

AN EXPLAINABLE TWO-STAGE MULTI-DOMAIN MACHINE LEARNING FRAMEWORK FOR PERSONALIZED TYPE 2 DIABETES RISK ASSESSMENT WITH FORWARD TEMPORAL VALIDATION USING NHANES

Apurva Anand Jadhav¹, Pranjali Kasture¹

¹Department of Information Technology, Thakur College of Engineering & Technology, Mumbai, India

*Corresponding author: Apurva Jadhav. Email: apurvajadhav9076@gmail.com

ABSTRACT

Background: Type 2 diabetes mellitus (T2DM) is a heterogeneous metabolic disorder shaped by demographic, anthropometric, hemodynamic, socioeconomic, and lifestyle-related factors. Machine learning methods can integrate these heterogeneous predictors to support individualized risk assessment; however, published prediction studies frequently emphasize discrimination while providing limited evidence on calibration, clinical utility, explainability, and temporal transportability.

Objective: To develop and evaluate an explainable two-stage multi-domain machine learning framework for personalized T2DM risk assessment, quantifying discrimination, calibration, clinical utility, explainability, subgroup performance, and forward temporal transportability across independent NHANES cycles.

Methods: NHANES 2015–2016 was used for model development and NHANES 2017–2018 for forward temporal validation. Adults with an available diabetes-status outcome (NHANES variable DIQ010) and relevant predictor information were included. The Stage-1 model used demographic, anthropometric, blood-pressure, socioeconomic, physical-activity, smoking, and alcohol-related variables while excluding HbA1c and fasting glucose. Feature engineering incorporated clinically motivated anthropometric and hemodynamic interaction terms. An XGBoost classifier was optimized using stratified cross-validation and randomized hyperparameter search, benchmarked against class-weighted XGBoost, LightGBM, CatBoost, and a soft-voting ensemble. The operating threshold was selected from out-of-fold development predictions and then fixed for temporal validation. Performance was evaluated with ROC-AUC, PR-AUC, sensitivity, specificity, F1 score, Brier score, calibration intercept/slope, bootstrap 95% confidence intervals (2,000 resamples), and decision curve analysis. SHAP values were used for interpretation. A secondary, exploratory Stage-2 model added HbA1c and fasting glucose for clinical refinement.

Results: The development cohort comprised 5,856 participants, including 845 with T2DM (14.43%); the temporal validation cohort comprised 5,678 participants from NHANES 2017–2018 (878 events, 15.46%). The Stage-1 model achieved ROC-AUC 0.8083 (95% CI 0.7947–0.8212) and PR-AUC 0.4119 (95% CI 0.3805–0.4453) in forward temporal validation. At the development-derived threshold of 0.2195, sensitivity was 64.46% (95% CI 61.26–67.48%), specificity was 78.33% (95% CI 77.13–79.45%), and F1 was 0.4557 (95% CI 0.4311–0.4795). Brier score was 0.1088; the temporal calibration intercept and slope were 0.1783 and 1.1103, respectively, consistent with modest systematic underestimation of absolute risk alongside a well-preserved risk gradient. Decision curve analysis showed net benefit over treat-all and treat-none strategies across threshold probabilities from approximately 0.010 to 0.705. SHAP analysis ranked age, age×BMI, age×waist-circumference, waist circumference, race/ethnicity, and BMI-to-waist ratio among the leading contributors. In a separate, secondary internal analysis using an 80/20 split of NHANES 2017–2018 alone, adding HbA1c and fasting glucose to the same predictor set (Stage 2) increased ROC-AUC to 0.9186 and PR-AUC to 0.7218; because this analysis used a different cohort/split than the forward-temporal-validation result, it is presented as an exploratory clinical-refinement analysis rather than a direct head-to-head comparison with the Stage-1 temporal-validation estimate.

Conclusions: An explainable multi-domain XGBoost framework maintained stable discrimination across NHANES survey cycles while integrating calibration, threshold optimization, decision-curve analysis, subgroup assessment, and explainable machine learning. These results support the potential utility of multi-domain XGBoost for non-glycemic T2DM status/risk assessment and suggest that glycemic biomarkers can subsequently refine risk estimates, pending confirmation under the same temporal-validation design. Because NHANES is cross-sectional, this study characterizes diabetes-status classification and temporal transportability rather than prospective prediction of incident T2DM.

KEYWORDS: Type 2 diabetes mellitus; machine learning; XGBoost; explainable artificial intelligence; SHAP; risk assessment; temporal validation; NHANES; clinical decision support; calibration

1 INTRODUCTION

1.1 Epidemiological background

Type 2 diabetes mellitus (T2DM) remains one of the most prevalent chronic metabolic disorders worldwide. According to the 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas, global diabetes prevalence continues to rise, with substantial regional heterogeneity and a large proportion of cases remaining undiagnosed [17, 19]. In the United States, National Health and Nutrition Examination Survey (NHANES) data have repeatedly demonstrated that diabetes prevalence differs meaningfully by race, ethnicity, and socioeconomic status, with disproportionately higher burden among non-Hispanic Black, Hispanic, and Asian American subgroups relative to non-Hispanic White adults [16, 20]. This heterogeneity in both risk-factor distribution and disease burden motivates modeling approaches capable of representing nonlinear, multi-domain relationships between predictors and diabetes status, rather than relying on a small number of predefined additive risk factors.

Diabetes risk is influenced jointly by demographic characteristics, adiposity, blood pressure, socioeconomic factors, physical activity, smoking, and alcohol-related behaviors. Consequently, approaches that integrate information across multiple domains may provide a more individualized representation of risk than models built on a single clinical measurement or a narrow risk-factor set.

1.2 Limitations of conventional risk scores

Conventional diabetes risk instruments, such as the Finnish Diabetes Risk Score (FINDRISC), have demonstrated that demographic and anthropometric variables can provide useful information for identifying individuals at elevated risk using simple, additive point systems [14, 21, 22]. Anthropometric measures of central adiposity, particularly waist circumference, have been shown in prospective cohorts to be at least as strongly, and in some analyses more strongly, associated with incident T2DM as overall body-mass index (BMI), reflecting the specific contribution of visceral fat distribution to insulin resistance [15, 23, 24]. However, traditional risk scores generally rely on predefined linear or additive relationships among a limited set of predictors and do not learn interactions directly from data. Machine learning methods provide an alternative by allowing nonlinear relationships and interactions between predictors to be estimated empirically.

1.3 Methodological pitfalls of machine learning risk-prediction studies

Despite this promise, methodological reviews of machine-learning-based clinical prediction research have identified recurring concerns, including inadequate handling of missing data, insufficient separation between model development and evaluation, limited external validation, and inconsistent reporting of model specification [4, 5, 25]. A widely cited systematic review found no consistent performance benefit of machine learning over logistic regression when methodological quality was accounted for, underscoring that algorithmic sophistication does not by itself guarantee a clinically useful model [6]. Reporting and risk-of-bias frameworks such as TRIPOD [5] and PROBAST [3], and their AI-specific extension PROBAST+AI [2], have been developed specifically to address these concerns by requiring transparent reporting of predictor handling, validation design, and applicability. The present study was designed with explicit attention to these recommendations: development and evaluation data were kept strictly separate, the operating threshold was selected using development data only, and forward temporal validation was used to assess transportability across survey cycles, consistent with established guidance on internal, internal-external, and external validation of prediction models [8, 26].

1.4 Two-stage screening: the Khattab et al. framework

A particularly relevant recent study by Khattab et al. developed and externally validated a two-stage machine-learning framework using the All of Us Research Program and UK Biobank [1]. Their Stage-1 model used XGBoost together with demographic, clinical, and polygenic predictors to estimate five-year incident T2DM, while Stage 2 incorporated HbA1c or fasting glucose to refine risk estimates. The reported Stage-1 AUROC was approximately 0.81 in All of Us and 0.82 in UK Biobank, and the study incorporated SHAP-based interpretation and subgroup evaluation. This two-stage structure provides a useful conceptual basis for distinguishing early assessment using information available without direct glycemic measurement from subsequent clinical refinement using glycemic biomarkers. The present study differs from Khattab et al. in several respects: it uses publicly accessible NHANES survey data rather than biobank cohorts with genotype data, it emphasizes multi-domain non-glycemic predictors including socioeconomic and lifestyle variables, it evaluates forward temporal transportability across NHANES survey cycles rather than prospective incident disease, and it explicitly integrates calibration, decision curve analysis, threshold optimization, subgroup analysis, and SHAP-based interpretation within a single, reproducible pipeline built entirely from public data.

1.5 Explainability and clinical utility

High discrimination alone is insufficient for a model intended to support clinical decisions. Explainable artificial intelligence methods such as SHapley Additive exPlanations (SHAP) [10, 27, 28] provide a theoretically grounded way to quantify the contribution of individual predictors to model outputs and can reveal nonlinear relationships and interactions learned by tree-based models. Clinical usefulness is conceptually distinct from statistical discrimination and from calibration: even a well-discriminating and well-calibrated model may not be useful if acting on its predictions does not improve net clinical benefit relative to simpler default strategies. Decision curve analysis (DCA) addresses this by quantifying net benefit across a range of probability thresholds, allowing a model to be compared against treat-all and treat-none strategies [9, 29]. Calibration, in

turn, has been described as an under-examined but critical property of predictive models applied in medicine, since a model that ranks patients well (high discrimination) can nonetheless produce systematically miscalibrated absolute risk estimates [7, 30]. This study therefore evaluates the proposed framework using discrimination, calibration, and decision-analytic criteria jointly, rather than relying on ROC-AUC alone.

