Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and Temporal Validation of an Explainable Machine Learning Model for Identifying Colorectal Adenoma in Patients With Metabolic Syndrome: A Retrospective Study

Qi Kong
Third Affiliated Hospital of Beijing University of Chinese Medicine
Wenyu Ge
Beijing University of Chinese Medicine
Yilin Fang
Guangyuan Hospital of Traditional Chinese Medicine
Shufei Rao
Third Affiliated Hospital of Beijing University of Chinese Medicine
Bingjie Duan
Third Affiliated Hospital of Beijing University of Chinese Medicine
Jianyun Wang
Third Affiliated Hospital of Beijing University of Chinese Medicine
Zaijian Wang
Third Affiliated Hospital of Beijing University of Chinese Medicine
Aiqing Han
aqhan@hotmail.com

Beijing University of Chinese Medicine
Huixia Li
Third Affiliated Hospital of Beijing University of Chinese Medicine

Research Article

Keywords: Colorectal adenoma, Metabolic syndrome, Machine learning, Clinical prediction model, Explainable artificial intelligence

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10974357/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Colorectal adenoma (CRA) is a precursor to colorectal cancer. Metabolic syndrome (MetS) is associated with an increased likelihood of colorectal neoplasia. We aimed to develop and temporally validate an explainable machine learning model to estimate the probability of CRA in patients with MetS.

Methods

Demographic, clinical, and laboratory data were retrospectively collected from patients with MetS who underwent colonoscopy between January 2024 and June 2026. Based on colonoscopic and histopathological findings, patients were classified into a CRA group and a normal colonoscopy group. Patients from January 2024 to December 2025 were assigned to the training set, whereas those from January to June 2026 formed the temporal test set. All preprocessing, feature selection, and model development were restricted to the training set. Mutual information was used for preliminary screening, followed by a grid-search-optimized GBDT and five-fold cross-validated permutation importance. Three base models were developed: Elastic Net Logistic Regression, XGBoost, and Extra Trees. Calibration and ensemble modeling yielded 11 candidate configurations. The final model and classification threshold were determined using five-fold cross-validation in the training set. Model performance was evaluated using AUC, sensitivity, specificity, Brier score, decision curve analysis, and SHAP.

Results

A total of 774 patients were included: 467 (60.3%) with CRA and 307 (39.7%) with normal colonoscopy findings. Two-stage feature selection identified 10 predictors. Prognostic nutritional index (PNI), hemoglobin, and age had the highest mean permutation importance. ExtraTrees_Isotonic was selected as the final model with a classification threshold of 0.5563. In the temporal test set, this model achieved an AUC of 0.811 (95% CI: 0.740–0.875), a sensitivity of 0.819, a specificity of 0.705, and a Brier score of 0.176. Decision curve analysis showed that the model provided greater net benefit than the treat-all and treat-none strategies across most evaluated thresholds. SHAP analysis showed that sex, history of constipation, and PNI contributed most to model predictions.

Conclusions

The ExtraTrees_Isotonic model showed good discrimination and potential clinical net benefit across a range of threshold probabilities. The model's generalizability and clinical utility require further validation in independent external cohorts and prospective studies.

Background

Colorectal cancer (CRC) remains a major challenge in global cancer prevention and control. According to GLOBOCAN 2022 estimates, CRC is the third most commonly diagnosed cancer and the second leading cause of cancer-related death worldwide, accounting for 9.6% of all new cancer cases and 9.3% of all cancer deaths, respectively [1]. Based on GLOBOCAN 2020 data and global demographic projections, the annual global burden of CRC is projected to reach approximately 3.2 million new cases and 1.6 million deaths by 2040 [2]. Most CRCs arise from colorectal adenomas (CRAs), making the detection and removal of these precursor lesions during screening colonoscopy a key strategy for interrupting the adenoma–carcinoma sequence and preventing CRC [3]. Currently, recommendations regarding the initiation of screening in average-risk individuals remain predominantly age-based, although the recommended starting age differs across regions and guidelines, generally ranging from 40 to 50 years [4–7]. Although age is an established determinant of CRA risk, considerable interindividual variation in CRA risk remains beyond age alone [8]. Age alone may therefore be insufficient to assess individual CRA risk, particularly in clinical populations with distinct risk profiles.

