Cancer Screening and Prevention **2026** | Epub
DOI: 10.14218/CSP.2026.00016

Original Article

Modeling Current Colorectal Cancer Screening Status Among Adults Aged 45–75 Years in the 2024 BRFSS: A Comparative Evaluation of Ensemble Machine Learning Models and Logistic Regression

Philips N. Okeagu[1*], Godisgreat O. Okeke[2], Chidindu P. Anagwu[3] and Joseph Atarere[4]

[1]*Department of Oral Health Policy and Epidemiology, Harvard School of Dental Medicine, Boston, MA, USA;* [2]*Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA;* [3]*Redus Center for Digestive Health, Lekki, Lagos, Nigeria;* [4]*Department of Gastroenterology, University of Colorado Hospital, Aurora, CO, USA*

Received: July 29, 2026 | **Revised:** September 14, 2026 | **Accepted:** September 23, 2026 | **Published online:** October 28, 2026

Abstract

Background and objectives: We compared logistic regression, random forest, and extreme gradient boosting (XGBoost) for classifying current colorectal cancer screening status among U.S. adults aged 45–75 years.

Methods: We analyzed 2024 Behavioral Risk Factor Surveillance System data for 227,647 adults aged 45–75 years with determinable screening status. Data were split into stratified training and held-out test sets before imputation. Missing predictors were handled with 20 multiply imputed datasets; imputation and preprocessing were fitted in training data and applied to held-out data. Model performance was assessed using five-fold cross-validation and the held-out test set.

Results: In the test set, areas under the receiver operating characteristic curve were 0.771 for logistic regression, 0.778 for random forest, and 0.781 for XGBoost; corresponding areas under the precision-recall curve were 0.884, 0.888, and 0.890. Brier scores were 0.152, 0.153, and 0.149, respectively. At model-specific Youden thresholds, sensitivity ranged from 75.5% to 78.6% and specificity from 62.5% to 66.3%.

Conclusions: XGBoost has the highest discrimination and lowest Brier score, but differences among models are small. The models classify current screening status in this analytic cohort; external validation is needed before use in screening outreach.

Introduction

Colorectal cancer (CRC) is one of the most commonly diagnosed cancers and a leading cause of cancer-related death in the United States.[1] When men and women are considered together, CRC is the third most commonly diagnosed cancer and the second leading cause of cancer-related death.[1,2] Screening reduces CRC incidence and mortality through detection and removal of precancerous lesions and identification of cancer at an earlier stage.[1] The U.S. Preventive Services Task Force (USPSTF) recommends CRC screening for adults aged 45–75 years using several available strategies, including colonoscopy, fecal immunochemical testing, and stool DNA testing.[2] Despite these recommendations, many eligible adults remain inadequately screened, with disparities related to socioeconomic conditions and access to healthcare.[3,4]

CRC screening status is associated with demographic characteristics, socioeconomic conditions, healthcare access, health behaviors, and medical history.[3,4] These relationships can be examined using regression-based methods, particularly logistic regression.[5] Logistic regression is well suited to estimating associations and produces readily interpretable models, but nonlinear relationships and complex interactions must generally be specified in advance.[6] Machine learning methods can model some of these relationships more flexibly.[7] Random forests, for example, combine multiple

Keywords: Colorectal neoplasms; Colorectal cancer screening; Machine learning; Classification; Logistic regression; Random forest; XGBoost; Behavioral Risk Factor Surveillance System.
***Correspondence to:** Philips N. Okeagu, Department of Oral Health Policy and Epidemiology, Harvard School of Dental Medicine, 188 Longwood Avenue, Boston, MA 02115, USA. ORCID: https://orcid.org/0009-0003-2021-6052. Tel: +1-857-707-9259, E-mail: philipsokeagu@hsph.harvard.edu
How to cite this article: Okeagu PN, Okeke GO, Anagwu CP, Atarere J. Modeling Current Colorectal Cancer Screening Status Among Adults Aged 45–75 Years in the 2024 BRFSS: A Comparative Evaluation of Ensemble Machine Learning Models and Logistic Regression. *Cancer Screen Prev*. Published online: Oct 28, 2026. DOI: 10.14218/CSP.2026.00016.

© 2026 The Author(s). This article has been published under the terms of Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0), which permits noncommercial unrestricted use, distribution, and reproduction in any medium, provided that the following statement is provided. "This article has been published in *Cancer Screening and Prevention* at https://doi.org/10.14218/CSP.2026.00016 and can also be viewed on the Journal's website at https://www.xiahepublishing.com/journal/csp".

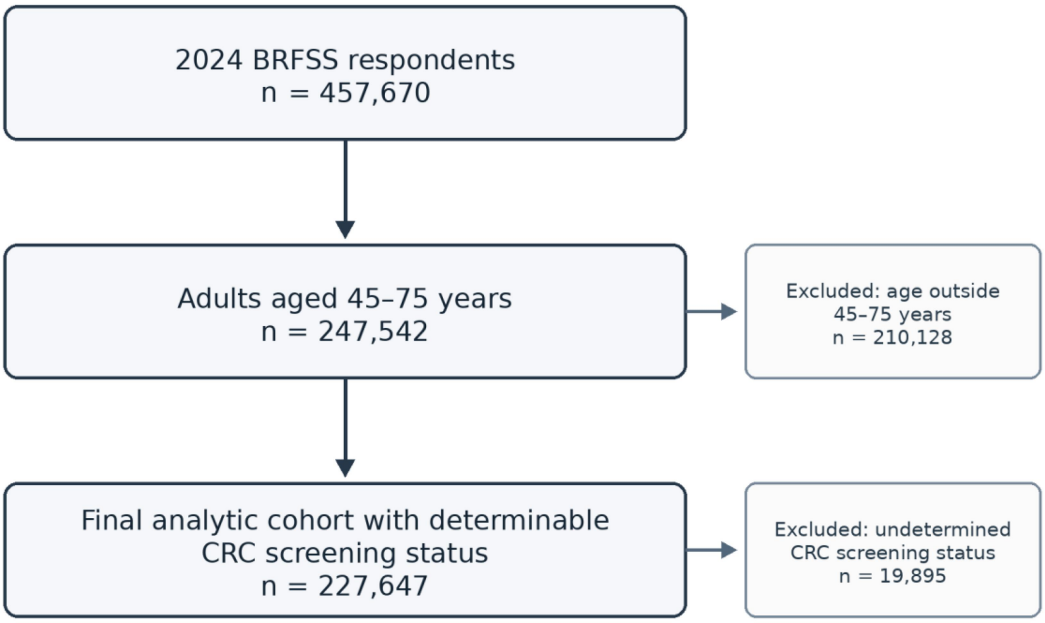


Fig. 1. Participant selection flow diagram. The analytic sample was restricted to BRFSS respondents aged 45–75 years with determinable colorectal cancer (CRC) screening status. BRFSS, Behavioral Risk Factor Surveillance System.

decision trees, while extreme gradient boosting (XGBoost) builds trees sequentially to correct errors from earlier trees.[8,9]

Greater model complexity, however, does not guarantee better classification. Studies comparing machine learning with logistic regression in structured clinical data have often found little improvement in discrimination despite substantially more complex models.[7,10] Whether ensemble methods offer a meaningful advantage therefore depends on the data and the classification task. This question is relevant to CRC screening because a model used for outreach would need to identify adults who are not up-to-date with screening accurately enough to justify the additional complexity of implementation.

