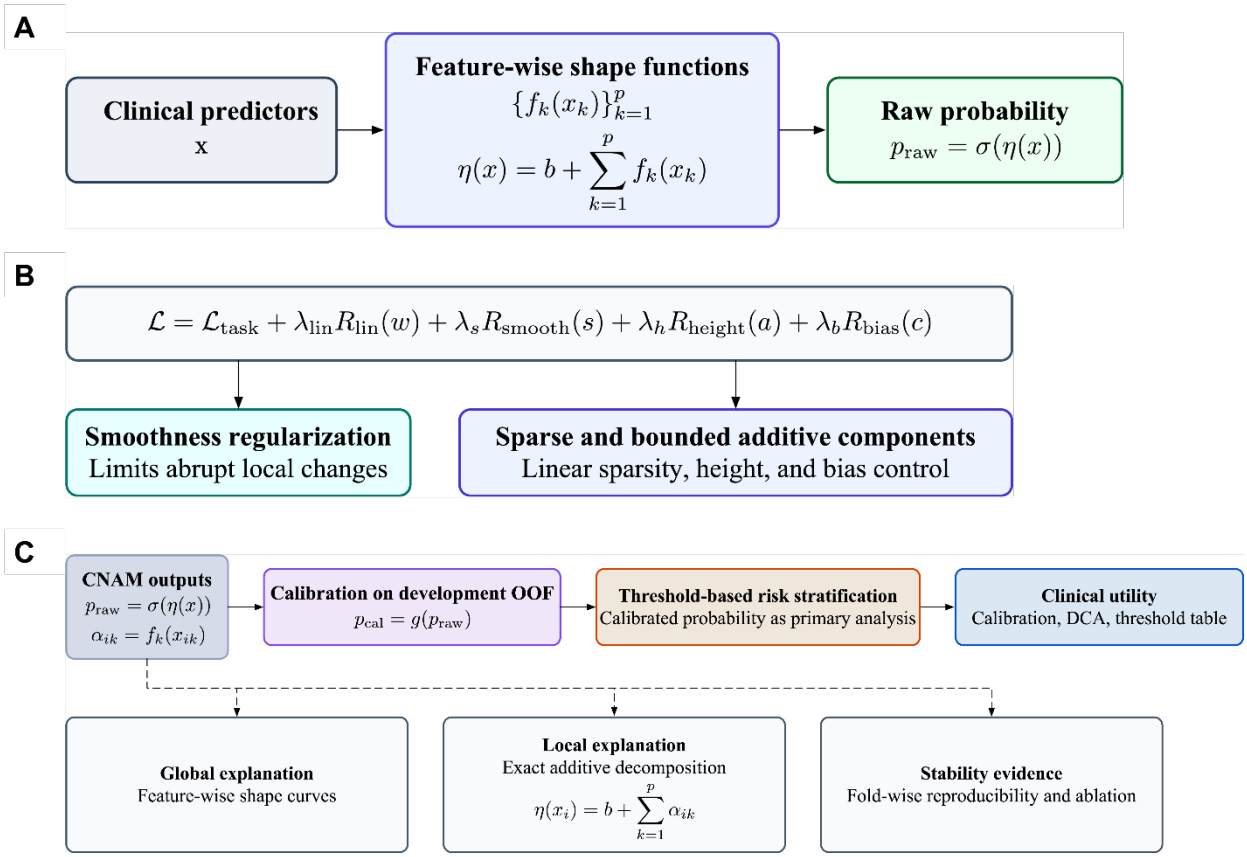


Figure 2. Detailed schematic view of the proposed knowledge-constrained neural additive model (CNAM) modeling framework. (A) Step-based additive representation from clinical predictors to the additive logit and raw probability. (B) Knowledge-guided learning through a constrained objective with linear sparsity, smoothness, amplitude, and bias penalties. (C) Compact decision-support-oriented summary of calibrated decision support and structural explanation.
Abbreviations: DCA: Decision curve analysis; OOF: Out-of-fold.

Table 2. Locked-test performance of logistic regression, gradient-boosted trees, explainable boosting machine, and the proposed CNAM

Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Sens.	Spec.	PPV	NPV
Logistic regression	0.7300 (0.6460–0.8100)	0.3991 (0.2580–0.5530)	0.6410	0.6126	0.2252	0.9067
Tree	0.7167 (0.6170–0.8100)	0.3411 (0.2210–0.5020)	0.7179	0.6171	0.2478	0.9257
EBM	0.7323 (0.6420–0.8140)	0.4002 (0.2610–0.5620)	0.7436	0.5991	0.2458	0.9301
CNAM	0.7350 (0.6470–0.8160)	0.4127 (0.2620–0.5570)	0.8205	0.5405	0.2388	0.9449

Note: Threshold-dependent metrics were computed using calibrated probabilities at the primary operating point.
Abbreviations: CI: Confidence interval; CNAM: Knowledge-constrained neural additive model; EBM: Explainable boosting machine; NPV: Negative predictive value; PPV: Positive predictive value; PR-AUC: Precision–recall-area under the curve; ROC-AUC: Receiver operating characteristic-area under the curve; Sens.: Sensitivity; Spec.: Specificity.