Metabolic syndrome (MetS) is characterized by a cluster of interrelated metabolic abnormalities, including central obesity, elevated blood pressure, elevated fasting glucose, hypertriglyceridemia, and reduced high-density lipoprotein cholesterol (HDL-C) levels [9]. MetS is highly prevalent worldwide, with prevalence estimates ranging from 12.5% to 31.4% depending on the diagnostic criteria used [10]. A more recent global systematic review and modelling analysis estimated that, in 2023, the prevalence of MetS was 31.0% in women and 25.7% in men [11]. Beyond its well-established cardiometabolic consequences, accumulating evidence indicates that MetS has been associated with a higher likelihood of CRA [12–14]. Individual MetS components may also differ in their associations with CRA [13–17]. These findings suggest that the likelihood of CRA varies among individuals with MetS, supporting individualized risk assessment in this population.

Several clinical prediction models have been developed to estimate the likelihood of CRA in general populations and selected high-risk groups [18–22], including patients with type 2 diabetes mellitus [18], metabolic dysfunction-associated steatotic liver disease [19], and chronic liver disease [21]. These models differ substantially in the populations studied, outcomes predicted, and predictors included. Some models rely on specialized biomarkers or information unavailable before colonoscopy. For example, He et al. developed a model using the relative abundances of cross-kingdom gut microorganisms derived from shotgun metagenomic sequencing, whereas Ma et al. incorporated baseline colonoscopic and pathological characteristics, including adenoma size, location, and degree of dysplasia. Such requirements may limit the accessibility of these models and their applicability to routine pre-colonoscopy identification of CRA [23–24]. Tran et al. used metabolic syndrome and its components to develop machine learning and Cox models for incident gastrointestinal cancer [25]. The applicability of existing models to patients with MetS remains uncertain. Few studies have developed CRA prediction models specifically for this population using routinely available pre-colonoscopy data.

To address this gap, we aimed to develop and temporally validate a machine learning model based on routinely available clinical data to estimate the likelihood of CRA in patients with MetS.

Methods

Data Source

This retrospective study included 774 patients with MetS who underwent colonoscopy at the Endoscopy Center, Department of Gastroenterology and Hepatology, Third Affiliated Hospital of Beijing University of Chinese Medicine, between January 2024 and June 2026. The study protocol was approved by the Ethics Committee of the Third Affiliated Hospital of Beijing University of Chinese Medicine(Approval No. BZYSY-2025KYKTPJ-78) and was conducted in accordance with the Declaration of Helsinki. Given the retrospective study design, the requirement for informed consent was waived by the Ethics Committee. All data were de-identified before analysis. The study design and workflow are shown in Fig. 1.

Study Population

MetS was defined according to the modified National Cholesterol Education Program Adult Treatment Panel III (NCEP ATP III) criteria [26]. Because waist circumference was not routinely recorded in this retrospective cohort, the obesity component was operationalized using a body mass index (BMI) cutoff of ≥ 25.0 kg/m². This cutoff was adopted from the 2004 Chinese Diabetes Society criteria [27]. The use of BMI as a substitute for waist circumference has also been reported in previous studies [28–31].