Machine learning has been used to classify adherence to other cancer-related preventive and treatment behaviors, including lung cancer screening and endocrine therapy.[11,12] Evidence specific to CRC screening is more limited, particularly for direct comparisons between ensemble methods and logistic regression in a large U.S. survey sample. The Behavioral Risk Factor Surveillance System (BRFSS) contains information on CRC screening as well as demographic, socioeconomic, healthcare access, behavioral, and health characteristics, providing a large dataset in which these approaches can be compared.[13]

In this study, we compared logistic regression, random forest, and XGBoost for classifying current CRC screening status among adults aged 45–75 years in the 2024 BRFSS. We evaluated discrimination and calibration and examined the predictors that contributed most to each model. We were particularly interested in whether the added complexity of the ensemble models produced a meaningful improvement over logistic regression and how well the models identified adults who were not up-to-date with CRC screening.

Materials and methods

Data source

Data were obtained from the 2024 BRFSS, an annual, state-based, cross-sectional telephone survey conducted by the Centers for Disease Control and Prevention (CDC) in collaboration with state and territorial health departments.[13] The 2024 public-use dataset included data from 49 states, the District of Columbia, Guam, Puerto Rico, and the U.S. Virgin Islands; Tennessee was not included because insufficient data were collected to meet the minimum requirements for the public-use dataset.[13] BRFSS uses a complex, stratified, multistage probability sampling design to collect information on health behaviors, chronic conditions, healthcare access, and preventive health practices among noninstitutionalized U.S. adults aged ≥ 18 years.[13]

Study population

The 2024 BRFSS included 457,670 respondents. We restricted the sample to adults aged 45–75 years, corresponding to the age range for routine CRC screening recommended by the USPSTF (n = 247,542).[2] We then excluded 19,895 respondents for whom CRC screening status could not be determined from the CDC-derived CRC screening recommendation variable (_CRCREC3). The final analytic cohort included 227,647 respondents (Fig. 1). No formal sample-size calculation was performed; all respondents in the 2024 BRFSS public-use dataset who met the eligibility criteria and had determinable CRC screening status were included.

The USPSTF recommendation applies to asymptomatic adults

at average risk for CRC.[2] BRFSS did not permit complete ascertainment of all criteria defining this population for the present analysis, including current CRC symptoms and all high-risk conditions such as prior adenomatous polyps, inflammatory bowel disease, or hereditary CRC syndromes. The analytic cohort was therefore defined by age and availability of CRC screening status and should not be interpreted as a strictly verified asymptomatic, average-risk population.

Outcome definition

The outcome was current CRC screening status based on the CDC-derived _CRCREC3 variable. Respondents classified by the CDC as having completed at least one recommended CRC screening test within the recommended interval were classified as up-to-date (coded as 1). Respondents who had not completed a recommended test within the specified interval or had never undergone a recommended CRC screening test were classified as not up-to-date (coded as 0). Respondents with insufficient information to determine screening status were excluded.

For adults aged 45–75 years, _CRCREC3 incorporates fecal occult blood testing within the previous year, sigmoidoscopy within the previous 5 years, colonoscopy within the previous 10 years, stool DNA testing within the previous 3 years, virtual colonoscopy within the previous 5 years, or sigmoidoscopy within the previous 10 years combined with fecal occult blood testing within the previous year.[13]

Up-to-date screening was designated as the positive class for binary classification analyses, consistent with the original modeling workflow. Because respondents who were not up-to-date represent the group most relevant to screening outreach, performance for correctly identifying this group was also examined through specificity at the evaluated classification thresholds.

Predictor selection

Candidate predictors were selected a priori based on previous literature, clinical relevance, and availability in BRFSS.[4] The 41 candidate predictors covered demographic characteristics, socioeconomic status, healthcare access and utilization, health behaviors, chronic medical conditions, functional health, and geographic characteristics.

Demographic variables included age, sex, race/ethnicity, and marital status. Socioeconomic variables included education, household income, employment, home ownership, and veteran status. Healthcare variables included insurance status and type, presence of a personal healthcare provider, cost-related barriers to care, and time since the last routine checkup. Health behaviors included smoking, e-cigarette use, alcohol use, physical activity, and body mass index. Medical and functional health variables included chronic disease history, general health status, physical and mental health days, and functional limitations. State and urban/rural classification were also included.

Where multiple BRFSS variables measured similar constructs, CDC-derived or collapsed variables were preferentially selected to reduce redundancy and sparse categories. A complete list of candidate predictors, BRFSS variable names, and coding definitions is provided in Supplementary Table. 1.

Candidate predictors were reviewed for potential overlap with the outcome definition. Variables representing receipt, timing, or type of CRC screening, including the component screening variables used to construct _CRCREC3, were not included as predictors. Broader measures of healthcare utilization, such as time since the last routine checkup and presence of a personal healthcare provider, were retained because they measure healthcare engagement rather than CRC screening itself.

Data preprocessing

Categorical predictors were converted to indicator variables using treatment coding, and continuous predictors were retained as numeric variables. Imputation, encoding, and other preprocessing steps were estimated using training data and applied to the corresponding validation or held-out test data. Predictors with sparse categories were retained when clinically relevant, and near-zero variance filtering was not performed.

Multicollinearity was assessed for the logistic regression model using variance inflation factors; all variance inflation factors were < 5.

The BRFSS dataset was initially subset to the outcome, candidate predictors, and three survey-design variables: the final sampling weight, primary sampling unit, and stratification variable. Survey-design variables were retained during data preparation but were not included as model predictors.

Missingness among candidate predictors ranged from 0% to 14.9%, with household income having the highest proportion of missing values. Missing predictor data were addressed using multiple imputation by chained equations with 20 imputations.[14] Continuous variables were imputed using predictive mean matching, and categorical variables were imputed using logistic or polytomous regression, as appropriate.

The train-test split was performed before imputation. Within each cross-validation iteration, imputation models were estimated using only the training fold and applied to the validation predictors. For final evaluation, imputation models were estimated using the training set and applied to the held-out test predictors; test outcomes were not used to impute predictors.

Each model was fitted and used to generate predictions across the 20 completed datasets. For each participant, predicted probabilities were averaged across imputations before performance was calculated.

Data partitioning and survey design

The analytic cohort was randomly partitioned into training (80%; n = 182,118) and test (20%; n = 45,529) sets using stratified sampling on CRC screening status. The test set was held out from model development. Hyperparameter tuning was performed within the training set using five-fold stratified cross-validation.