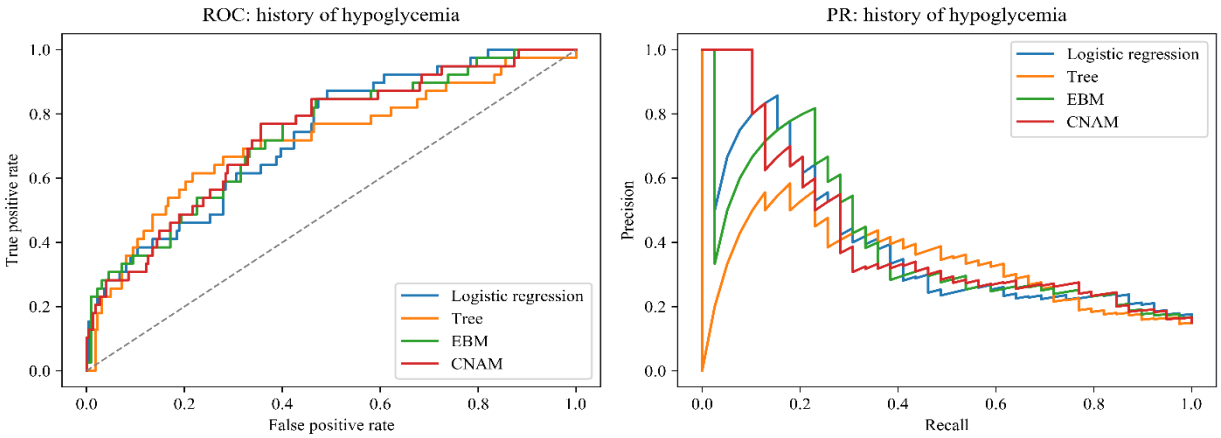


Figure 3. Receiver operating characteristic (ROC) and precision–recall (PR) curves on the locked test set for logistic regression, gradient-boosted trees, explainable boosting machine (EBM), and the proposed knowledge-constrained neural additive model (CNAM)