1.6 Study objective and contributions

The objective of this study was to develop and evaluate an explainable two-stage multi-domain machine-learning framework for personalized T2DM risk/status assessment using publicly available NHANES data. The specific contributions of this work are:

1. Integration of demographic, anthropometric, blood-pressure, socioeconomic, physical-activity, smoking, and alcohol-related information into a single multi-domain Stage-1 model.
2. Development of a Stage-1 model that deliberately excludes HbA1c and fasting glucose, isolating a non-glycemic early-assessment signal.
3. Feature engineering using nonlinear and interaction-oriented representations of age, adiposity, waist circumference, and blood pressure, motivated by known clinical relationships [14, 15].
4. Leakage-aware hyperparameter optimization and operating-threshold selection using development data only, followed by forward temporal validation from NHANES 2015–2016 to NHANES 2017–2018.
5. Comprehensive evaluation using discrimination, calibration, Brier score, bootstrap confidence intervals, and decision curve analysis, together with SHAP-based interpretation and subgroup performance analysis.
6. A secondary Stage-2 model incorporating HbA1c and fasting glucose to characterize the incremental value of glycemic biomarkers for clinical refinement.

Because NHANES is a cross-sectional survey repeated in independent biennial cycles, the outcome examined in this study represents diabetes status/risk assessment and its transportability across survey cycles, rather than prospective prediction of incident T2DM.

2 MATERIALS AND METHODS

2.1 Study design and two evaluation cohorts

This study used a retrospective machine-learning development design based on publicly available NHANES data [18, 31], combining two distinct evaluation cohorts that are kept explicitly separate throughout this manuscript to avoid conflating them, as follows.

(i) Forward temporal validation cohort. The Stage-1 model was trained on the complete NHANES 2015–2016 cycle (development cohort) and evaluated, without any retraining or refitting, on the complete, independently collected NHANES 2017–2018 cycle (temporal validation cohort). This is the primary analysis of the paper and the only analysis in which the evaluation data were collected in a survey cycle entirely subsequent to, and independent of, the training data. All threshold selection and hyperparameter tuning for this analysis used development-cohort data only, consistent with recommendations that temporal (‘internal–external’) validation provides stronger evidence of transportability than a random train/test split drawn from a single dataset [8].

(ii) Internal cross-sectional benchmarking cohort. A second, separate set of analyses, algorithm comparison (original, class-weighted, LightGBM, CatBoost, and ensemble models), Stage-2 clinical refinement, and subgroup performance, was conducted using a single NHANES cycle (2017–2018 predictor and outcome data, $n = 5,678$) split once into an 80% training partition ($n = 4,543$) and a 20% stratified hold-out test partition ($n = 1,136$). This internal benchmarking analysis was used to compare candidate algorithms and to explore the incremental value of glycemic biomarkers, but, unlike (i), it does not constitute an independent-cycle temporal validation, because both the training and test partitions were drawn from the same NHANES cycle. Results from this internal cohort are labeled explicitly as such throughout the Results section and are not combined numerically with the forward-temporal-validation estimates.

2.2 Study population and outcome definition

Adults aged 18 years or older were eligible for inclusion. The diabetes-status outcome was derived from the NHANES diabetes questionnaire variable DIQ010. Participants reporting a diabetes diagnosis were assigned a positive outcome, and participants reporting no diagnosis were assigned a negative outcome; responses corresponding to ‘don’t know’, refused, or other non-classifiable categories were treated as missing for the target variable and excluded. Because the outcome is a self-reported diagnosis rather than an independently adjudicated or longitudinally confirmed diagnosis, this study reports diabetes-status classification/assessment throughout rather than diabetes ‘prediction’ in the prospective sense. The development cohort (NHANES 2015–2016) consisted of 5,856 participants, including 845 with T2DM (prevalence 14.43%). The forward temporal validation cohort (NHANES 2017–2018, full cycle) consisted of 5,678 participants, including 878 with T2DM (prevalence 15.46%). The internal benchmarking cohort used the same 5,678 NHANES 2017–2018 participants, split 80/20 (4,543 training /1,136 held-out test).

2.3 Predictor domains

Predictors were drawn from multiple domains available in NHANES public-use files:

- **Demographic:** age (RIDAGEYR), sex (RIAGENDR), race/ethnicity (RIDRETH3), educational attainment (DMDEDUC2), and income-to-poverty ratio (INDFMPIR).

- **Anthropometric:** body-mass index (BMXBMI), body weight (BMXWT), height (BMXHT), waist circumference (BMXWAIST), and derived obesity/waist indicators.
- **Blood pressure:** mean systolic (SBP_MEAN) and diastolic (DBP_MEAN) blood pressure, pulse pressure, mean arterial pressure (MAP), and elevated/very-high blood-pressure indicators.
- **Lifestyle:** physical-activity questionnaire items (PAQ605–PAQ665), smoking status (SMQ020, SMQ040), and alcohol-use items (ALQ101, ALQ110, ALQ120Q, ALQ130).

2.4 Stage-1 predictor restriction

A central design feature of the framework was the separation of early, non-glycemic risk assessment from direct glycemic measurement. The Stage-1 model therefore excluded HbA1c and fasting plasma glucose, as well as the diabetes-questionnaire items used to define the outcome itself. This restriction distinguishes non-glycemic risk assessment, which could in principle be obtained without a blood draw dedicated to glycemic testing, from current glycemic-status detection, consistent with the Stage-1/Stage-2 conceptual separation proposed by Khattab et al. [1].

2.5 Feature engineering

Additional derived predictors were constructed to capture nonlinear relationships and clinically motivated interactions, consistent with prior evidence that central adiposity and its interaction with age carry information beyond BMI alone [15]. Engineered variables included: obesity and severe-obesity indicators; an elevated-waist indicator; BMI-to-waist ratio; pulse pressure; mean arterial pressure; elevated and very-high blood-pressure indicators; age squared; age $\times$ BMI; age $\times$ waist circumference; BMI $\times$ sex; and BMI $\times$ systolic blood pressure. Missingness indicators were incorporated for selected variables to retain information about the pattern of missing data rather than only its magnitude. Algorithm 1 summarizes the feature-engineering and preprocessing pipeline applied identically to the development and temporal-validation cohorts.

Algorithm 1 Stage-1 feature engineering and leakage-aware preprocessing

Require: Raw NHANES predictor tables D_{train} (2015–2016), D_{valid} (2017–2018)

Ensure: Model-ready design matrices $X_{\text{train}}, X_{\text{valid}}$

- 1: **for** each dataset $D \in \{D_{\text{train}}, D_{\text{valid}}\}$ **do**
 - 2: Restrict D to adults ($\text{RIDAGEYR} \geq 18$)
 - 3: Compute SBP_MEAN, DBP_MEAN as means of repeated readings
 - 4: Derive PULSE_PRESSURE $\leftarrow \text{SBP_MEAN} - \text{DBP_MEAN}$
 - 5: Derive MAP $\leftarrow \text{DBP_MEAN} + (\text{SBP_MEAN} - \text{DBP_MEAN})/3$
 - 6: Derive OBESITY, SEVERE_OBESITY, HIGH_WAIST indicators from BMXBMI, BMXWAIST
 - 7: Derive BMI_WAIST_RATIO $\leftarrow \text{BMXBMI} / \text{BMXWAIST}$
 - 8: Derive HIGH_BP, VERY_HIGH_BP indicators from SBP_MEAN, DBP_MEAN
 - 9: Derive AGE_SQUARED, AGE_BMI, AGE_WAIST, BMI_SEX, BMI_SBP interaction terms
 - 10: Add a missingness indicator for each predictor with missing values
 - 11: Exclude LBXGH (HbA1c), LBXGLU (fasting glucose), and DIQ0* items
 - 12: **end for**
 - 13: One-hot encode categorical predictors (drop-first) using categories observed in D_{train}
 - 14: Fit median imputer M on D_{train} only
 - 15: $X_{\text{train}} \leftarrow M.\text{transform}(D_{\text{train}})$; $X_{\text{valid}} \leftarrow M.\text{transform}(D_{\text{valid}})$ $\{M$ is NOT refit on $D_{\text{valid}}\}$
 - 16: **return** $X_{\text{train}}, X_{\text{valid}}$
-

2.6 Data preprocessing

Categorical predictors were converted into indicator variables using one-hot encoding (drop-first parameterization). Missing numerical values were imputed using median imputation with accompanying missingness indicators. The imputation transformation was fitted exclusively on the development data (NHANES 2015–2016) and subsequently applied unchanged to the temporal validation observations (NHANES 2017–2018), so that no information from the validation cohort influenced the imputation, scaling, or encoding of the development data (Algorithm 1, lines 13–14).