Survey weights, primary sampling units, and stratification variables were not incorporated into model fitting. The primary objective was to compare the classification performance of logistic regression, random forest, and XGBoost under the same analytic framework rather than to estimate nationally representative CRC screening prevalence or population regression parameters. The resulting performance estimates therefore describe classification within the analytic cohort and should not be interpreted as design-weighted estimates of model performance in the U.S. population.

Model development

Three binary classification approaches were evaluated: logistic

regression, random forest, and XGBoost. Logistic regression was implemented using the glm engine. Random forest was implemented using the ranger engine, and XGBoost was implemented using the xgboost engine.[5,8,9]

Both ensemble models were specified with 500 trees. Random search was used to tune random forest and XGBoost hyperparameters using five-fold stratified cross-validation within the training set.

Class imbalance

Up-to-date screening represented approximately 73% of the analytic cohort, resulting in moderate class imbalance. The synthetic minority oversampling technique (SMOTE) was used during model development to increase representation of the not-up-to-date class during training.[15] SMOTE was applied only to training observations within each cross-validation fold. Synthetic observations were not introduced into the held-out test set, which retained the observed outcome distribution.

Threshold-independent model performance was assessed using the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (PR-AUC), and Brier score. Threshold-dependent performance was evaluated separately using model-specific Youden-derived classification thresholds. Predicted probabilities were not recalibrated after SMOTE.

Hyperparameter tuning

For random forest, the number of predictors considered at each split (*mtry*) and the minimum number of observations required for a terminal node (*min_n*) were tuned. The number of trees was fixed at 500.

For XGBoost, tuned parameters included tree depth (*tree_depth*), learning rate (*learn_rate*), minimum node size (*min_n*), minimum loss reduction required for an additional split (*loss_reduction*), and the proportion of observations sampled for each boosting iteration (*sample_prop*). The number of boosting trees was fixed at 500.

The selected random forest parameters were *mtry* = 5 and *min_n* = 14. The selected XGBoost parameters were *tree_depth* = 3, *learn_rate* = 0.268, *min_n* = 19, *loss_reduction* = 0, and *sample_prop* = 0.690.

The available analysis records do not preserve the parameter search ranges, number of candidate combinations, or metric used to select the final settings.

Threshold selection

For each model, a classification threshold was selected by maximizing Youden's *J* statistic (sensitivity + specificity − 1) using cross-validated predictions from the training set.[16] The resulting model-specific threshold was then applied to the held-out test set to calculate sensitivity, specificity, accuracy, and Youden's *J*.

Youden's *J* was used as a common criterion for model comparison because it gives equal weight to sensitivity and specificity. The resulting thresholds were used for comparative evaluation and were not intended to represent recommended operational thresholds for CRC screening outreach. In practice, an outreach program may assign different consequences to missing a respondent who is not up-to-date and contacting someone who is already screened.

For descriptive comparison, specificity was also calculated at the conventional probability threshold of 0.50. AUROC, PR-AUC, and Brier score were calculated independently of a classification threshold.

Model evaluation

Final model performance was evaluated in the held-out test set. Discrimination was assessed using AUROC and PR-AUC. Because up-to-date screening was designated as the positive class, the reported PR-AUC describes precision-recall performance for the up-to-date class.

Threshold-dependent performance at the model-specific Youden-derived thresholds was summarized using sensitivity, specificity, accuracy, F1 score, and Youden's *J*. F1 scores were calculated for the up-to-date class, which was designated as the positive class. Sensitivity represents the proportion of up-to-date respondents correctly classified, whereas specificity represents the proportion of not-up-to-date respondents correctly classified.

Calibration was evaluated using the Brier score and calibration plots comparing predicted probabilities with observed outcome proportions. Lower Brier scores indicate smaller overall differences between predicted probabilities and observed outcomes.

Performance estimates were compared descriptively. Confidence intervals and formal pairwise statistical tests comparing model performance were not calculated; numerical differences between models were therefore not interpreted as evidence of statistical superiority.

Variable importance and model interpretation

Variable importance was evaluated separately for each modeling approach. For logistic regression, importance was represented by the absolute Wald *z*-statistic for each coefficient, as returned by the vip package for a generalized linear model. For random forest, permutation importance measured the change in model performance after individual predictors were randomly permuted. For XGBoost, gain represented the contribution of predictors to improvements in tree-splitting performance. The 20 highest-ranked predictors from each model were visualized.

Because these importance measures have different definitions and scales, their magnitudes were not compared directly across models. Cross-model comparisons were limited to whether predictors consistently appeared among the highest-ranked variables. Variable importance was interpreted as predictive contribution to the fitted model and not as evidence of a causal effect on CRC screening status.

Reporting guidelines

Reporting was guided by the TRIPOD+AI statement and the Strengthening the Reporting of Observational Studies in Epidemiology statement.[17] Completed reporting checklists are provided in Supplementary Table. 2 and 3.

Software

Analyses were performed in R version 4.4.2 using the tidymodels framework, including recipes, parsnip, workflows, tune, and yardstick. Additional packages included dplyr, tidyr, and purrr for data management; haven for BRFSS data import; ranger for random forest; xgboost for gradient boosting; probably for threshold selection; vip for variable importance; ggplot2 for

figures; mice for multiple imputation; and themis for SMOTE. A fixed random seed was specified before data partitioning, with additional seeds used during model tuning.

Results

Study population

The final analytic cohort included 227,647 adults aged 45–75 years. Of these, 167,374 (73.5%) were up-to-date with CRC screening and 60,273 (26.5%) were not up-to-date (Table 1).

Respondents who were up-to-date with CRC screening were older than those who were not up-to-date (mean age, 62.8 *vs.* 57.4 years; $P < 0.001$). Non-Hispanic White respondents accounted for 80% of the up-to-date group and 71% of the not-up-to-date group, whereas Hispanic respondents accounted for 6.2% and 12%, respectively. College graduates represented 47% of respondents who were up-to-date and 36% of those who were not. Among respondents with available income data, 66% of those who were up-to-date reported annual household income of ≥ $50,000 compared with 55% of those who were not up-to-date.

Differences were also observed in healthcare access and utilization. Health insurance coverage was reported by 98% of respondents who were up-to-date compared with 90% of those who were not, and Medicare coverage was more common in the up-to-date group (42% *vs.* 27%). Absence of a personal healthcare provider was more common among respondents who were not up-to-date (19% *vs.* 3.4%), as were cost-related barriers to care (14% *vs.* 5.6%). A routine checkup within the previous year was reported by 92% of respondents who were up-to-date compared with 71% of those who were not. The complete distribution of baseline characteristics is presented in Table 1.

Comparative model performance

Logistic regression, random forest, and XGBoost were evaluated in the held-out test set of 45,529 respondents (Table 2). AUROC values were 0.771 for logistic regression, 0.778 for random forest, and 0.781 for XGBoost (Fig. 2). Corresponding PR-AUC values were 0.884, 0.888, and 0.890, respectively. XGBoost had the numerically highest AUROC and PR-AUC, although the absolute differences among models were small. The AUROC difference between XGBoost and logistic regression was 0.010. Formal statistical comparisons of model performance were not performed.