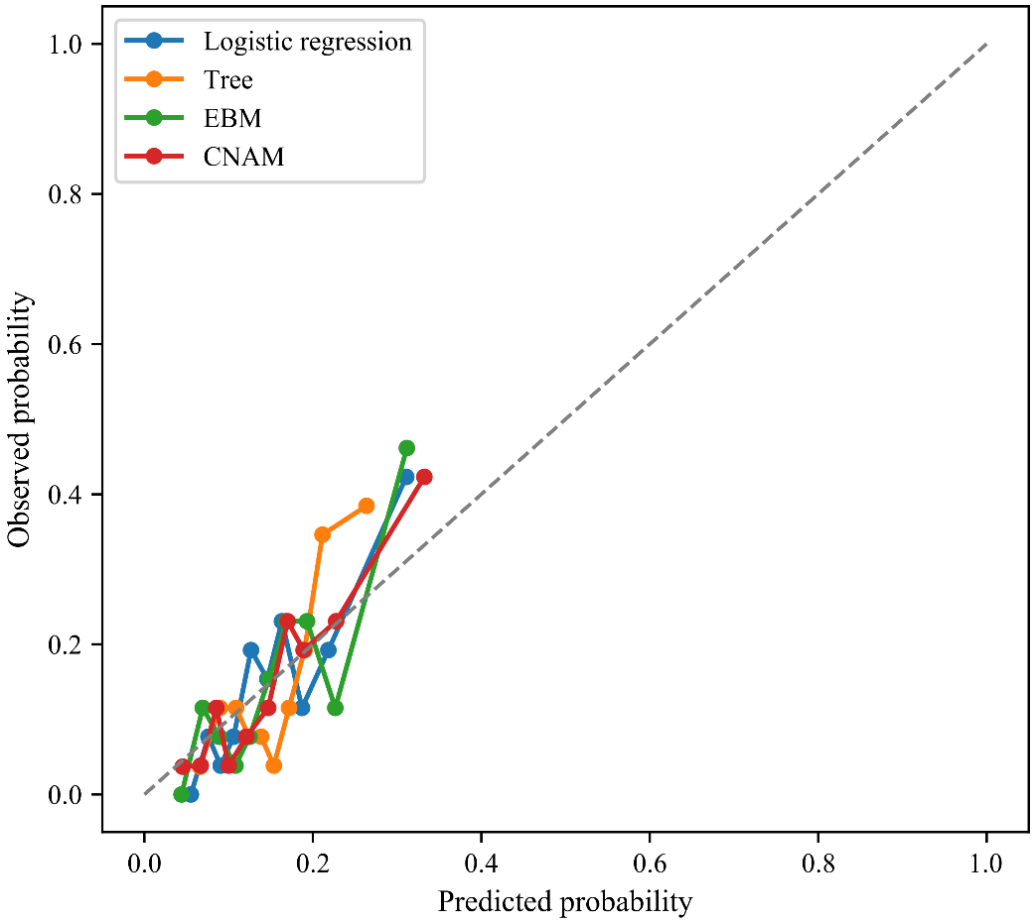


Figure 4. Calibration curves on the locked test set after sigmoid calibration fitted on development out-of-fold predictions. Post hoc calibration improved probability reliability, especially for CNAM.
Abbreviations: CNAM: Knowledge-constrained neural additive model; EBM: Explainable boosting machine.

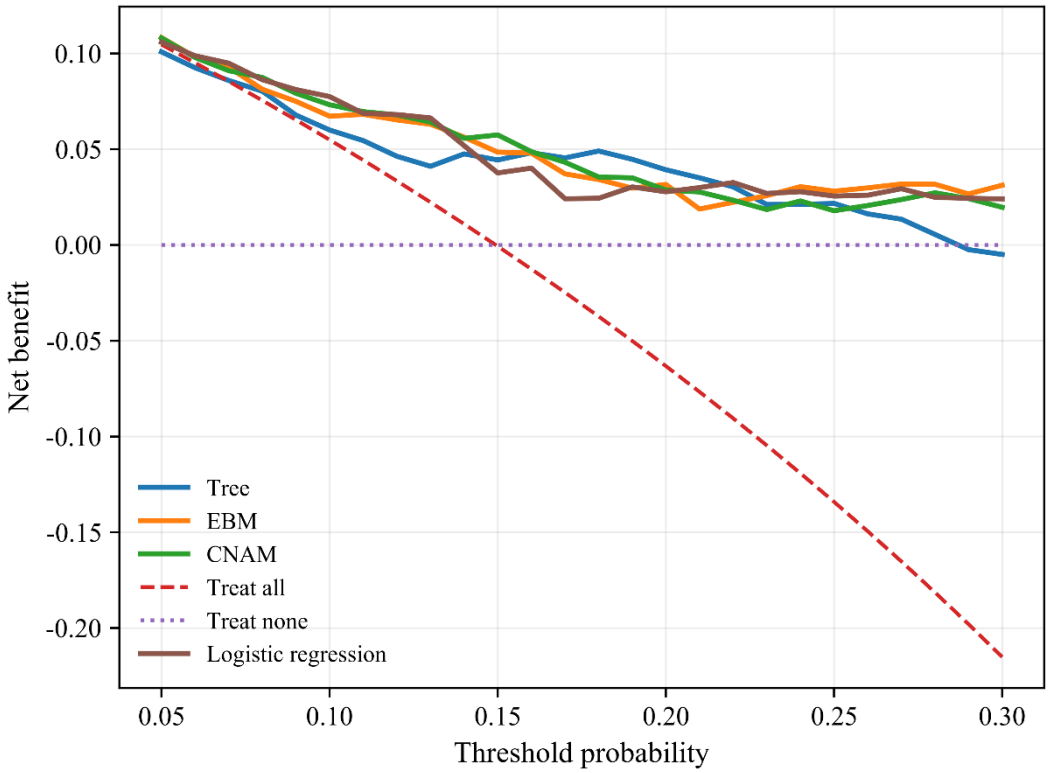


Figure 5. Decision curve analysis on the locked test set. CNAM and logistic regression had broadly similar net-benefit profiles across the evaluated threshold range, and calibrated CNAM retained positive net benefit across clinically relevant threshold probabilities.
Abbreviations: CNAM: Knowledge-constrained neural additive model; EBM: Explainable boosting machine.

3.4. Constraint ablation and explanation stability

Ablation analysis showed that knowledge-based constraints improved explanation behavior without a significant loss in predictive performance. Compared to the unconstrained NAM variant, the smoothness-only and smoothness-plus-monotonicity CNAM variants produced similar test ROC-AUC values, and the monotonicity-constrained model achieved the highest cross-validated PR-AUC among the three ablation settings. The zero-violation rate of the monotonicity-constrained variant is an expected result of the constrained parameterization, not an independent empirical performance gain. Its role is to keep the selected feature effects clinically plausible in direction. By contrast, the unconstrained NAM and smoothness-only CNAM variants showed nonzero monotonicity violation rates on the same features, providing context for the usefulness of the constraint in explanation plausibility (Table S9).