MetS was diagnosed when at least three of the following five criteria were met: (a) obesity, defined as BMI ≥ 25.0 kg/m²; (b) elevated blood pressure, defined as systolic blood pressure ≥ 130 mmHg, diastolic blood pressure ≥ 85 mmHg, or current antihypertensive treatment; (c) elevated triglycerides, defined as triglyceride levels ≥ 1.7 mmol/L (150 mg/dL) or relevant treatment for hypertriglyceridemia; (d) reduced HDL-C, defined as HDL-C levels < 1.03 mmol/L in men or < 1.29 mmol/L in women; and (e) elevated fasting plasma glucose, defined as fasting plasma glucose ≥ 6.1 mmol/L (110 mg/dL) or current glucose-lowering treatment.

All participants were adults aged ≥ 18 years with MetS who underwent complete colonoscopy with adequate bowel preparation and had sufficient clinical, colonoscopic, and histopathological data to determine eligibility and classify the study outcome. All detected polypoid lesions were endoscopically resected and examined histopathologically to determine whether they were adenomatous. Based on the histopathological findings, patients were classified into two groups: the CRA group, comprising patients with histologically confirmed colorectal adenomas, and the normal colonoscopy group, comprising patients with no colorectal polyps or neoplastic lesions detected on colonoscopy. Patients were excluded if they had colorectal cancer, inflammatory bowel disease, a history of colorectal surgery, hereditary polyposis syndromes, severe cardiac, hepatic, renal, or neurological dysfunction, other malignancies, hematological diseases, or autoimmune diseases.

Study Variables

Specifically, the following data were collected: (1) demographic and clinical characteristics, including age, sex, height, weight, body mass index (BMI; calculated as weight in kilograms divided by height in meters squared), smoking status, history of alcohol consumption, history of constipation, history of diarrhea, history of hypertension, type 2 diabetes mellitus, dyslipidemia, non-alcoholic fatty liver disease, and a first-degree family history of colorectal cancer or other digestive system cancers. Height and weight were collected solely to calculate BMI and were not included as separate candidate predictors. (2) Laboratory parameters, including white blood cell count, red blood cell count, hemoglobin, platelet count, lymphocyte count, monocyte count, neutrophil count, alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transferase (GGT), alkaline phosphatase (ALP), total bilirubin (TBIL), direct bilirubin (DBIL), total protein, albumin-to-globulin ratio (A/G ratio), total bile acids, fasting plasma glucose, urea, homocysteine, serum creatinine, uric acid, total cholesterol, triglycerides, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, and C-reactive protein. Based on the original clinical and laboratory variables, and considering clinical significance and routine availability, a total of 10 derived features were constructed. Hematologic, inflammatory, immune, and nutritional indices included the neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), lymphocyte-to-monocyte ratio (LMR), systemic immune-inflammation index (SII = neutrophil count × platelet count / lymphocyte count), and prognostic nutritional index (PNI). Serum albumin was derived from total protein (TP) and the albumin-to-globulin ratio (A/G) using the following formula: albumin = TP × A/G / (1 + A/G). PNI was then calculated as albumin (g/L) + 5 × total lymphocyte count (10^9/L). Hepatic enzyme-derived and metabolic indices included the AST-to-ALT ratio (AST/ALT), triglyceride-to-HDL-C ratio (TG/HDL-C), total cholesterol-to-HDL-C ratio (TC/HDL-C), non-HDL cholesterol (non-HDL-C = TC − HDL-C), and triglyceride–glucose index (TyG). The TyG index was calculated as ln[TG (mg/dL) × fasting plasma glucose (mg/dL) / 2]. TG and fasting plasma glucose values were converted from mmol/L to mg/dL before calculation. The 38 original variables and 10 derived indices constituted a total of 48 candidate predictors.