2.7 Model development and algorithm comparison (internal benchmarking cohort)

XGBoost [11, 32, 33, 34] was selected as the primary learning algorithm because of its strong empirical performance on structured/tabular clinical data and its compatibility with SHAP-based interpretation. Within the internal NHANES 2017–2018 benchmarking cohort described in Section 2.1(ii), a class-weighted XGBoost model, a LightGBM classifier [12], a CatBoost classifier [13], and a soft-voting ensemble of these models were additionally trained on the 80% training partition and evaluated on the 20% held-out test partition ($n = 1,136$), to verify that the selected algorithm was not idiosyncratically favorable and to characterize discrimination–calibration trade-offs across gradient-boosting implementations. These comparisons inform algorithm selection but are internal to a single NHANES cycle and are reported separately from the forward-temporal-validation results (Section 2.1(i)).

2.8 Hyperparameter optimization for the temporally validated Stage-1 model

Hyperparameters for the Stage-1 model used in forward temporal validation were optimized using randomized search with five-fold stratified cross-validation on the NHANES 2015–2016 development data only, using

ROC-AUC as the optimization objective. The final model used 200 estimators, maximum tree depth of 3, learning rate 0.03, minimum child weight 10, subsample fraction 0.80, column-subsampling fraction 0.90, gamma 0.30, L1 regularization (reg_alpha) 0, and L2 regularization (reg_lambda) 1 (Supplementary Table S1). The relatively shallow trees and conservative learning rate were selected by the randomized search itself and are consistent with a bias toward regularized, less overfit-prone configurations given the moderate sample size and class imbalance of the development cohort.

2.9 Operating-threshold selection

Because the outcome was imbalanced (14.43% prevalence in the development cohort) and because using the independent validation cohort to select an operating threshold would bias its subsequent evaluation, the classification threshold used for forward temporal validation was determined using out-of-fold predictions generated during cross-validation on the NHANES 2015–2016 development data only. The F1-optimal threshold identified in this manner was 0.21954 and was fixed before being applied, without modification, to the NHANES 2017–2018 temporal validation cohort. Algorithm 2 summarizes the full development-to-temporal-validation procedure, including this leakage-aware threshold selection step.

2.10 Model evaluation

Discrimination was assessed using the area under the receiver-operating-characteristic curve (ROC-AUC) and the area under the precision–recall curve (PR-AUC), the latter being particularly informative under class imbalance. Threshold-dependent metrics included sensitivity, specificity, precision, and F1 score, computed at the pre-specified development-derived threshold. Overall probabilistic accuracy was assessed with the Brier score. Calibration was assessed by regressing the observed outcome on the linear predictor (logit of predicted probability) and reporting the resulting calibration intercept and calibration slope [7]: at a fixed slope, a positive intercept indicates that observed outcome frequency exceeds predicted risk (systematic underestimation of risk), a negative intercept indicates systematic overestimation, and a slope of exactly 1 indicates an appropriately scaled risk gradient without implying that the model’s discrimination or overall calibration is otherwise ideal. Bootstrap resampling with 2,000 iterations was used to estimate 95% confidence intervals for ROC-AUC, PR-AUC, sensitivity, specificity, and F1 score in the temporal validation cohort. Clinical utility was assessed using decision curve analysis, which quantifies net benefit across a range of threshold probabilities and compares the model against treat-all and treat-none default strategies [9].

Algorithm 2 Leakage-aware development and forward temporal validation of the Stage-1 model

Require: X_{train} (NHANES 2015–2016), y_{train} , X_{valid} (NHANES 2017–2018), y_{valid}

Ensure: Trained model f , operating threshold t^* , temporal validation metrics

- 1: Split X_{train} into $K = 5$ stratified folds
 - 2: $\theta^* \leftarrow \text{RandomizedSearchCV}(\text{XGBoost}, X_{\text{train}}, y_{\text{train}}, \text{folds} = K, \text{scoring} = \text{ROC-AUC})$
 - 3: $\text{oof_probs} \leftarrow \mathbf{0}^{|y_{\text{train}}|}$
 - 4: **for** $k = 1$ to K **do**
 - 5: Fit $f_k \leftarrow \text{XGBoost}(\theta^*)$ on folds $\neq k$
 - 6: $\text{oof_probs}[\text{fold } k] \leftarrow f_k.\text{predict_proba}(\text{fold } k)$
 - 7: **end for**
 - 8: $t^* \leftarrow \text{argmax}_t F_1(y_{\text{train}}, \text{oof_probs} \geq t)$ {threshold fixed here, using D_{train} ONLY}
 - 9: $f \leftarrow \text{XGBoost}(\theta^*).\text{fit}(X_{\text{train}}, y_{\text{train}})$ {final model refit on all of D_{train} }
 - 10: $p_{\text{valid}} \leftarrow f.\text{predict_proba}(X_{\text{valid}})$ {no refitting on X_{valid} }
 - 11: $\text{metrics} \leftarrow \text{Evaluate}(y_{\text{valid}}, p_{\text{valid}}, t^*)$ {ROC-AUC, PR-AUC, Sens., Spec., F1, Brier}
 - 12: $(\text{intercept}, \text{slope}) \leftarrow \text{CalibrationRegression}(y_{\text{valid}}, p_{\text{valid}})$
 - 13: $\text{NB} \leftarrow \text{DecisionCurveAnalysis}(y_{\text{valid}}, p_{\text{valid}})$
 - 14: $\text{CI} \leftarrow \text{BootstrapCI}(y_{\text{valid}}, p_{\text{valid}}, t^*, n_{\text{resamples}} = 2000)$
 - 15: **return** $f, t^*, \text{metrics}, (\text{intercept}, \text{slope}), \text{NB}, \text{CI}$
-

2.11 Explainable artificial intelligence

SHAP (SHapley Additive exPlanations) values [10] were computed for the final, temporally validated Stage-1 XGBoost model using a tree-based SHAP explainer to quantify the contribution of each predictor to individual model predictions. Global feature importance was summarized using the mean absolute SHAP value across the evaluation sample, and predictor-level contributions were additionally aggregated into demographic, anthropometric, blood-pressure, and lifestyle domains to characterize the relative contribution of each domain to model output. Because engineered interaction terms (e.g., age $\times$ BMI, age $\times$ waist circumference) overlap conceptually with their constituent variables, individual SHAP magnitudes were interpreted descriptively as measures of model behavior rather than as independent causal effect estimates.

2.12 Subgroup analysis

Subgroup performance by age (18–39, 40–59, 60+ years), BMI category (normal, overweight, obese), and sex was evaluated using the internal NHANES 2017–2018 benchmarking cohort’s 20% held-out test partition ($n = 1,136$; Section 2.1(ii)), at the corresponding internally selected operating threshold, rather than on the full forward-temporal-validation cohort ($n = 5,678$). This distinction is stated explicitly because it explains why subgroup sample sizes in Table 4 ($396 + 319 + 421 = 1,136$ for age; $564 + 572 = 1,136$ for sex; $285 + 332 + 421$

= 1,038 for BMI, reflecting missing BMI values in a small number of records) are substantially smaller than the full temporal validation cohort. ROC-AUC, PR-AUC, sensitivity, specificity, and F1 score were calculated within each subgroup; because some subgroups, particularly the youngest age band (8 events), contained few outcome events, subgroup estimates were interpreted descriptively rather than as confirmatory comparisons. Repeating this subgroup analysis on the full forward-temporal-validation cohort at its own development-derived threshold was not performed in the present pipeline and is identified as a priority for future revision of this work.

2.13 Secondary internal Stage-2 analysis (clinical refinement)

A secondary, exploratory Stage-2 model incorporated HbA1c (LBXGH) and fasting plasma glucose (LBXGLU) in addition to the full Stage-1 predictor set. This analysis was conducted entirely within the internal NHANES 2017–2018 benchmarking cohort described in Section 2.1(ii) ($n = 5,678$; 80/20 stratified split, 4,543 training / 1,136 test), using an independently specified XGBoost configuration (500 estimators, maximum depth 5, learning rate 0.03, subsample 0.80, column-subsample 0.80, with scale_pos_weight set to the training-set class-imbalance ratio) evaluated at the conventional default probability threshold of 0.50 rather than a development-derived optimal threshold. Missing HbA1c/fasting-glucose values were handled by the same median-imputation procedure as Stage-1 predictors, fitted on the 80% training partition. Calibration, decision curve analysis, and bootstrap confidence intervals were not computed for this analysis, and it was not evaluated under the forward-temporal-validation design used for Stage 1. Accordingly, Stage-2 results are reported as a secondary, hypothesis-generating clinical-refinement analysis illustrating the incremental discriminative value of glycemic biomarkers once measured, rather than as a temporally validated extension of the Stage-1 model, and its performance is not compared numerically head-to-head with the Stage-1 temporal-validation estimate.

2.14 Statistical and methodological framework

Throughout, the study emphasized strict separation of development and evaluation data within the forward-temporal-validation design (predictor imputation fitted only on development data; operating threshold selected only from development out-of-fold predictions), joint evaluation of discrimination and calibration, assessment of clinical usefulness via decision curve analysis, and explicit labeling of analyses that used a single-cycle internal split rather than independent-cycle temporal validation, in line with methodological guidance for machine-learning-based clinical prediction models [2, 3, 5, 6, 8].