The stability of the learned CNAM shape functions across outer folds is shown in Figure 6. Pairwise fold-wise correlations were highest for HbA1c (0.975), age at diagnosis (0.945), and diabetes duration (0.900), indicating reproducible nonlinear patterns for these clinically central predictors. Stability was lower for age (0.770) and BMI (0.680), so these relationships should be interpreted more cautiously. Supplementary comparison with EBM showed a feature-dependent pattern: CNAM was more stable for HbA1c, age at diagnosis, and diabetes duration, broadly similar for BMI, and less stable for age. The constraints imposed on age, diabetes duration, and age at diagnosis were based on a priori clinical expectations. The monotonicity analysis is therefore reported as a constraint-adherence check within the fitted model class, rather than as evidence of independent predictive improvement. De-identified case-level explanations still showed exact additive decomposition at the patient level, allowing clinicians to inspect the feature contributions underlying individual predictions.

3.5. Additional robustness analyses

Supplementary analyses provide additional checks for the main findings. Table S1 reports the ablation comparison of CNAM variants. Table S2 and Figure S1 summarize subgroup PR-AUC patterns across demographic and medication-defined strata, although some subgroup estimates had wide confidence intervals. Table S3 and

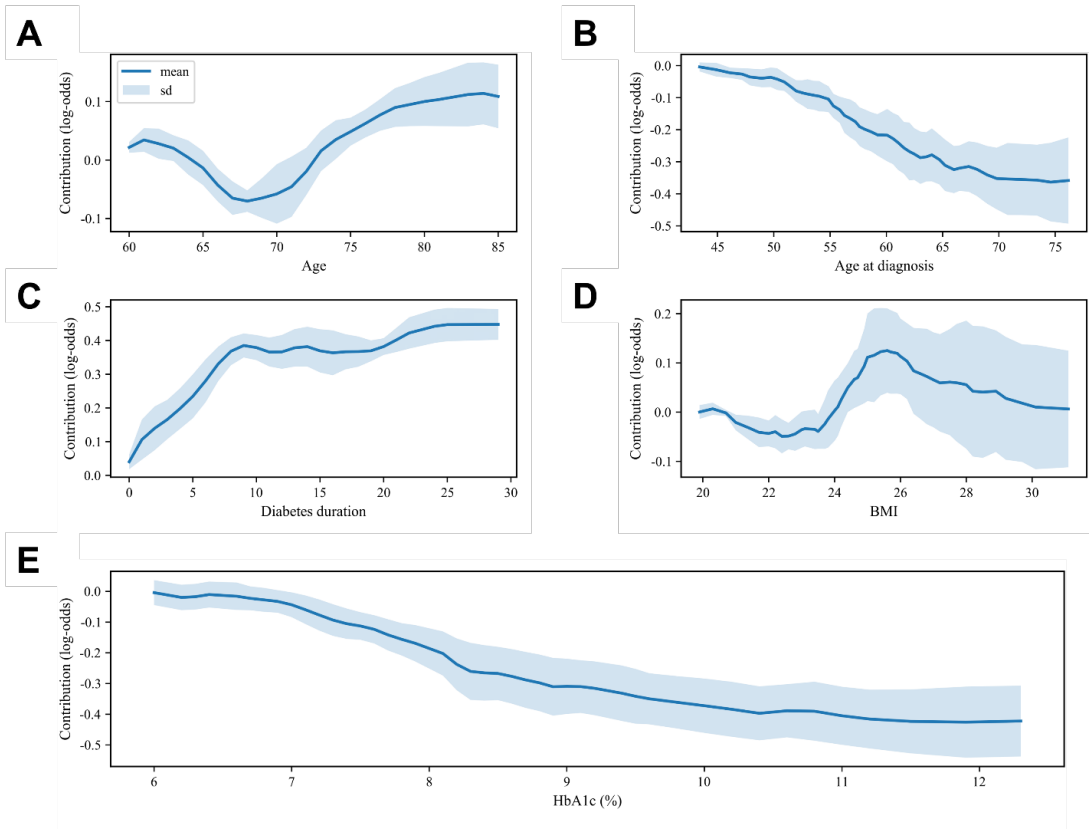


Figure 6. Stability of knowledge-constrained neural additive model (CNAM) shape functions across outer folds for key continuous predictors. (A) Age. (B) Age at diagnosis. (C) Diabetes duration. (D) Body mass index (BMI). (E) Glycated hemoglobin (HbA1c). The most clinically salient feature effects remained reproducible across repeated model fits.