Data Preprocessing and Feature Selection

Before model development, data quality was assessed. Patients with more than 30% missing clinical data were excluded. For variables with sporadic missing data, continuous variables were imputed with the median, while categorical variables were imputed with the mode. The dataset was divided chronologically according to colonoscopy date. Patients who underwent colonoscopy between January 2024 and December 2025 were assigned to the training set, whereas those who underwent colonoscopy between January and June 2026 were assigned to the temporal test set. Continuous variables were Winsorized using the 1st and 99th percentile values calculated within each training fold, followed by Yeo–Johnson transformation and standardization. Binary variables retained their original 0/1 coding. All preprocessing parameters were estimated within the training folds of cross-validation and applied to the corresponding validation folds. ADASYN oversampling was used only in the training folds. No

resampling was performed in the validation folds or the temporal test set, which retained their original class distributions. Upon completion of the model development protocol, the preprocessing pipeline and model were refit on the full training set, then fixed and applied to the temporal test set. Feature selection was conducted on the training set, with a total of 48 candidate features included. First, mutual information (MI) was used for preliminary screening, and features with MI > 0 were retained. GridSearchCV was then used to optimize the hyperparameters of a gradient boosting decision tree (GBDT) model. Permutation importance was calculated using five-fold cross-validation. Features with a mean permutation importance greater than 0 and positive importance in at least 50% of cross-validation folds were retained. No minimum number of retained features was prespecified.

Model Development and Selection

Based on the selected features, three base models were developed: Elastic Net Logistic Regression, Extreme Gradient Boosting (XGBoost), and Extra Trees. These represented a regularized linear model, a sequential gradient-boosting tree model, and a randomized tree ensemble model, respectively. Each model was trained using the preprocessing pipeline described above. Hyperparameters were optimized using Optuna's tree-structured Parzen estimator (TPE) sampler, with cross-validated AUC in the training set as the optimization objective. The uncalibrated version of each base model was retained. Platt (sigmoid) scaling and isotonic regression were then applied separately to each base model, forming six calibrated models. These calibrated models were treated as separate candidate configurations and were not used to construct the ensemble models. The three optimized uncalibrated base classifiers were subsequently used as base learners to construct stacking and soft-voting ensemble models. The stacking model used logistic regression as the meta-learner and combined the out-of-fold predictions of the three base learners. The soft-voting model averaged the predicted probabilities of the three base learners. A total of 11 candidate model configurations were evaluated, including three uncalibrated base models, six calibrated models, and two ensemble models. Model selection and classification thresholds were based on out-of-fold predictions from stratified five-fold cross-validation in the training set. Both were locked before evaluation in the temporal test set. For each candidate model, the classification threshold was set to the predicted probability cutoff that yielded the maximum Youden index. The primary goal was to identify as many patients with CRA as possible before colonoscopy while maintaining the ability to distinguish individuals with normal colonoscopy findings. Models were first ranked by sensitivity. Ties were resolved by AUC and then specificity.

Model Evaluation and Interpretation

The final model was evaluated in the temporal test set. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC), with 95% confidence intervals (CIs) estimated using 1,000 bootstrap resamples. Sensitivity and specificity were reported at the classification threshold locked during training. The Brier score and calibration curve were used to assess overall prediction accuracy and calibration. Balanced accuracy, F1 score, Matthews correlation coefficient (MCC), positive

predictive value, negative predictive value, and the confusion matrix were reported as additional performance measures. AUCs were compared using the DeLong test, with Holm adjustment for multiple comparisons. Decision curve analysis (DCA) was employed to evaluate the clinical net benefit of the model across different threshold probabilities. SHAP (SHapley Additive exPlanations) analysis was performed on the temporal test set to quantify the contribution of each feature to model predictions. Global feature importance plots, SHAP summary plots, and dependence plots were generated for visualization. SHAP results were used only to interpret model predictions and not to infer causality.

Statistical Analysis

Data analysis was performed using SPSS 27.0 and Python 3.11 software. Continuous variables were assessed for normality using the Shapiro-Wilk test. Normally distributed continuous variables were presented as mean ± standard deviation and were compared using the independent-samples *t* test. For variables with non-normal distributions in either group, data were presented as median and interquartile range and were compared using the Mann–Whitney *U* test. Categorical variables were presented as number (percentage) and were compared using the chi-square test or Fisher's exact test. All tests were two-sided, and $P < 0.05$ was considered statistically significant.