2.15 Software

Analyses were implemented in Python using scikit-learn for preprocessing, cross-validation, and calibration metrics; XGBoost, LightGBM, and CatBoost for gradient-boosting model implementations; and the SHAP library for model interpretation. Modeling code and outputs were organized across reproducible analysis notebooks covering data preparation, the internal algorithm-benchmarking and Stage-2 analyses, the forward-temporal-validation pipeline, and explainability.

3 Results

3.1 Cohort characteristics

The development cohort (NHANES 2015–2016) comprised 5,856 eligible participants, of whom 845 (14.43%) reported a T2DM diagnosis. The temporal validation cohort (NHANES 2017–2018) comprised 5,678 eligible participants, of whom 878 (15.46%) reported a T2DM diagnosis (Table 1). The modestly higher prevalence in the later cycle is consistent with previously documented year-on-year increases in national diabetes prevalence estimates [16, 17] and provided a realistic test of transportability under natural distributional drift between survey cycles.

Table 1: Cohort characteristics and outcome prevalence.

Cohort	N	T2DM events	Non-T2DM	Prevalence (%)
NHANES 2015–2016 (development)	5,856	845	5,011	14.43
NHANES 2017–2018 (temporal validation)	5,678	878	4,800	15.46

3.2 Model development and comparison (internal NHANES 2017–2018 benchmarking cohort)

Within the internal NHANES 2017–2018 benchmarking cohort (Methods, Section 2.1(ii)), the default-threshold XGBoost model achieved ROC-AUC 0.7795 but very low sensitivity (14.77%) because of class imbalance. Class-weighted XGBoost, LightGBM, and CatBoost achieved broadly similar ROC-AUC (0.7756–0.7900) with different sensitivity–specificity trade-offs, and a soft-voting ensemble of these models did not outperform the best individual learner. Following feature engineering and randomized hyperparameter optimization, an optimized XGBoost model achieved ROC-AUC 0.7971 and PR-AUC 0.3933 on the 20% held-out internal test partition ($n = 1,136$), with sensitivity 71.59% and specificity 70.42% at its own selected operating point (Table 2). A reverse-direction robustness check, in which this same 2017–2018-trained model was applied at its untuned default threshold (0.50) to the complete, independent NHANES 2015–2016 cohort ($n = 5,856$), yielded ROC-AUC 0.8012 but low sensitivity (14.44%) and high specificity (97.15%), illustrating again that discrimination (ROC-AUC) was reasonably preserved across cycles in this direction as well, while threshold-dependent metrics were highly sensitive to the untuned default threshold rather than to a loss of discrimination per se. This internal-cohort model, its hyperparameters, and its default-threshold evaluation are distinct from the separately optimized, development-only-thresholded Stage-1 model used for the primary forward-temporal-validation analysis (Section 3.3).

Table 2: Model comparison across the three evaluation designs used in this study. *Internal benchmarking cohort: model trained on an 80% partition of NHANES 2017–2018 and evaluated on the corresponding 20% held-out partition ($n = 1,136$) from the same cycle. †Forward temporal validation, the primary analysis of this study: model trained on the complete NHANES 2015–2016 cohort ($n = 5,856$) and evaluated without retraining on the complete, independent NHANES 2017–2018 cohort ($n = 5,678$), at a threshold selected from development data only. ‡Reverse-direction robustness check: the internal-benchmarking model trained on NHANES 2017–2018 (80% partition) evaluated, at the untuned default threshold of 0.50, on the complete NHANES 2015–2016 cohort ($n = 5,856$). Rows marked with the same symbol were evaluated under the same design; results across different symbols should not be compared as if from the same evaluation.

Model / experiment	ROC-AUC	PR-AUC	Sens.	Spec.	F1	Brier
Original XGBoost, default threshold*	0.7795	0.3653	0.1477	0.9646	0.2203	0.1174
Class-weighted XGBoost*	0.7756	0.3464	0.5341	0.8052	0.4114	0.1492
LightGBM*	0.7767	0.3771	0.4318	0.8427	0.3772	0.1404
CatBoost*	0.7900	0.3696	0.5909	0.7667	0.4127	0.1535
Soft-voting ensemble*	0.7870	0.3754	0.3466	0.8948	0.3609	0.1250
Optimized XGBoost (internal test)*	0.7971	0.3933	0.7159	0.7042	0.4300	0.1112
Forward temporal validation†	0.8083	0.4119	0.6446	0.7833	0.4557	0.1088
Secondary Stage-2 (+HbA1c, FPG)*	0.9186	0.7218	0.6648	0.9260	0.6429	0.0779
Reverse-direction check‡	0.8012	0.3642	0.1444	0.9715	0.2198	0.1162

3.3 Forward temporal validation

When the optimized Stage-1 model, together with its development-derived operating threshold (0.21954), was applied without retraining to the independent NHANES 2017–2018 cohort, ROC-AUC was 0.8083 (95% CI 0.7947–0.8212) and PR-AUC was 0.4119 (95% CI 0.3805–0.4453) (Table 3; Figure 1, Figure 2). Sensitivity was 64.46% (95% CI 61.26–67.48%), specificity was 78.33% (95% CI 77.13–79.45%), and F1 score was 0.4557 (95% CI 0.4311–0.4795). Overall accuracy was 76.19% and precision was 35.24%. The corresponding confusion matrix (Table 4) showed 566 true positives, 312 false negatives, 1,040 false positives, and 3,760 true negatives. Notably, ROC-AUC in the temporal validation cohort (0.8083) was numerically similar to, rather than lower than, the internal optimized-model estimate (0.7971), indicating that the ranking of participants by predicted risk was preserved across survey cycles despite the model receiving no information from the later cycle during training or threshold selection.

Table 3: Forward temporal validation performance (NHANES 2015–2016 → 2017–2018) with 95% bootstrap confidence intervals (2,000 resamples).

Metric	Estimate	95% CI
ROC-AUC	0.8083	0.7947–0.8212
PR-AUC	0.4119	0.3805–0.4453
Sensitivity	0.6446	0.6126–0.6748
Specificity	0.7833	0.7713–0.7945
F1 score	0.4557	0.4311–0.4795
Brier score	0.1088	
Calibration intercept	0.1783	
Calibration slope	1.1103	
Accuracy	0.7619	
Precision (PPV)	0.3524	
Operating threshold	0.21954	

Table 4: Confusion matrices for forward temporal validation and for a reverse-direction robustness check evaluated at the conventional default threshold of 0.50.

Dataset	TN	FP	FN	TP
Forward temporal validation, $t = 0.21954$	3,760	1,040	312	566
Reverse-direction check (trained 2017–2018 → evaluated on full 2015–2016), default $t = 0.50$	4,868	143	723	122

3.4 Calibration and clinical utility

Temporal calibration analysis, based on regressing observed outcome on the linear predictor of the Stage-1 model, yielded a calibration intercept of 0.1783 and a calibration slope of 1.1103, with a Brier score of 0.1088 (Figure 3). Under this parameterization, a positive intercept indicates that, at a given predicted linear-predictor value, the observed rate of T2DM in the temporal validation cohort was somewhat higher than the rate predicted by the model; that is, the model tended to modestly underestimate absolute risk in NHANES 2017–2018 relative to the risk levels learned from NHANES 2015–2016. The calibration slope of 1.1103 was close to, but not identical to, the ideal value of 1, indicating a broadly comparable but not perfectly matched spread of predicted risk relative to observed risk. These findings are reported descriptively rather than characterized as ‘good’ or

‘appropriate’ calibration in isolation, consistent with the recommendation that calibration be interpreted quantitatively rather than qualitatively [7, 35, 36]; in an applied setting, this pattern of a near-unit slope with a positive intercept would typically be addressed with a simple recalibration (e.g., re-estimating the intercept, or Platt scaling) before predicted probabilities are presented as absolute risk. Decision curve analysis indicated greater net benefit for the model than either the treat-all or treat-none strategy across threshold probabilities from approximately 0.010 to 0.705 (Figure 4). At the selected operating threshold of 0.21954, model net benefit was 0.0479, compared with -0.0844 for the treat-all strategy and 0.0000 for treat-none, indicating that using the model to guide a hypothetical intervention decision would have been preferable to intervening in either everyone or no one across this range.

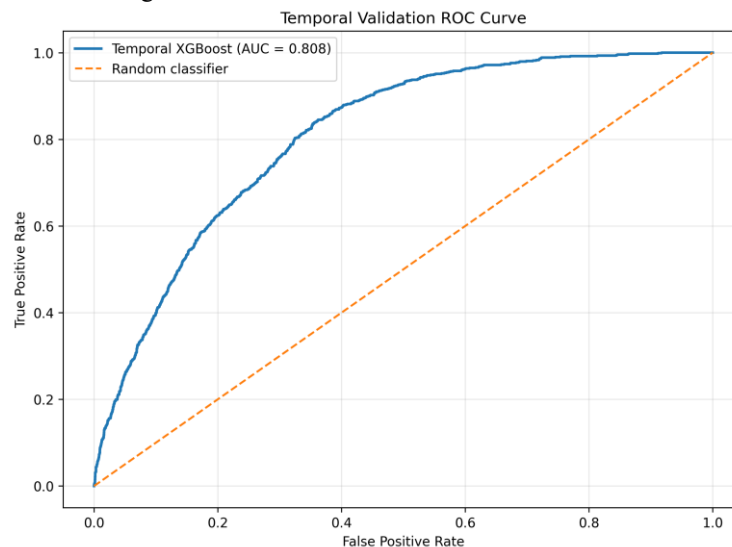


Figure 1: Receiver-operating-characteristic curve for forward temporal validation (NHANES 2015–2016 → 2017–2018).