Brier scores were 0.152 for logistic regression, 0.153 for random forest, and 0.149 for XGBoost. XGBoost had the numerically lowest Brier score, although differences across models were small. Calibration plots showed generally close agreement between predicted and observed probabilities for logistic regression and XGBoost, whereas random forest showed greater deviation at lower predicted probabilities (Fig. 3).

Up-to-date screening was coded as the positive class in the modeling workflow. Accordingly, the reported PR-AUC values describe precision-recall performance for the up-to-date class and should not be interpreted as precision-recall performance for identifying respondents who were not up-to-date.

Classification performance at Youden-derived thresholds

The thresholds that maximized Youden's *J* statistic in cross-validated training predictions were 0.71 for logistic regression, 0.72 for random forest, and 0.74 for XGBoost (Table 3). These model-specific thresholds were then applied to the held-out test set.

At these thresholds, sensitivity for the up-to-date class was 78.6% for logistic regression, 75.5% for random forest, and 75.8% for XGBoost. Specificity was 62.5% for logistic regression and 66.3% for both random forest and XGBoost. F1 scores for the up-to-date positive class were 0.82, 0.80, and 0.81, respectively. Because up-to-date screening was designated as the positive class, specificity represents the proportion of respondents who were not up-to-date who were correctly classified as not up-to-date. Youden's *J* values were 0.412, 0.418, and 0.422 for logistic regression, random forest, and XGBoost, respectively. Accuracy at the Youden-derived thresholds was 74.4%, 73.0%, and 73.3%, respectively.

Compared with the conventional 0.50 threshold, the Youden-derived thresholds increased specificity from 33.8% to 62.5% for logistic regression, from 28.4% to 66.3% for random forest, and from 38.5% to 66.3% for XGBoost. This increase in the proportion of not-up-to-date respondents correctly classified was accompanied by lower sensitivity for the up-to-date class.

Model specifications

The selected random forest hyperparameters were *mtry* = 5 and *min_n* = 14. For XGBoost, the selected parameters were *tree_depth* = 3, *learn_rate* = 0.268, *min_n* = 19, *loss_reduction* = 0, and *sample_prop* = 0.690. Both ensemble models used 500 trees.

Variable importance

Age was the highest-ranked predictor of CRC screening status in all three models (Figs. 4, 5 and 6). Time since the last routine medical checkup and absence of a personal healthcare provider were also among the highest-ranked predictors across logistic regression, random forest, and XGBoost.

Other predictors that ranked highly across models included retirement status, college education, Medicare coverage, arthritis, and home rental status. Body mass index was also highly ranked in the XGBoost model. Healthcare access and utilization variables, including recency of routine care, personal healthcare provider status, and insurance-related measures, appeared consistently among the higher-ranked predictors across modeling approaches.

Because variable importance was quantified differently for each model, the magnitude of importance scores was not compared directly across approaches. Absolute Wald *z*-statistics were used for logistic regression, permutation importance for random forest, and gain for XGBoost. The recurrence of several predictors among the highest-ranked variables reflects consistency in predictive contribution across models and should not be interpreted as evidence of causal influence on CRC screening status.

Discussion

In this analysis of 227,647 adults aged 45–75 years, random forest and XGBoost provided only modest improvements over logistic regression for classifying current CRC screening status. XGBoost had the highest AUROC (0.781) and PR-AUC (0.890) and the lowest Brier score (0.149), but the absolute differences across models were small. The AUROC difference between XGBoost

Table 1. Baseline characteristics of adults aged 45–75 years by colorectal cancer screening status, BRFSS 2024

Characteristic	Overall *N* = 227,647	Up-to-date *N* = 167,374	Not up-to-date *N* = 60,273	*P* value
Age, years	61.4 (8.7)	62.8 (8.1)	57.4 (9.0)	< 0.001
Sex				< 0.001
Male	106,769 (47%)	77,031 (46%)	29,738 (49%)	
Female	120,878 (53%)	90,343 (54%)	30,535 (51%)	
Race/ethnicity				< 0.001
White NH	174,152 (78%)	132,329 (80%)	41,823 (71%)	
Black NH	17,272 (7.7%)	12,798 (7.8%)	4,474 (7.6%)	
Other NH	10,158 (4.5%)	6,300 (3.8%)	3,858 (6.6%)	
Multiracial NH	4,443 (2.0%)	3,014 (1.8%)	1,429 (2.4%)	
Hispanic	17,405 (7.8%)	10,289 (6.2%)	7,116 (12%)	
Missing	4,217	2,644	1,573	
Education				< 0.001
Less than HS	12,196 (5.4%)	6,815 (4.1%)	5,381 (9.0%)	
HS graduate	52,686 (23%)	36,117 (22%)	16,569 (28%)	
Some college	62,575 (28%)	46,056 (28%)	16,519 (28%)	
College graduate	99,322 (44%)	77,929 (47%)	21,393 (36%)	
Missing	868	457	411	
Household income				< 0.001
Less than $15,000	11,320 (5.8%)	6,917 (4.8%)	4,403 (8.7%)	
$15,000 to < $25,000	17,221 (8.9%)	11,460 (8.0%)	5,761 (11%)	
$25,000 to < $35,000	19,923 (10%)	13,747 (9.6%)	6,176 (12%)	
$35,000 to < $50,000	23,510 (12%)	16,889 (12%)	6,621 (13%)	
$50,000 to < $100,000	58,801 (30%)	45,098 (32%)	13,703 (27%)	
$100,000 to < $200,000	45,753 (24%)	35,234 (25%)	10,519 (21%)	
$200,000 or more	17,118 (8.8%)	13,561 (9.5%)	3,557 (7.0%)	
Missing	34,001	24,468	9,533	
Has health insurance	212,908 (96%)	160,783 (98%)	52,125 (90%)	< 0.001
Missing	6,329	4,070	2,259	
Insurance type				< 0.001
Employer/union	77,709 (36%)	55,966 (35%)	21,743 (42%)	
Private plan	19,482 (9.2%)	14,080 (8.8%)	5,402 (10%)	
Medicare	81,369 (38%)	67,437 (42%)	13,932 (27%)	
Medigap	193 (< 0.1%)	151 (< 0.1%)	42 (< 0.1%)	
Medicaid	14,054 (6.6%)	8,602 (5.4%)	5,452 (10%)	
CHIP	37 (< 0.1%)	22 (< 0.1%)	15 (< 0.1%)	
Military/VA	8,366 (3.9%)	6,770 (4.2%)	1,596 (3.1%)	
IHS	733 (0.3%)	418 (0.3%)	315 (0.6%)	
State plan	5,736 (2.7%)	3,719 (2.3%)	2,017 (3.9%)	
Other government	5,228 (2.5%)	3,618 (2.3%)	1,610 (3.1%)	

Table 1. Baseline characteristics of adults aged 45–75 years by colorectal cancer screening status, BRFSS 2024 (*continued*)