Figure S2 summarize calibration-metric sensitivity and decision-curve sensitivity under sigmoid and isotonic mapping. Table S4 summarizes seed sensitivity under the fixed data split. Figure S3 compares CNAM and EBM global shape functions for key numerical predictors. Additional analyses showed that raw variable-level missingness was limited to BMI (5.07%), HbA1c (3.80%), and age at diagnosis (1.73%) (Table S5); MICE and median imputation produced only modest differences for the evaluated baseline models (Table S6); no participant exceeded the prespecified > 20% missing-predictor exclusion threshold (Table S7); the CNAM-versus-EBM shape-stability comparison was feature-dependent (Table S8); and the monotonicity-constrained variant achieved zero violation rate on the evaluated constrained features (Table S9). Figure S4 reports the seed-sensitivity plots. Note S1 and Algorithm S1 clarify the definition of the recorded-history outcome, the severity-level limitation, and the pooled development out-of-fold calibration pipeline. Tables S10–S13 report calibration confidence intervals, ECE binning sensitivity, alternative CNAM operating points, and net-benefit differences between CNAM and logistic regression. Figures S5 and S6 provide the smoothed calibration curve with bootstrap confidence bands and the calibration-perturbation decision-curve sensitivity analysis.

4. Discussion

This study addresses an important practical gap in hypoglycemia risk stratification for older adults with type 2 diabetes. Rather than proposing another discrimination-only predictor, it develops a clinically adapted additive neural framework in which domain knowledge, explanation structure, and decision-support-oriented probability evaluation are designed together. The results suggest that considering predictive performance, calibration, and explainability jointly may be more informative than treating them as independent add-ons in this clinical setting.

With broader external and temporal validation, the proposed framework could help inform future decision-support research for identifying older adults with type 2 diabetes who appear more vulnerable to hypoglycemia

and who may warrant closer medication review or individualized treatment reassessment. At present, its value is limited to interpretable stratification, explanation of contributing factors, and support for individualized treatment review within the present study setting.

Several clinical predictors were consistently associated with hypoglycemia history. Younger age at diagnosis and longer diabetes duration likely capture cumulative disease burden and prolonged exposure to hypoglycemia-prone treatment strategies. Medication exposure was particularly important: insulin and insulin secretagogues were strongly associated with a history of hypoglycemia, reinforcing the practical importance of treatment review in older adults. Associations with family history and chronic complications further suggest that hypoglycemia vulnerability reflects not only current treatment but also broader disease phenotype and complication burden.

Structurally, the proposed CNAM acts as a clinically oriented middle ground between conventional interpretable methods and purely accuracy-driven black-box predictors. It did not dramatically outperform every comparator on every metric, but it achieved competitive discrimination with numerically favorable point estimates on the locked test set while offering exact feature-wise decomposition and more clinically plausible response shapes than an unconstrained additive alternative. Given the substantial overlap in confidence intervals, these findings should be interpreted as evidence of a favorable balance between interpretability and performance rather than statistically confirmed superiority.

Calibration was central to whether the model could be used for decision support. Raw CNAM probabilities were poorly calibrated, whereas sigmoid calibration reduced the Brier score from 0.209 to 0.114 and the expected calibration error from 0.305 to 0.019 on the locked test set. This result illustrates why probability reliability has to be checked directly, not inferred from discrimination. It is also consistent with recent clinical AI evaluations, which show that highly parameterized models can remain poorly calibrated on structured prediction tasks unless probability reliability is explicitly assessed.[21] Related work on bias in medical AI likewise highlights that unreliable or distorted probability estimates can adversely affect downstream clinical decision-making, even when ranking performance appears acceptable.[22] The raw CNAM probability scale was not clinically usable before calibration. Its Brier score of 0.2085 exceeded the no-information Brier score implied by the observed 15.1% prevalence, approximately 0.128, indicating that uncalibrated probabilities could be misleading despite acceptable ranking performance. CNAM should therefore not be used as a probability-generating clinical decision-support tool without independent validation and, ideally, local recalibration in the target clinical setting.

The explainability analyses further support the framework's interpretability rationale. The constrained additive structure yields exact feature-wise contributions rather than post hoc approximation, while the stability analysis shows that several clinically salient feature effects remain reproducible across repeated model fits. This emphasis on intrinsic rather than post hoc interpretability is also consistent with recent critiques of digital-health AI, which argue that unstable surrogate explanations can mislead high-stakes clinical reasoning.[23] Recent clinician-facing discussions in medical AI likewise emphasize that interpretable predictions are especially important when model outputs may influence downstream care decisions.[24] In addition, broader reviews of explainable machine learning in clinical disciplines highlight both the promise and the limitations of opening black-box models for medical use.[10] In other words, the proposed model is interpretable not only because it can be visualized, but also because explanation stability can be examined quantitatively and compared across resampling and ablation settings.

The present study nevertheless includes several design elements intended to improve internal evaluation rigor within the current cohort, including a large multicenter sample of older adults, clinically accessible predictors, a locked hold-out test set, nested cross-validation, explicit calibration analysis, decision curve analysis, and multiple robustness checks, even though broader generalizability still requires external and temporal validation.