3.5 Subgroup performance

Subgroup analysis was conducted on the internal NHANES 2017–2018 benchmarking cohort’s 20% held-out test partition ($n = 1,136$; Methods, Section 2.12), not on the full forward-temporal-validation cohort ($n = 5,678$); this is why subgroup counts in Table 5 sum to a substantially smaller total than the primary temporal validation sample, and it should be kept in mind when comparing Table 5 with Tables 2 and 3. In sex-stratified analysis, ROC-AUC was 0.7983 among males and 0.7943 among females, indicating similar discrimination between sexes. In BMI-stratified analysis, ROC-AUC was 0.8464 in the normal-BMI group, 0.7312 in the overweight group, and 0.8034 in the obese group. Age-stratified analysis showed ROC-AUC of 0.7610 for ages 18–39, 0.7213 for ages 40–59, and 0.6301 for ages 60 years or older; the youngest group contained only 8 positive events (2.02% prevalence), so its performance estimates, including a sensitivity of 0.000 at the fixed threshold, should be interpreted with caution rather than as evidence that the model fails in younger adults per se.

3.6 Explainability

SHAP analysis identified age as the highest-ranked individual contributor to model output, followed by the age $\times$ BMI interaction, the age $\times$ waist-circumference interaction, waist circumference itself, race/ethnicity, BMI-to-waist ratio, the BMI $\times$ systolic-blood-pressure interaction, mean diastolic blood pressure, pulse pressure, and income-to-poverty ratio (Table 6; Figure 5, Figure 6). When mean absolute SHAP values were aggregated by predictor domain, anthropometric variables accounted for the largest share of model output (45.17%), followed by demographic variables (33.27%), blood-pressure variables (14.02%), lifestyle variables (6.92%), and other/missingness indicators (0.61%) (Table 7). Supplementary dependence plots (Figures S1–S5) show that predicted risk increased approximately monotonically with age, BMI, and waist circumference, with the age $\times$ BMI and age $\times$ waist interactions indicating that the marginal contribution of adiposity to predicted risk was more pronounced at older ages, a pattern broadly consistent with the epidemiological association between central adiposity, aging, and insulin resistance [15]. Because engineered interaction variables overlap conceptually with their component variables, individual SHAP magnitudes should not be interpreted as independent or causal effect estimates.

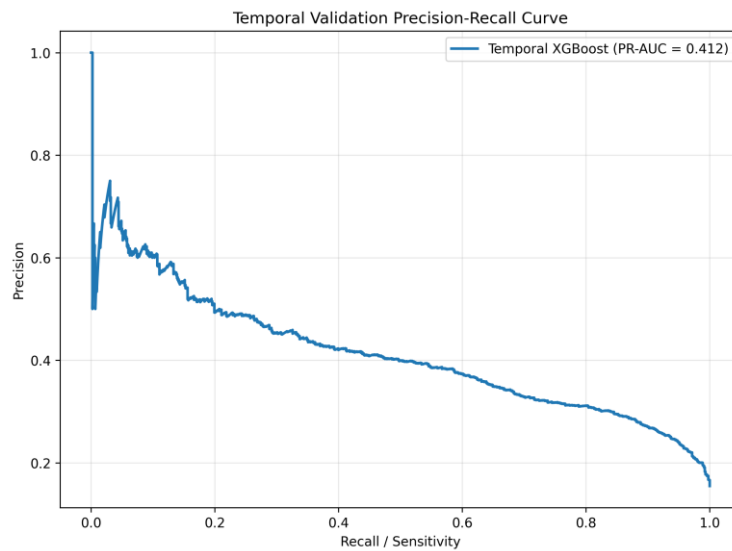


Figure 2: Precision–recall curve for forward temporal validation.

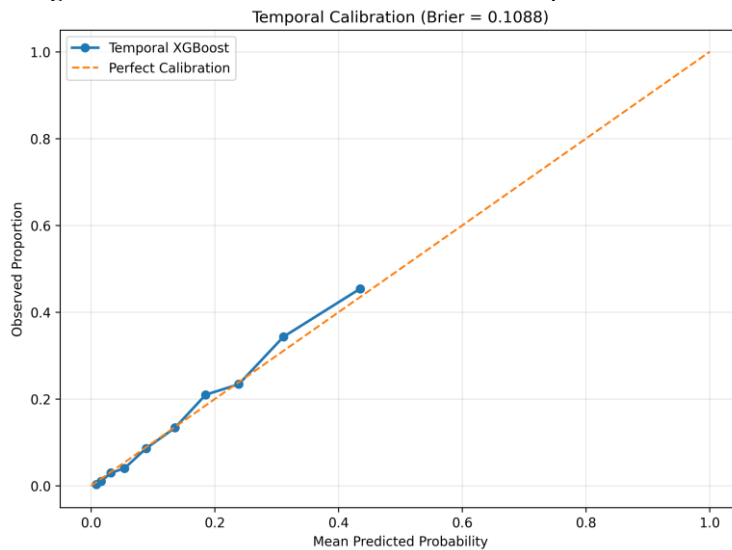


Figure 3: Calibration curve for forward temporal validation.

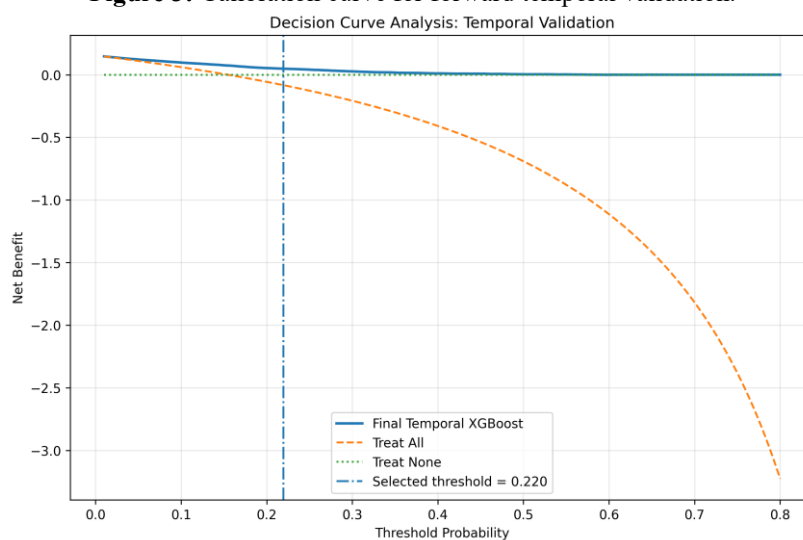


Figure 4: Decision curve analysis comparing the Stage-1 model with treat-all and treat-none strategies.

Table 5: Subgroup performance by age, body-mass-index category, and sex, evaluated on the internal NHANES 2017–2018 benchmarking cohort’s held-out test partition ($n = 1,136$), not the full forward-temporal-validation cohort (see Methods, Section 2.12).

Subgroup	N	Events	Prev.(%)	ROC-AUC	PR-AUC	Sens.	Spec.	F1
Age 18–39	396	8	2.02	0.7610	0.2375	0.000	1.000	0.000

Subgroup	N	Events	Prev.(%)	ROC-AUC	PR-AUC	Sens.	Spec.	F1
Age 40–59	319	48	15.05	0.7213	0.3584	0.5208	0.7712	0.3704
Age 60+	421	120	28.50	0.6301	0.4263	0.8417	0.2625	0.4560
BMI: Normal	285	33	11.58	0.8464	0.4176	0.5152	0.8968	0.4474
BMI: Overweight	332	47	14.16	0.7312	0.2487	0.6383	0.6561	0.3429
BMI: Obese	421	85	20.19	0.8034	0.5091	0.8235	0.5893	0.4778
Sex: Male	564	94	16.67	0.7983	0.4006	0.7447	0.6894	0.4516
Sex: Female	572	82	14.34	0.7943	0.4157	0.6829	0.7184	0.4058

3.7 Secondary internal Stage-2 analysis (clinical refinement)

In the secondary, exploratory Stage-2 analysis conducted entirely within the internal NHANES 2017–2018 benchmarking cohort (80/20 split; Methods, Section 2.13), addition of HbA1c and fasting glucose to the Stage-1 predictor set increased discrimination relative to the internal-cohort Stage-1 comparator models in Table 2, with ROC-AUC of 0.9186 and PR-AUC of 0.7218 at the default 0.50 threshold; sensitivity was 66.48%, specificity was 92.60%, F1 was 0.6429, and Brier score was 0.0779. Because this analysis used a different dataset, split, model configuration, and (untuned, default) threshold than the forward-temporal-validation Stage-1 result reported in Section 3.3, its ROC-AUC of 0.9186 should not be read as a direct improvement over the temporally validated Stage-1 estimate of 0.8083; the two numbers characterize different evaluation designs and are not comparable head-to-head. Interpreted on its own terms, the Stage-2 result is consistent with direct glycemic measurements providing substantial incremental discriminative information once non-glycemic risk stratification has been performed, and motivates a future Stage-2 analysis conducted under the same forward-temporal-validation design as Stage 1 (development-only threshold selection, calibration, decision curve analysis, and evaluation on the independent NHANES 2017–2018 cohort), which was not performed in the present study.