Characteristic	Overall *N* = 227,647	Up-to-date *N* = 167,374	Not up-to-date *N* = 60,273	*P* value
Missing	14,740	6,591	8,149	
Has personal healthcare provider				< 0.001
Yes—one	125,459 (55%)	93,775 (56%)	31,684 (53%)	
Yes—more than one	83,856 (37%)	67,007 (40%)	16,849 (28%)	
No	16,751 (7.4%)	5,670 (3.4%)	11,081 (19%)	
Missing	1,581	922	659	
Cost-related barrier to care	17,469 (7.7%)	9,308 (5.6%)	8,161 (14%)	< 0.001
Missing	666	349	317	
Time since last routine checkup				< 0.001
< 1yr	195,272 (87%)	153,320 (92%)	41,952 (71%)	
1–2 years	15,948 (7.1%)	9,287 (5.6%)	6,661 (11%)	
2–5 years	7,101 (3.2%)	2,658 (1.6%)	4,443 (7.6%)	
≥ 5 years	6,926 (3.1%)	1,182 (0.7%)	5,744 (9.8%)	
Missing	2,400	927	1,473	
Smoking status				< 0.001
Current—daily	21,546 (9.6%)	12,960 (7.8%)	8,586 (14%)	
Current—some days	7,513 (3.3%)	4,710 (2.8%)	2,803 (4.7%)	
Former	67,777 (30%)	51,863 (31%)	15,914 (27%)	
Never	127,953 (57%)	95,811 (58%)	32,142 (54%)	
Missing	2,858	2,030	828	
Any physical activity (past 30 days)	173,204 (76%)	129,819 (78%)	43,385 (72%)	< 0.001
Missing	540	337	203	
Diabetes				< 0.001
Yes	42,007 (18%)	32,675 (20%)	9,332 (16%)	
Yes, gestational	1,388 (0.6%)	923 (0.6%)	465 (0.8%)	
No	176,988 (78%)	128,352 (77%)	48,636 (81%)	
Prediabetes	6,907 (3.0%)	5,195 (3.1%)	1,712 (2.8%)	
Missing	357	229	128	
Cardiovascular disease	23,609 (10%)	18,526 (11%)	5,083 (8.5%)	< 0.001
Missing	2,158	1,469	689	
Arthritis	96,989 (43%)	77,769 (47%)	19,220 (32%)	< 0.001
Missing	1,074	732	342	
Non-melanoma skin cancer history	21,542 (9.5%)	18,539 (11%)	3,003 (5.0%)	< 0.001
Missing	1,384	1,029	355	
Other cancer history	31,367 (14%)	26,603 (16%)	4,764 (7.9%)	< 0.001
Missing	1,117	813	304	
Urban/rural residence				< 0.001
Urban	186,972 (85%)	138,653 (86%)	48,319 (83%)	
Rural	32,930 (15%)	23,299 (14%)	9,631 (17%)	

Values are presented as mean (SD) or *n* (%). Percentages are calculated among respondents with non-missing data for each characteristic. Comparisons were performed using the Wilcoxon rank-sum test for continuous variables and Pearson's chi-square test for categorical variables. Values and comparisons are unweighted and are not intended as population-level estimates. BRFSS, Behavioral Risk Factor Surveillance System; CHIP, Children's Health Insurance Program; HS, high school; IHS, Indian Health Service; NH, non-Hispanic; SD, standard deviation; VA, Veterans Affairs.

Table 2. Performance comparison of logistic regression and ensemble machine learning models for classifying current colorectal cancer screening status

Model	AUROC	PR-AUC*	Brier score
Logistic regression	0.771	0.884	0.152
Random forest	0.778	0.888	0.153
XGBoost	0.781	0.890	0.149

Model performance was evaluated in the held-out test set (*n* = 45,529) following an 80/20 stratified train-test split. Up-to-date screening was designated as the positive class; therefore, PR-AUC describes precision-recall performance for the up-to-date class. *PR-AUC refers to the area under the precision-recall curve with up-to-date screening as the positive class. The Brier score measures the mean squared difference between predicted probabilities and observed outcomes, with lower values indicating better probabilistic accuracy. AUROC, area under the receiver operating characteristic curve; PR-AUC, area under the precision-recall curve; XGBoost, extreme gradient boosting.

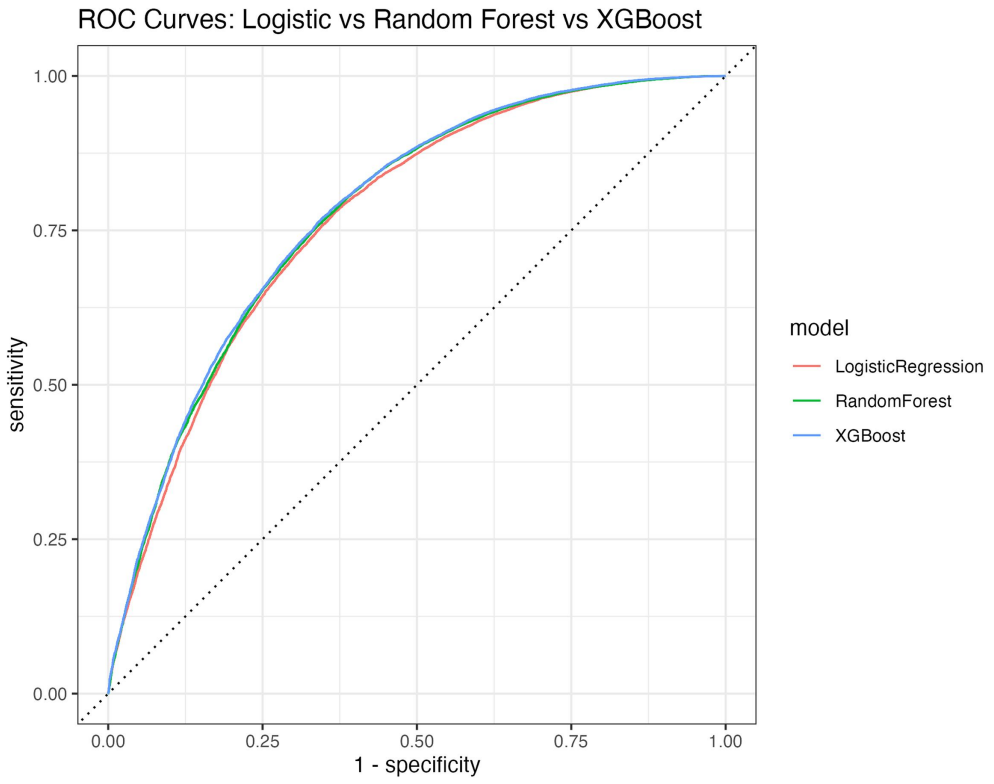


Fig. 2. Receiver operating characteristic (ROC) curves for logistic regression, random forest, and XGBoost models classifying current colorectal cancer screening status. ROC curves show model discrimination in the held-out test dataset (*n* = 45,529), plotting sensitivity against 1 − specificity across classification thresholds. The dotted diagonal line represents no discrimination (AUROC = 0.50). AUROC values were 0.771 for logistic regression, 0.778 for random forest, and 0.781 for XGBoost. AUROC, area under the receiver operating characteristic curve; XGBoost, extreme gradient boosting.