These findings should nevertheless be interpreted in light of several limitations. First, the data were cross-sectional and based on recorded hypoglycemia history rather than prospectively observed future events, so temporal precedence cannot be established. Some predictors, including treatment-related variables, may partly reflect responses to prior hypoglycemia rather than pure antecedent risk markers. The model should therefore be interpreted as an interpretable stratification tool for hypoglycemia history and vulnerability within the observed cohort, not as a causal model or a temporally validated predictor of future events. Second, the cohort was derived from Chinese outpatient settings only and was not evaluated in independent external, temporal, inpatient, nursing-home, or rural primary-care cohorts, so generalizability and broader clinical applicability remain to be established. This boundary is also consistent with emerging dataset-equity guidance, which emphasizes that clinically used AI systems should be evaluated across diverse populations and care settings before broader

adoption.[25] Third, the locked test set contained a limited number of positive cases, which increases uncertainty in both discrimination and calibration estimates. Fourth, review-driven missing-data audits showed limited variable-level missingness and materially unchanged conclusions under the evaluated sensitivity analyses, but the current paper-facing model package was built from a preprocessed complete train/test representation, so future end-to-end pipelines should preserve missing-data handling transparently from the raw source layer. Finally, several clinically relevant confounders, including renal function, cognitive impairment, frailty, nutritional status, and social context factors, were unavailable or inconsistently captured. The learned feature effects should therefore be interpreted as associations within the observed feature space rather than fully adjusted causal effects. Recent geriatric diabetes literature further underscores that frailty, cognitive decline, and other late-life vulnerability states can substantially reshape both hypoglycemia susceptibility and treatment targets, reinforcing the importance of treating these omitted factors as unresolved sources of confounding in the present analysis.[26,27] Recent elderly-focused inpatient hypoglycemia models likewise suggest that predictor salience and model behavior can differ across care settings, further reinforcing the need for broader validation beyond the current outpatient cohort.[28,29]

Recorded hypoglycemia history also depends on ascertainment. It may reflect the actual occurrence of an event, patient recall, glucose monitoring behavior, symptom recognition, and clinical documentation practices. In older adults, nocturnal hypoglycemia and impaired symptom awareness may further contribute to under-ascertainment. The observed 15.1% prevalence may therefore underestimate the true latent burden of hypoglycemia. The learned shape functions should be interpreted as associations with recorded hypoglycemia history, not with all underlying hypoglycemic events. The calibration intercept may also reflect local monitoring and documentation practices, so absolute probability estimates require external validation and local recalibration before use in other settings.

The single 3.9 mmol/L clinical alert threshold also means that the outcome does not distinguish among Level 1, Level 2, and Level 3 hypoglycemia. Participants with different severity profiles may therefore be grouped into one composite endpoint. This limitation prevents severity-specific modeling and should be addressed in future datasets that record biochemical severity, severe episodes requiring assistance, and nocturnal events.

Future work should prioritize external and temporal validation, richer confounder capture, clinician-centered assessment of explanation usefulness, and broader benchmarking against recent biomedical and bioinformatics machine-learning studies in clinically diverse settings. Recent territory-wide modeling of severe hypoglycemia in older adults also highlights the value of comparing interpretable additive approaches with larger, more heterogeneous EHR-based systems[6], while current trustworthy-healthcare AI guidance emphasizes robustness, usability, and post-deployment monitoring as essential next steps for clinically oriented prediction models.[30] Recent monotonic and interpretable deep-tabular architectures further indicate that clinically constrained feature behavior remains an active methodological direction beyond the present model family.[31,32] In addition, the reviewer-highlighted bioinformatics studies on graph-convolutional and neighborhood-regularized representation learning illustrate a broader biomedical context in which structurally constrained machine learning may support future multimodal extensions beyond the present tabular setting.[33,34] Broader recent AI-in-health literature likewise emphasizes that successful clinical translation depends on validation across heterogeneous cohorts, institutions, and workflow contexts, particularly when hidden demographic structure may affect apparent transportability.[35,36] Current methodological guidance also makes clear that internal reproducibility and external transportability are distinct goals, and that clinically relevant prediction models require dedicated external-validation studies beyond the present phase of development.[37] Recent international guidance on predictive medical AI further emphasizes that calibration, decision consequences, and operating-threshold behavior should be evaluated alongside discrimination when models are intended to inform medical decisions.[38] Related work on probabilistic transport failure also shows how latent variables and site-specific proxy structure can limit portability across institutions[39], while recent geriatric guideline synthesis further reinforces the need for individualized treatment targets in older and frail adults with type 2 diabetes.[40]

5. Conclusion

This study presents a CNAM framework for clinically interpretable stratification of recorded hypoglycemia history in older adults with type 2 diabetes. In this cross-sectional outpatient cohort, constrained additive modeling combined competitive discrimination, improved calibration after post hoc probability mapping, and structurally grounded explanations. These findings support knowledge-constrained additive modeling as a reasonable option for interpretable stratification when both transparent reasoning and probability reliability are

needed, while severity-specific outcome capture, external validation, temporal validation, and local recalibration remain necessary before broader clinical use.