Table 6: Top 15 SHAP features and mean absolute SHAP values, with domain assignment.

Rank	Feature	Domain	Mean SHAP
1	Age (RIDAGEYR)	Demographic	0.3918
2	Age × BMI	Anthropometric	0.3106
3	Age × waist circumference	Anthropometric	0.2646
4	Waist circumference (BMXWAIST)	Anthropometric	0.2294
5	Race/ethnicity (RIDRETH3)	Demographic	0.1528
6	BMI-to-waist ratio	Anthropometric	0.1225
7	BMI × systolic BP	Blood pressure	0.1165
8	Mean diastolic BP	Blood pressure	0.0834
9	Pulse pressure	Blood pressure	0.0721
10	Income-to-poverty ratio (INDFMPIR)	Demographic	0.0644
11	Body weight (BMXWT)	Anthropometric	0.0521
12	Body-mass index (BMXBMI)	Anthropometric	0.0494
13	Education level (DMDEDUC2)	Demographic	0.0430
14	BMI × sex	Anthropometric	0.0413
15	Age squared	Demographic	0.0374

Table 7: Aggregated SHAP contribution by predictor domain.

Domain	Mean SHAP	Contribution (%)
Anthropometric	1.0350	45.17
Demographic	0.7624	33.27
Blood pressure	0.3212	14.02
Lifestyle	0.1587	6.92
Other / missingness indicators	0.0141	0.61

4 DISCUSSION

The principal finding of this study is that an explainable multi-domain XGBoost framework maintained, rather than lost, discrimination when evaluated on a later, independently collected NHANES survey cycle. The Stage-1 model achieved ROC-AUC 0.8083 in forward temporal validation, compared with an internal optimized-model ROC-AUC of 0.7971 in the earlier development cycle. Because the temporal validation cohort was not used for model development, hyperparameter optimization, or threshold selection, this estimate is more relevant to assessing transportability than the internal test result, and its similarity to (rather than substantial degradation from) the development-cycle estimate suggests that the learned ranking of participants by non-glycemic risk factors was relatively stable across the two survey cycles examined here. This pattern is broadly consistent with the transportability reported by Khattab et al. for their Stage-1 model across the All of Us Research Program and UK Biobank (AUROC approximately 0.81–0.82), despite substantial differences in population, predictor set, and outcome definition [1], and lends some support to the more general observation that adequately regularized

gradient-boosting models built on demographic and anthropometric predictors can generalize across related but non-identical populations when development and evaluation data are kept strictly separate [6, 8].

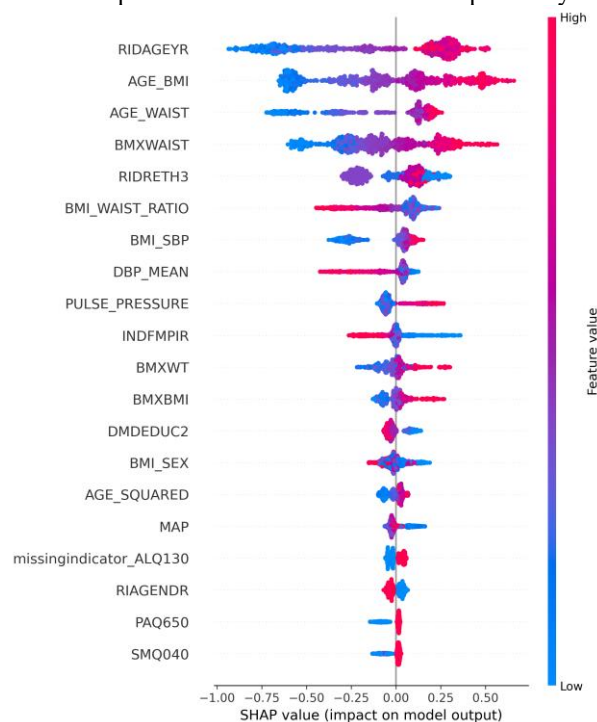


Figure 5: SHAP beeswarm plot showing the distribution and direction of feature contributions for the Stage-1 model.

The operating threshold illustrates the importance of evaluating prediction models beyond the conventional default probability threshold of 0.50. Applying a default threshold in this class-imbalanced setting produced very low sensitivity (Table 4), whereas the development-derived threshold of 0.21954 produced a markedly more clinically usable balance of sensitivity (64.46%) and specificity (78.33%) in the later validation cohort. This finding reinforces the broader methodological point that threshold selection should be explicit, pre-specified from development data, and linked to the intended screening application, rather than left at a default value inherited from generic classification settings.

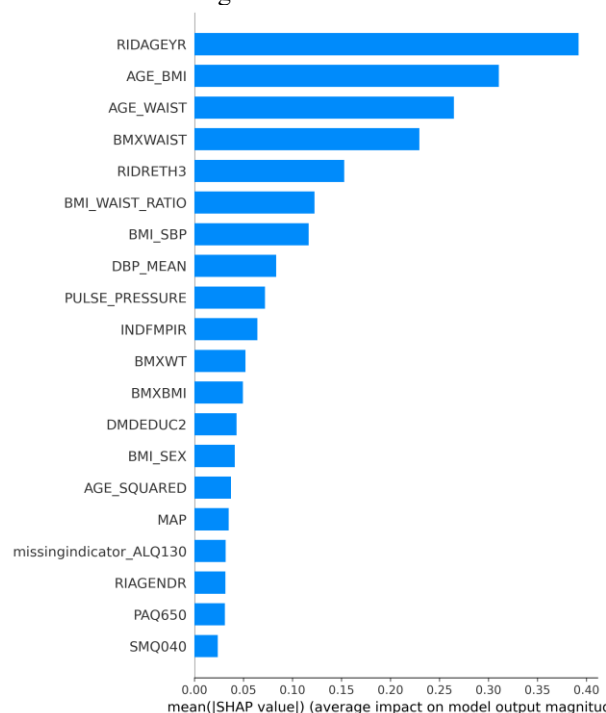


Figure 6: Global SHAP feature importance (mean absolute SHAP value) for the Stage-1 model.

Calibration and clinical utility provide complementary evidence to discrimination. The temporal calibration slope of 1.1103 was close to, though not exactly equal to, the ideal value of 1, while the positive calibration intercept of 0.1783 indicates that the model tended to modestly underestimate absolute risk in the NHANES 2017–2018 validation cohort relative to the risk levels learned from NHANES 2015–2016. In a deployment

setting, this pattern would warrant simple recalibration (for example, re-estimating the intercept, or Platt scaling) before probabilities are presented as absolute risk estimates, and calibration should continue to be reported and monitored quantitatively rather than summarized only as ‘good’ or ‘poor’ [7]. Decision curve analysis demonstrated greater net benefit than both the treat-all and treat-none strategies across a broad range of threshold probabilities (approximately 0.01–0.70), supporting the interpretation of the model as a potential risk-stratification tool rather than relying on ROC-AUC in isolation, consistent with methodological guidance emphasizing that discrimination alone is an incomplete measure of clinical value [9].

SHAP analysis indicated that age, waist circumference, and their interaction with BMI dominated the model’s learned representation of diabetes status, together accounting for the largest share of aggregated domain-level SHAP contribution (anthropometric and demographic domains combined, 78.4%). This pattern is consistent with a substantial epidemiological literature linking central adiposity, and its interaction with aging, to insulin resistance and incident T2DM, and with evidence that waist circumference can carry information about diabetes risk beyond BMI alone [15]. The prominent contribution of race/ethnicity is consistent with well-documented disparities in diabetes prevalence across racial and ethnic groups in NHANES and other U.S. national survey data [16, 17], and underscores the importance of explicitly reporting subgroup performance (Table 5) rather than presenting only aggregate discrimination, in line with applicability considerations emphasized by prediction-model risk-of-bias frameworks [2, 3]. As emphasized throughout, SHAP contributions describe the behavior of the fitted model and should not be interpreted as estimates of causal effects, particularly for the engineered interaction terms that conceptually overlap with their component variables.

Subgroup analysis (conducted on the internal NHANES 2017–2018 hold-out partition, $n = 1,136$; Methods, Section 2.12) revealed heterogeneous discrimination across age strata, with ROC-AUC declining from 0.7610 in the youngest group to 0.6301 in participants aged 60 years or older. Discrimination was lower in the older age stratum; several explanations are plausible, including a compression of achievable risk separation as non-glycemic risk factors become near-universal at older ages, unmeasured confounding, or sample-specific variation given the relatively small subgroup sizes, but the mechanism underlying this heterogeneity cannot be established from the present cross-sectional analysis and should be investigated directly, ideally by repeating the subgroup analysis on the full forward-temporal-validation cohort. The very small number of events in the youngest stratum ($n = 8$) further limits the precision of subgroup estimates in that group specifically and should not be over-interpreted.