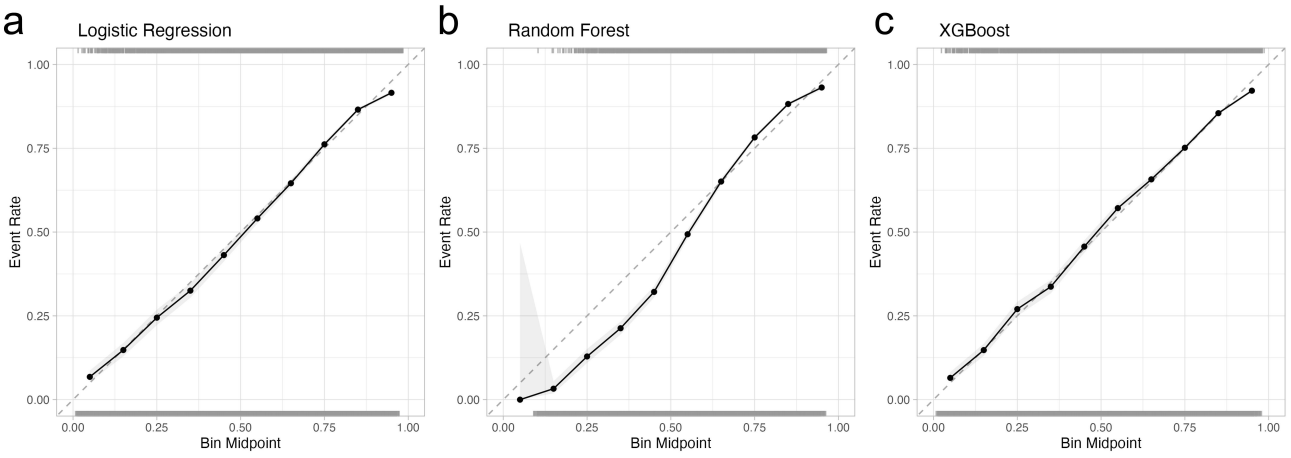


Fig. 3. Calibration plots for logistic regression, random forest, and XGBoost models classifying current colorectal cancer screening status. (a) Logistic regression, (b) random forest, and (c) XGBoost. Each panel plots the midpoint of each predicted-probability bin (x-axis) against the observed proportion of up-to-date respondents within each bin (y-axis) in the held-out test dataset (n = 45,529). The dashed diagonal line represents perfect calibration, where predicted probabilities equal observed proportions. XGBoost, extreme gradient boosting.

Table 3. Classification performance at model-specific Youden-derived probability thresholds

Model	Youden-derived threshold	Sensitivity (%)	Specificity (%)	Youden's *J*	Accuracy (%)	F1 score
Logistic regression	0.71	78.6	62.5	0.412	74.4	0.82
Random forest	0.72	75.5	66.3	0.418	73.0	0.80
XGBoost	0.74	75.8	66.3	0.422	73.3	0.81

Model-specific thresholds were selected by maximizing Youden's *J* statistic (sensitivity + specificity − 1) using cross-validated training predictions and were then applied to the held-out test set. Up-to-date screening was designated as the positive class; sensitivity and F1 score are reported for this class, whereas specificity is the proportion of not-up-to-date respondents correctly classified. These thresholds were used for model comparison and should not be interpreted as recommended operational thresholds for screening outreach. XGBoost, extreme gradient boosting.

and logistic regression was only 0.010. These findings are consistent with prior work showing that machine-learning methods do not necessarily provide substantial gains over logistic regression when applied to structured clinical or population data.[10] In this setting, added model complexity did not translate into a large improvement in classification performance.

Performance for respondents who were not up-to-date with screening is particularly relevant because this is the group that screening outreach programs would seek to identify. At the Youden-derived thresholds, specificity ranged from 62.5% for logistic regression to 66.3% for random forest and XGBoost, meaning that approximately one-third of respondents who were not up-to-date were still misclassified. The conventional 0.50 threshold performed substantially worse for this group, with specificity ranging from 28.4% to 38.5%. Raising the classification threshold improved identification of respondents who were not up-to-date but reduced sensitivity for those who were already up-to-date. This tradeoff illustrates why a threshold selected for statistical comparison should not automatically be adopted for implementation. An outreach program with limited capacity may choose a different threshold from one seeking to reach as many unscreened individuals as possible.

Calibration was broadly similar across models. Logistic regression and XGBoost showed relatively close agreement between predicted and observed probabilities, while random forest showed greater deviation at lower predicted probabilities. Differences in Brier scores were also small. This matters because a model used to prioritize outreach would need reasonably accurate probabilities, not simply good ranking of individuals. Even a model with acceptable discrimination may perform poorly in practice if predicted probabilities are poorly calibrated in the population where it is deployed. Local calibration would therefore be necessary before using any of these models to allocate outreach resources.[17]

Age was the highest-ranked predictor in all three models despite restriction of the sample to adults aged 45–75 years. Older adults have had more opportunities to undergo screening and may have more frequent contact with the healthcare system, Medicare coverage, and chronic conditions that increase interaction with clinicians. Measures of healthcare engagement were also consistently prominent. Time since the last routine checkup and absence of a personal healthcare provider ranked highly across modeling approaches, consistent with prior evidence linking access to regular medical care with CRC screening participation.[4] Other highly ranked characteristics, including Medicare coverage and retirement status, may similarly capture differences in healthcare contact and access. These findings should not be