Acknowledgments

We are grateful to the Committee on Elderly Endocrine and Metabolism of the Chinese Geriatric Society and to all the participants in this study for their valuable contributions.

Funding

This study was funded by the Science and Technology Project for the High-Quality Development of China Medical University, with support from the Department of Biotechnology Development and Industrialization (grant 2023JH2/20200110).

Conflict of interest

The authors declare no competing interests.

Author contributions

Conceptualization: Difei Wang, Longfeng Sun, Hui Tian, Shuangtong Yan, Xinyu Miao
Formal analysis: Zheng Rong, Hui Si, Wei Bai, Xin Kang, Xin Hou
Investigation: Difei Wang, Longfeng Sun, Hui Tian, Shuangtong Yan, Xinyu Miao
Methodology: Difei Wang, Longfeng Sun, Hui Tian, Shuangtong Yan, Xinyu Miao
Writing–original draft: Zheng Rong, Hui Si, Wei Bai
Writing–review & editing: Xin Kang, Xin Hou
All authors approved the final manuscript.

Ethics approval and consent to participate

This multicenter cross-sectional study was approved by the Medical Ethics Committee of the Chinese PLA General Hospital under approval number S2015-038-01. Written informed consent was obtained from all participants before study participation.

Consent for publication

Not applicable, as no individually identifiable data, images, or personal information are included in this manuscript.

Availability of data

De-identified data are not publicly available because of privacy and institutional restrictions. Analysis code and configuration files are available from the corresponding author upon reasonable request.

Further disclosure

During the revision process, AI-assisted tools were used only to support language editing, readability improvement, and organization of reviewer-response drafts. All scientific content, data analysis, interpretation, conclusions, and final decisions were reviewed and approved by the authors.

References

1. ElSayed NA, Aleppo G, Bannuru RR, *et al.* 6. Glycemic goals and hypoglycemia: Standards of Care in Diabetes—2024. *Diabetes Care*. 2024;47(suppl 1):S111-S125.

   doi: 10.2337/dc24-S006

2. Ahmad E, Lim S, Lamptey R, Webb DR, Davies MJ. Type 2 diabetes. *Lancet*. 2022;400(10365):1803-1820.

   doi: 10.1016/S0140-6736(22)01655-5

3. Davies MJ, Lim S, Slater T, *et al.* Type 2 diabetes mellitus. *Nat Rev Dis Primers*. 2026;12(1):13.

   doi: 10.1038/s41572-026-00687-w

4. Tan SHC, Loh WJ, Lim SC. Precision medicine in diabetes care. *Curr Opin Endocrinol Diabetes Obes*. 2025;32(1):12-19.

   doi: 10.1097/MED.0000000000000894

5. Mellor J, Kuznetsov D, Heller S, *et al.* Risk factors and prediction of hypoglycaemia using the Hypo-RESOLVE cohort: a secondary analysis of pooled data from insulin clinical trials. *Diabetologia*. 2024;67(9):1873-1885.

   doi: 10.1007/s00125-024-06177-6

6. Shi M, Yang A, Lau ESH, *et al.* A novel electronic health record-based, machine-learning model to predict severe hypoglycemia leading to hospitalizations in older adults with diabetes: a territory-wide cohort and modeling study. *PLoS Med*. 2024;21(4):e1004369.

   doi: 10.1371/journal.pmed.1004369

7. Mathioudakis NN, Abusamaan MS, Shakarchi AF, *et al.* Development and validation of a machine learning model to predict near-term risk of iatrogenic hypoglycemia in hospitalized patients. *JAMA Netw Open*. 2021;4(1):e2030913.

   doi: 10.1001/jamanetworkopen.2020.30913

8. Kulzer B, Freckmann G, Ziegler R, Schnell O, Glatzer T, Heinemann L. Nocturnal hypoglycemia in the era of continuous glucose monitoring. *J Diabetes Sci Technol*. 2024;18(5):1052-1060.

   doi: 10.1177/19322968241267823

9. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215.

   doi: 10.1038/s42256-019-0048-x

10. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can J Cardiol*. 2022;38(2):204-213.

doi: 10.1016/j.cjca.2021.09.004

11. Ling S, Zaccardi F, Lawson C, Seidu SI, Davies MJ, Khunti K. Glucose control, sulfonylureas, and insulin treatment in elderly people with type 2 diabetes and risk of severe hypoglycemia and death: an observational study. *Diabetes Care*. 2021;44(4):915-924.

doi: 10.2337/dc20-0876

12. Shabestari M, Mehrabbeik A, Barbieri S, Marques-Vidal P, Heshmati-Nasab P, Azizi R. Predictive factors of hypoglycemia in type 2 diabetes: a prospective study using machine learning. *Sci Rep*. 2025;15(1):18143.

doi: 10.1038/s41598-025-03030-7

13. Hastie T, Tibshirani R. Generalized additive models. *Stat Sci*. 1986;1(3):297-310.

doi: 10.1214/ss/1177013604

14. Agarwal R, Melnick L, Frosst N, *et al.* Neural additive models: interpretable machine learning with neural nets. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. 2021;359:4699-4711.