In the secondary internal Stage-2 analysis, adding HbA1c and fasting glucose to the Stage-1 predictor set increased ROC-AUC to 0.9186 within the internal NHANES 2017–2018 benchmarking cohort. Because this analysis used a different dataset/split, model configuration, and untuned default threshold than the forward-temporal-validation Stage-1 estimate of 0.8083, it is presented as an exploratory clinical-refinement analysis rather than direct evidence that Stage 2 outperforms the temporally validated Stage-1 model. On its own terms, the result is consistent with the conceptual value of a two-stage architecture in which non-glycemic risk assessment, potentially usable in settings without immediate access to laboratory glycemic testing, is followed by clinical refinement once glycemic biomarkers become available, mirroring the rationale underlying the two-stage design of Khattab et al. [1]. A confirmatory Stage-2 analysis conducted under the same forward-temporal-validation design as Stage 1, including development-only threshold selection, calibration, and decision curve analysis on the independent NHANES 2017–2018 cohort, was not performed here and is identified as a specific priority for future work.

4.1 Strengths

This study has several methodological strengths. First, the primary Stage-1 model’s development, hyperparameter optimization, and threshold selection were performed exclusively on the earlier NHANES cycle, with forward temporal validation on an independently collected later cycle, addressing a common criticism of machine-learning prediction studies that rely solely on random train/test splits from a single dataset [4, 6, 8]; a reverse-direction robustness check (trained on 2017–2018, evaluated on the full 2015–2016 cohort) provided a complementary, if less rigorously thresholded, cross-cycle comparison. Second, evaluation of the Stage-1 model combined discrimination, calibration, bootstrap confidence intervals, decision curve analysis, subgroup performance, and SHAP-based explainability, in line with contemporary reporting and risk-of-bias guidance for prediction models using machine learning methods [2, 3, 5]. Third, the framework was built entirely from publicly accessible NHANES data, supporting reproducibility and independent re-analysis, and the underlying feature-engineering and forward-temporal-validation procedures are specified in reproducible algorithmic form (Algorithms 1–2). Fourth, the explicit Stage-1/Stage-2 separation isolates the contribution of non-glycemic predictors from that of direct glycemic biomarkers, clarifying what each stage of the framework can and cannot claim to demonstrate, even though the Stage-2 and subgroup analyses reported here were conducted under a less stringent internal-cohort design than Stage-1’s forward temporal validation.

5 Limitations

- NHANES is a cross-sectional survey; therefore, the present analysis characterizes diabetes-status assessment and its transportability across survey cycles, and does not establish prospective prediction of incident T2DM.
- The diabetes outcome is based on self-reported diabetes-status information (NHANES variable DIQ010) rather than longitudinally confirmed incident disease or laboratory-adjudicated diagnosis, and is therefore subject to recall and reporting bias.

- Forward temporal validation across two NHANES cycles provides evidence of transportability across survey periods collected using a consistent national sampling design, but is not equivalent to prospective clinical validation in a health-system or point-of-care setting.
- The study was performed using United States national survey data; performance, calibration, and the relative importance of specific predictors may differ in other populations, health systems, or countries.
- Subgroup estimates can be unstable when event counts are small, particularly in the youngest age stratum (8 events), and should be interpreted with corresponding caution.
- The subgroup and secondary Stage-2 analyses reported here were conducted on an 80/20 internal split of NHANES 2017–2018 alone, rather than on the full forward-temporal-validation cohort or under the same development-only threshold selection, calibration, and decision-curve-analysis design used for the primary Stage-1 result; consequently, subgroup and Stage-2 estimates in this study characterize an internal cross-sectional cohort and should not be interpreted as having received the same temporal-validation rigor as the headline Stage-1 result.
- The Stage-2 model incorporates direct glycemic measurements (HbA1c, fasting glucose) and therefore represents an exploratory clinical-refinement analysis rather than a non-glycemic early-risk model; because it was also evaluated under a different design than Stage 1 (Section 2.13), its performance should not be compared directly with the temporally validated Stage-1 result as evidence of a superior early-risk model.
- SHAP values describe associations and contributions learned by the fitted model and should not be interpreted as estimates of causal effects, particularly given the deliberately engineered, partially overlapping interaction terms used in this framework.
- Further external validation using independent, longitudinal clinical cohorts with adjudicated diabetes outcomes is required before any consideration of clinical deployment, consistent with applicability requirements articulated in prediction-model risk-of-bias frameworks [2, 3].

6 Conclusion

This study developed and evaluated an explainable two-stage multi-domain machine-learning framework for personalized T2DM status/risk assessment using publicly available NHANES data. The primary Stage-1 XGBoost model achieved ROC-AUC of 0.8083 (95% CI 0.7947–0.8212) in forward temporal validation from NHANES 2015–2016 to NHANES 2017–2018, with sensitivity of 64.46% and specificity of 78.33% at a pre-specified, development-derived threshold, and this discrimination was broadly corroborated by a reverse-direction check (ROC-AUC 0.8012) evaluating a separately trained model in the opposite temporal direction. Calibration analysis (indicating a modest, correctable underestimation of absolute risk), decision curve analysis, and SHAP-based interpretation provided complementary evidence beyond discrimination alone for this temporally validated Stage-1 model. A secondary, exploratory Stage-2 analysis conducted on an internal split of NHANES 2017–2018 (ROC-AUC 0.9186) and a subgroup analysis conducted on the same internal cohort suggest, respectively, that glycemic biomarkers could substantially refine risk estimates and that discrimination may vary by age; because both analyses used a different, less stringent evaluation design than the primary Stage-1 result, both should be treated as hypothesis-generating pending confirmation under the same forward-temporal-validation design. Future work should evaluate this framework in longitudinal, externally validated clinical cohorts with adjudicated diabetes outcomes, apply simple recalibration to address the observed calibration offset, repeat the subgroup and Stage-2 analyses on the full temporally validated cohort, and further investigate the mechanism underlying performance heterogeneity across age subgroups before any consideration of prospective clinical implementation.

DECLARATIONS

Ethics approval and consent to participate. This study used publicly available, de-identified NHANES data released by the U.S. Centers for Disease Control and Prevention/National Center for Health Statistics [18]; NHANES data collection was approved by the NCHS Research Ethics Review Board, and all original participants provided informed consent at the time of data collection. As a secondary analysis of publicly available de-identified data, this study did not require separate institutional ethics approval.

Consent for publication. Not applicable.

Availability of data and materials. The NHANES 2015–2016 and 2017–2018 datasets analyzed in this study are publicly available from the U.S. Centers for Disease Control and Prevention at <https://www.cdc.gov/nchs/nhanes/>. Derived analysis datasets, trained model artifacts, and code supporting the findings of this study are available from the corresponding author upon reasonable request. Code available on request.

Competing interests. The authors declare that they have no competing interests.

Funding. Not received.

Authors' contributions. All authors read and approved the final manuscript. All work done by corresponding author.

ACKNOWLEDGEMENTS. Not applicable

REFERENCES

- [1] Khattab A, et al. Development and validation of a two-stage machine learning model for personalised type 2 diabetes screening in the All of Us Research Program and UK Biobank. *BMJ Open*. 2026;16:e103171.
- [2] Moons KGM, 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.
- [3] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58.
- [4] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281.
- [5] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13:1.
- [6] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
- [7] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230.
- [8] Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247.
- [9] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
- [10] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
- [11] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016. p. 785–794.
- [12] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146–3154.
- [13] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst*. 2018;31:6638–6648.
- [14] Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*. 2003;26(3):725–731.
- [15] Wang Y, Rimm EB, Stampfer MJ, Willett WC, Hu FB. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. *Am J Clin Nutr*. 2005;81(3):555–563.
- [16] Cheng YJ, Kanaya AM, Araneta MRG, Saydah SH, Kahn HS, Gregg EW, et al. Prevalence of diabetes by race and ethnicity in the United States, 2011–2016. *JAMA*. 2019;322(24):2389–2398.
- [17] International Diabetes Federation. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. *Lancet Diabetes Endocrinol*. 2025 (published online ahead of print).
- [18] Centers for Disease Control and Prevention, National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES): survey methods and analytic guidelines, 2015–2018 cycles. Hyattsville, MD: NCHS; 2018.
- [19] Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, 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.
- [20] Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. *JAMA*. 2015;314(10):1021–1029.
- [21] Bang H, Edwards AM, Bombardier AS, Ballantyne CM, Brillion D, Callahan MA, et al. Development and validation of a patient self-assessment score for diabetes risk. *Ann Intern Med*. 2009;151(11):775–783.
- [22] Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ*. 2011;343:d7163.
- [23] Ross R, Neeland IJ, Yamashita S, Shai I, Seidell J, Magni P, et al. Waist circumference as a vital sign in clinical practice: a Consensus Statement from the IAS and ICCR Working Group on Visceral Obesity. *Nat Rev Endocrinol*. 2020;16(3):177–189.
- [24] Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*. 2006;444(7121):840–846.
- [25] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- [26] Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820.
- [27] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.