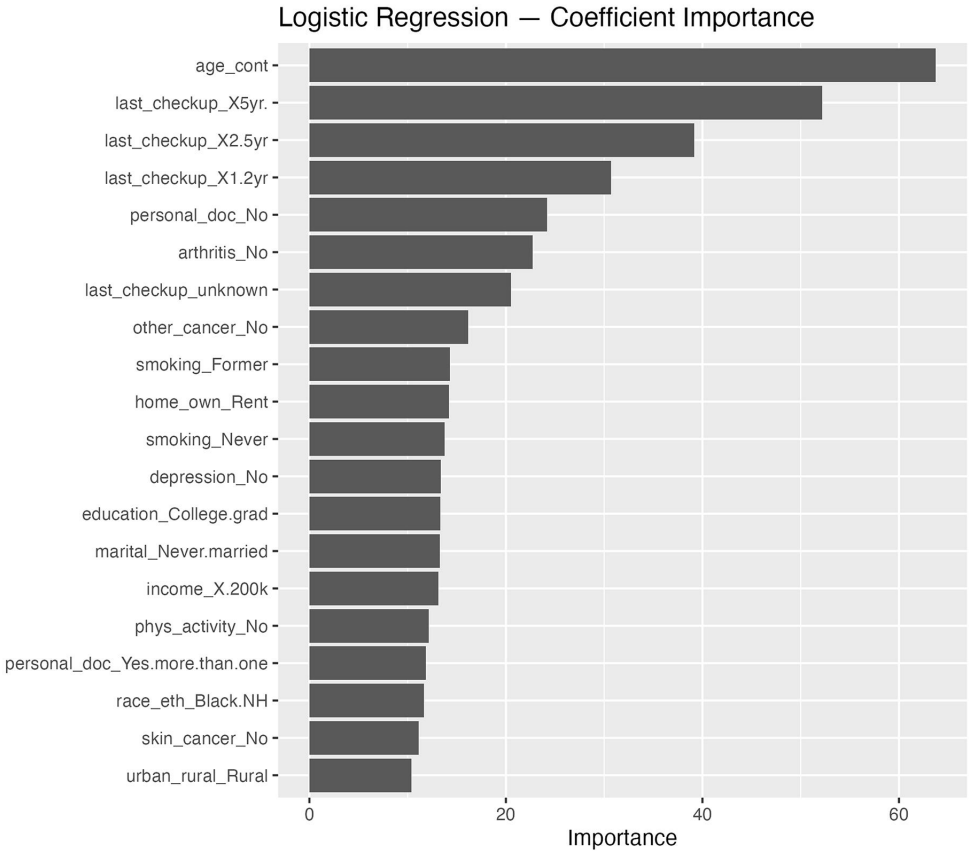


Fig. 4. Variable importance for the logistic regression model classifying current colorectal cancer screening status. Bars represent absolute Wald *z*-statistics for the 20 highest-ranked predictors, with larger values indicating greater contribution to model classification. Variable names reflect one-hot encoded categorical predictor levels. NH, non-Hispanic.

interpreted causally because the importance measures quantify contribution to classification rather than the effect of changing a predictor on screening behavior.

The modest advantage of XGBoost also has practical implications. Logistic regression is easier to interpret, reproduce, and implement, while XGBoost requires additional tuning and is less transparent. A small improvement in discrimination could still matter when a model is applied to a very large population, but this study does not establish that the observed difference would translate into more completed screenings or more efficient use of outreach resources. The appropriate model therefore depends on whether the incremental performance of a more complex approach is sufficient to justify the additional implementation burden. Before deployment, the required predictors would also need to be reliably available in the intended data system, and model performance, calibration, and operating thresholds would need to be evaluated locally.[17]

Several limitations should be considered. BRFSS data are self-reported and cross-sectional, so the models classify current screening status rather than predict future screening behavior. The cohort was defined by age and availability of the CDC-derived screening variable, and BRFSS did not permit complete verification that all respondents were asymptomatic and at average CRC risk.[2,13] Respondents with missing CRC screening status were excluded, which may have introduced selection bias. Missing predictor values were addressed using multiple imputation under a missing-at-random assumption that cannot be verified and may be particularly consequential for household income. Although imputation and preprocessing were fitted within training data and applied to held-out observations, performance was evaluated in a single random split from one survey year; external and temporal validation are needed.

The analysis also did not incorporate BRFSS survey weights, clustering, or stratification into model fitting. The reported performance therefore applies to the analytic cohort and should not be interpreted as design-weighted model performance for the U.S. population. Confidence intervals could not be calculated because individual-level test-set predictions were not available in the retained analysis files. Formal paired comparisons were not performed, so numerical differences between models should not be interpreted as evidence of statistical superiority. Performance was also not evaluated separately across sociodemographic subgroups, limiting conclusions about whether the models perform similarly across populations. Finally, development and evaluation used a single year of BRFSS data. Temporal and external validation are needed before these models are considered for

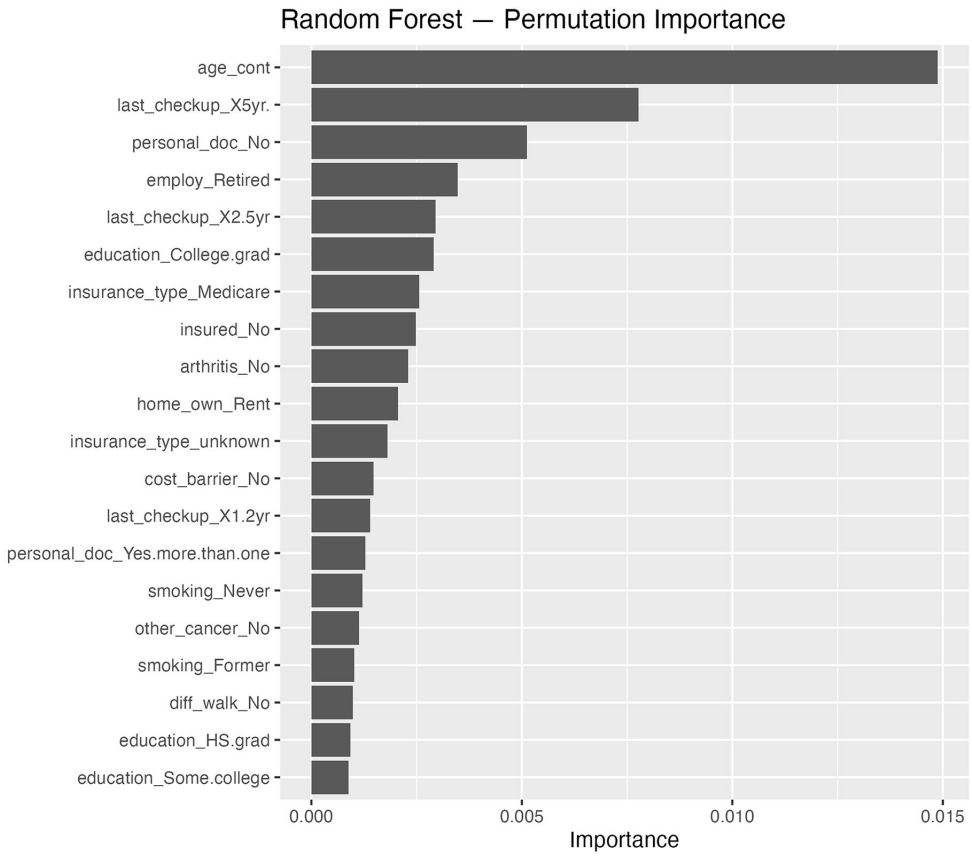


Fig. 5. Variable importance for the random forest model classifying current colorectal cancer screening status. Bars represent permutation-based variable importance for the 20 highest-ranked predictors, reflecting the change in model performance after random permutation of each predictor. Larger values indicate greater predictive contribution. Variable names reflect one-hot encoded categorical predictor levels. HS, high school.

routine use.[17]

Conclusions

Overall, random forest and XGBoost provide only modest improvements over logistic regression for classifying current CRC screening status. Healthcare access and utilization factors are consistently among the strongest predictors, while performance remains weaker for identifying respondents who were not up-to-date. Future work should determine whether these models maintain their performance in other populations and time periods and whether their use improves the efficiency or reach of CRC screening outreach.