15. Nori H, Caruana R, Bu Z, Shen JH, Kulkarni J. Accuracy, interpretability, and differential privacy via explainable boosting. In: *Proceedings of the 38th International Conference on Machine Learning*. Proceedings of Machine Learning Research; 2021:8227-8237.

16. You S, Ding D, Canini K, Pfeifer J, Gupta M. Deep lattice networks and partial monotonic functions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. 2017:2981-2989.

17. Mariotti E, Moral JMA, Gatt A. Exploring the balance between interpretability and performance with carefully designed constrainable neural additive models. *Inf Fusion*. 2023;99:101882.

doi: 10.1016/j.inffus.2023.101882

18. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.

doi: 10.1186/s12916-019-1466-7

19. Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.

doi: 10.1136/bmj-2023-078378

20. Moons KGM, Damen JAA, Kaul T, *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505.

doi: 10.1136/bmj-2024-082505

21. Brown KE, Yan C, Li Z, *et al.* Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *J Am Med Inform Assoc*. 2025;32(5):811-822.

doi: 10.1093/jamia/ocaf038

22. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. *PLoS Digit Health*. 2024;3(11):e0000651.

doi: 10.1371/journal.pdig.0000651

23. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750.

doi: 10.1016/S2589-7500(21)00208-9

24. Lu SC, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Front Oncol*. 2023;13:1129380.

doi: 10.3389/fonc.2023.1129380

25. Ganapathi S, Palmer J, Alderman JE, *et al.* Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med*. 2022;28(11):2232-2233.

doi: 10.1038/s41591-022-01987-w

26. Christiaens A, Sheppard JP. Risks of tight glucose control in older adults: time to focus beyond mortality. *Diabetes Care*. 2025;48(10):1716-1718.

doi: 10.2337/dci25-0068

27. Alkabbani W, Maxwell CJ, Marrie RA, *et al.* Associations of mid- and late-life severe hypoglycemic episodes with incident dementia among patients with type 2 diabetes: a population-based cohort study. *Diabetes Care*. 2023;46(2):331-340.

doi: 10.2337/dc22-1496

28. Zhang RT, Liu Y, Sun C, *et al.* Predicting hypoglycemia in elderly inpatients with type 2 diabetes: the ADOCHBIU model. *Front Endocrinol*. 2024;15:1366184.

doi: 10.3389/fendo.2024.1366184

29. Jia H, Zhang J. Machine learning-based prediction of hypoglycemia severity in hospitalized diabetic patients. *Front Endocrinol*. 2025;16:1634358.

doi: 10.3389/fendo.2025.1634358

30. Lekadir K, Frangi AF, Porras AR, *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554.

doi: 10.1136/bmj-2024-081554

31. Nguyen AP, Moreno DL, Le-Bel N, Rodríguez Martínez M. MonoNet: enhancing interpretability in neural networks via monotonic features. *Bioinform Adv*. 2023;3(1):vbad016.

doi: 10.1093/bioadv/vbad016

32. Richman R, Wüthrich MV. LocalGLMnet: interpretable deep learning for tabular data. *Scand Actuar J*. 2023;2023(1):71-95.

doi: 10.1080/03461238.2022.2081816

33. Kim K, Ha J. GMFLDA: improved prediction of lncRNA-disease association via graph convolutional network. *IEEE Access*. 2025;13:85330-85341.

doi: 10.1109/ACCESS.2025.3568461

34. Ha J, Kim K. Neighborhood-regularized matrix factorization for lncRNA-disease association identification. *Int J Mol Sci*. 2025;26(9):4283.

doi: 10.3390/ijms26094283

35. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38.

doi: 10.1038/s41591-021-01614-0

36. Gichoya JW, Banerjee I, Bhimireddy AR, *et al.* AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. 2022;4(6):e406-e414.

doi: 10.1016/S2589-7500(22)00063-2

37. Riley RD, Archer L, Snell KIE, *et al.* Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820.

doi: 10.1136/bmj-2023-074820

38. Van Calster B, Collins GS, Vickers AJ, *et al.* Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *Lancet Digit Health*. 2025;7(12):100916.

doi: 10.1016/j.landig.2025.100916

39. Lasko TA, Strobl EV, Stead WW. Why do probabilistic clinical models fail to transport between sites. *npj Digit Med*. 2024;7(1):53.

doi: 10.1038/s41746-024-01037-4

40. Bolt J, Carvalho V, Lin K, Lee SJ, Inglis C. Systematic review of guideline recommendations for older and frail adults with type 2 diabetes mellitus. *Age Ageing*. 2024;53(11):afae259.

doi: 10.1093/ageing/afae259