- [28]Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20:310.
- [29]Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
- [30]Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.
- [31]Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI. National Health and Nutrition Examination Survey, 2015–2018: sample design and estimation procedures. *Vital Health Stat 2*. 2020;(184):1–35.
- [32]Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Inf Process Syst*. 2022;35 (Datasets and Benchmarks Track).
- [33]Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019;19:211.
- [34]Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep*. 2020;10:11981.
- [35]Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76–86.
- [36]Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24(6):1052–1061.

Supplementary Material

Table 8: Supplementary Table S1. Final XGBoost hyperparameters and preprocessing specification.

Parameter	Value
Algorithm	XGBoost
Development dataset	NHANES 2015–2016
Temporal validation dataset	NHANES 2017–2018
Number of predictors	41
n_estimators	200
max_depth	3
learning_rate	0.03
min_child_weight	10
subsample	0.80
colsample_bytree	0.90
gamma	0.30
reg_alpha	0
reg_lambda	1
Decision threshold	0.21954

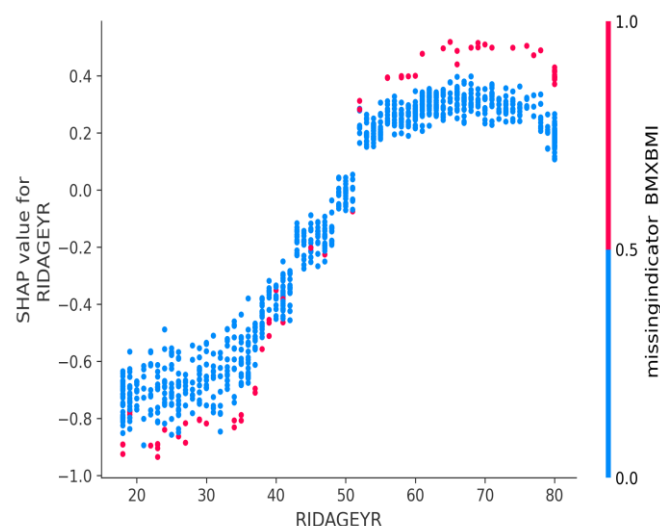


Figure 7: SHAP dependence plot for age (RIDAGEYR).

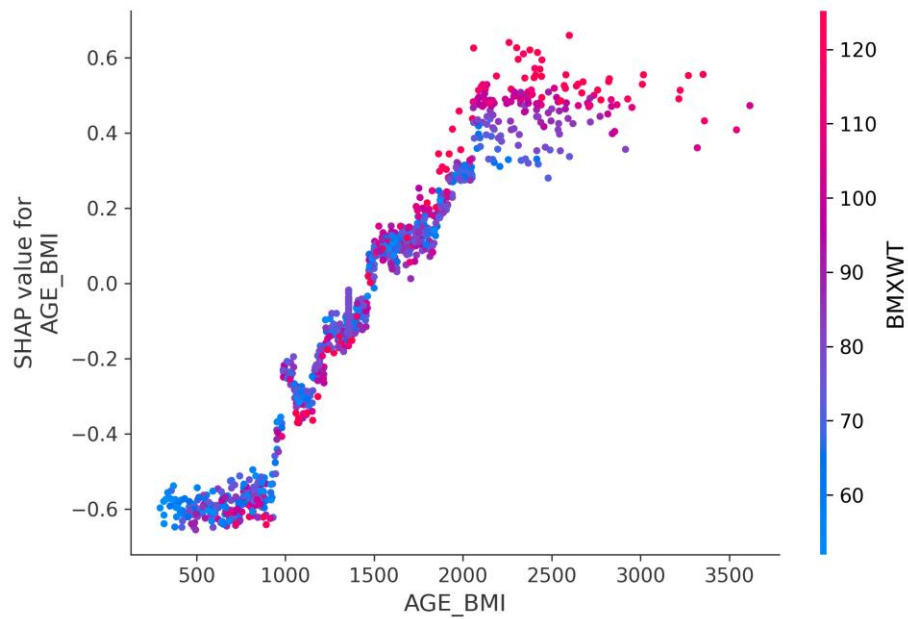


Figure 8: SHAP dependence plot for the age $\times$ BMI interaction term.

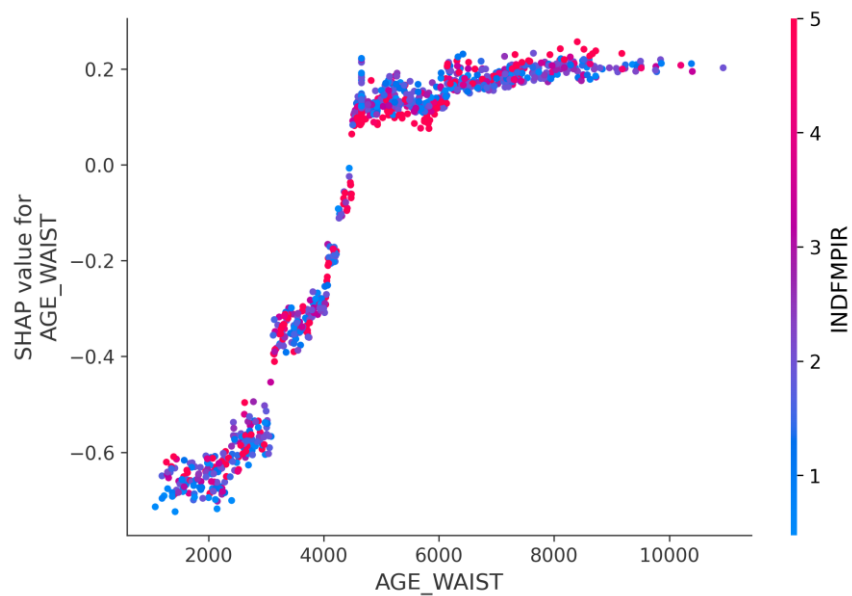


Figure 9: SHAP dependence plot for the age $\times$ waist-circumference interaction term.

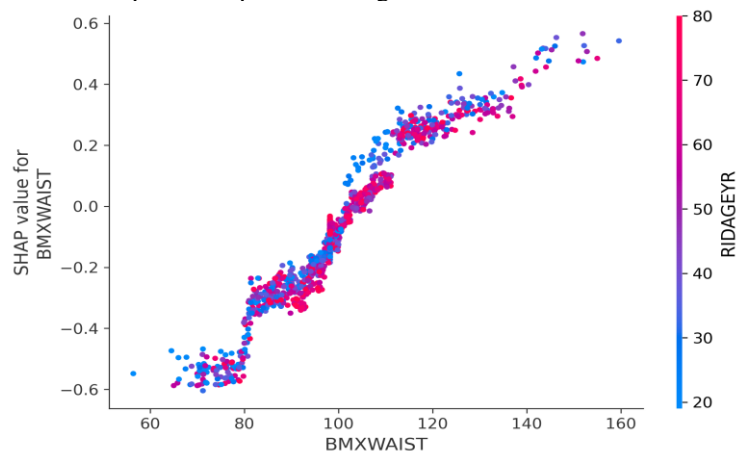


Figure 10: SHAP dependence plot for waist circumference (BMXWAIST).

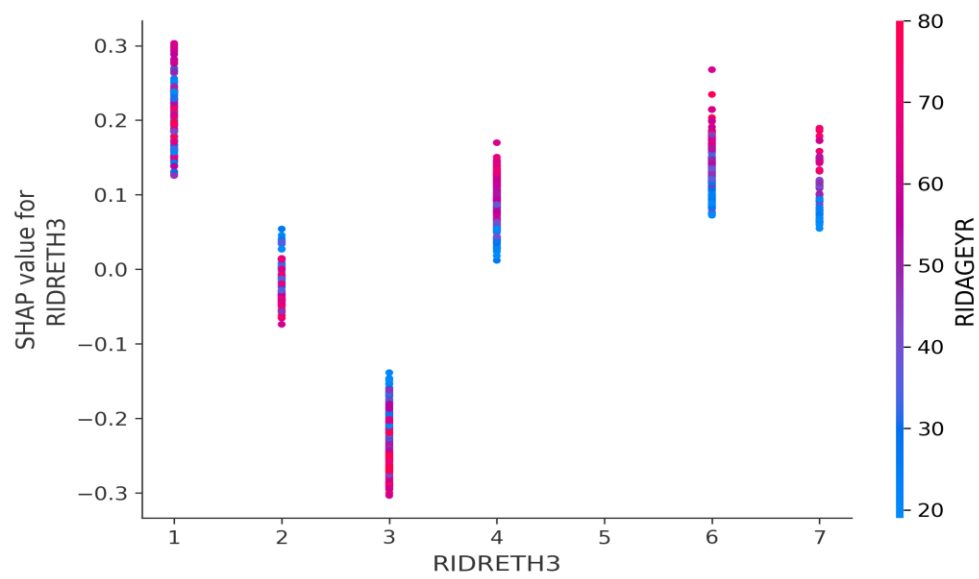


Figure 11: SHAP dependence plot for race/ethnicity (RIDRETH3).