Supporting information

Supplementary material for this article is available at https://doi.org/10.14218/CSP.2026.00016.

Acknowledgments

The authors acknowledge the CDC for providing access to the BRFSS data used in this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no conflicts of interest related to this work.

Author contributions

Conceptualization (PNO, GOO, CPA, JA), methodology (PNO), data analysis (PNO), visualization (PNO), interpretation of findings (PNO, GOO, CPA, JA), writing—original draft preparation (PNO), writing—review and editing (GOO, CPA, JA), critical revision of the manuscript (GOO, CPA, JA). All authors have approved the final version and publication of the manuscript.

Ethical statement

This study analyzed publicly available, de-identified data from the 2024 BRFSS. The analysis did not involve interaction with participants or access to identifiable private information.

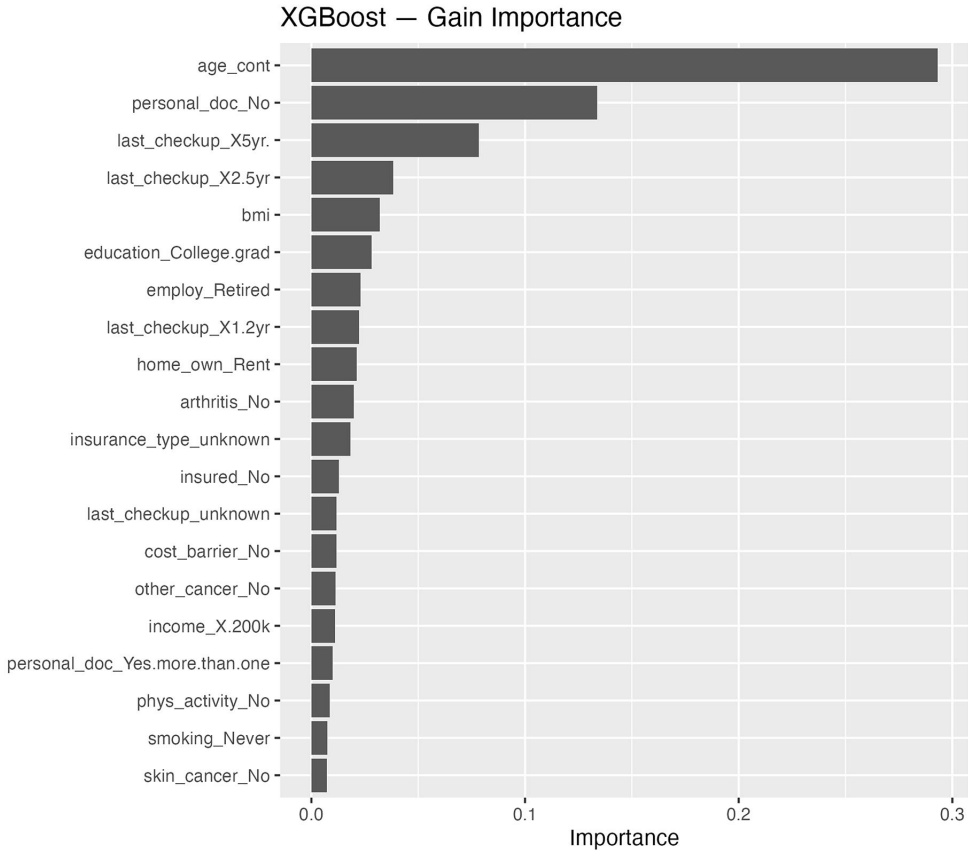


Fig. 6. Variable importance for the XGBoost model classifying current colorectal cancer screening status. Bars represent gain-based variable importance for the 20 highest-ranked predictors, reflecting each predictor's contribution to improvements in tree-splitting performance. Larger values indicate greater predictive contribution. Variable names reflect one-hot encoded categorical predictor levels. BMI, body mass index; XGBoost, extreme gradient boosting.

Therefore, institutional review board approval and additional informed consent were not required for this secondary analysis. The study was conducted in accordance with the principles of the Declaration of Helsinki (as revised in 2024).

Data sharing statement

The data analyzed in this study are publicly available from the CDC's BRFSS. The analytic code used for this study may be made available by the corresponding author upon reasonable request.

References

[1] Siegel RL, Wagle NS, Star J, Kratzer TB, Smith RA, Jemal A. Colorectal cancer statistics, 2026. CA Cancer J Clin 2026;76(2):e70067. DOI: 10.3322/caac.70067, PMID: 41769777.

[2] Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, Davis EM, *et al*. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2021;325(19):1965–1977. DOI: 10.1001/jama.2021.6238, PMID: 34003218.

[3] Centers for Disease Control and Prevention (CDC). Vital signs: colorectal cancer screening test use--United States, 2012. MMWR Morb Mortal Wkly Rep 2013;62(44):881–888. PMID: 24196665.

[4] Beydoun HA, Beydoun MA. Predictors of colorectal cancer screening behaviors among average-risk older adults in the United States. Cancer Causes Control 2008;19(4):339–359. DOI: 10.1007/s10552-007-9100-y, PMID: 18085415.

[5] Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, NJ: Wiley; 2013.

[6] Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018;319(13):1317–1318. DOI: 10.1001/jama.2017.18391, PMID: 29532063.

[7] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380(14):1347–1358. DOI: 10.1056/NEJMra1814259, PMID: 30943338.

[8] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal C, Shen D, Rastogi R, editors. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA, USA. New York: Association for Computing Machinery; 2016:785-794. DOI: 10.1145/2939672.2939785.

[9] Breiman L. Random Forests. Mach Learn 2001;45(1):5–32. DOI: 10.1023/A:1010933404324.

[10] 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. DOI: 10.1016/j.jclinepi.2019.02.004, PMID: 30763612.

[11] Bastani M, Chiuzan C, Silvestri G, Raoof S, Chusid J, Diefenbach M, *et al*. A predictive model for lung cancer screening nonadherence in a community setting health-care network. JNCI Cancer Spectr 2023;7(2):pkad019. DOI: 10.1093/jncics/pkad019, PMID: 37027213.

[12] Yerrapragada G, Siadimas A, Babaeian A, Sharma V, O'Neill TJ. Machine Learning to Predict Tamoxifen Nonadherence Among US Commercially Insured Patients With Metastatic Breast Cancer. JCO Clin Cancer Inform 2021;5:814–825. DOI: 10.1200/CCI.20.00102, PMID: 34383580.

[13] Centers for Disease Control and Prevention (CDC). 2024 BRFSS Survey Data and Documentation. Atlanta: U.S. Department of Health and Human Services; 2025. Available from: https://www.cdc.gov/brfss/annual_data/annual_2024.html. Accessed July 25, 2026.

[14] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45(3):1–67. DOI: 10.18637/jss.v045.i03.

[15] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002;16:321–357. DOI: 10.1613/jair.953.

[16] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–35. DOI: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3, PMID: 15405679.

[17] 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. DOI: 10.1136/bmj-2023-078378, PMID: 38626948.