Results

Study Population and Baseline Characteristics

A total of 774 patients with MetS were included, comprising 467 patients with CRA (60.3%) and 307 with normal colonoscopy findings (39.7%). Compared with the normal colonoscopy group, patients in the CRA group were older and had a higher BMI. Male sex, history of constipation and diarrhea, type 2 diabetes mellitus, dyslipidemia, smoking, and a history of alcohol consumption were more common in the CRA group (all $P < 0.05$). Regarding laboratory parameters, the CRA group had higher levels of WBC, RBC, hemoglobin, neutrophils, monocytes, ALT, GGT, total bilirubin, direct bilirubin, A/G ratio, serum creatinine, homocysteine, and PNI, whereas the AST/ALT ratio was lower (all $P < 0.05$). The complete baseline characteristics of the study population are presented in Table 1. Baseline characteristics of the training and temporal test sets are provided in Supplementary Table S1 in Additional file 1.

Table 1
Baseline characteristics of the study population

Variable	Normal colonoscopy group (n = 307)	CRA group (n = 467)	Statistic (Z/χ^2)	P value
Demographic and anthropometric characteristics				
Age (years)	59 (52–65)	61 (55–68)	−3.505	< 0.001
Sex, male	117 (38.1)	281 (60.2)	36.088	< 0.001
Female	190 (61.9)	186 (39.8)		
BMI (kg/m²)	25.20 (22.66–27.34)	25.69 (23.66–27.60)	−2.326	0.020
Hematologic parameters				
WBC (10^9/L)	5.54 (4.77–6.71)	5.92 (5.01–7.30)	−3.035	0.002
RBC (10^{12}/L)	4.48 (4.13–4.83)	4.69 (4.33–5.01)	−4.843	< 0.001
Hemoglobin (g/L)	136 (126–147)	144 (133–155)	−6.449	< 0.001
Platelets (10^9/L)	218 (178–260)	213 (181–256)	−0.010	0.992
Neutrophils (10^9/L)	3.26 (2.60–4.21)	3.54 (2.86–4.50)	−3.071	0.002
Lymphocytes (10^9/L)	1.79 (1.37–2.24)	1.84 (1.42–2.28)	−1.222	0.222
Monocytes (10^9/L)	0.32 (0.26–0.41)	0.35 (0.28–0.44)	−2.654	0.008
Liver-related parameters				
ALT (U/L)	20.90 (15.20–30.00)	22.90 (16.70–33.40)	−2.332	0.020
AST (U/L)	20.00 (16.80–24.00)	20.30 (17.20–24.60)	−1.330	0.183
ALP (U/L)	75 (62–91)	76 (63–92)	−0.387	0.699
GGT (U/L)	21.40 (15.00–33.50)	24.80 (17.00–38.70)	−3.121	0.002

Variable	Normal colonoscopy group (n = 307)	CRA group (n = 467)	Statistic (Z/χ^2)	P value
Demographic and anthropometric characteristics				
Total bilirubin (μmol/L)	11.50 (8.80–15.30)	12.20 (9.30–17.10)	−2.325	0.020
Direct bilirubin (μmol/L)	2.87 (2.20–3.95)	3.05 (2.23–4.32)	−1.965	0.049
Total protein (g/L)	69.50 (65.90–73.40)	70.40 (66.30–73.80)	−1.567	0.117
Albumin-to-globulin ratio	1.68 (1.50–1.83)	1.70 (1.50–2.00)	−2.494	0.013
Total bile acids (μmol/L)	3.10 (2.10–5.00)	3.10 (2.10–4.90)	−0.193	0.847
Renal and metabolic parameters				
Urea (mmol/L)	5.30 (4.50–6.20)	5.40 (4.50–6.30)	−0.266	0.791
Serum creatinine (μmol/L)	63.00 (52.50–72.20)	68.00 (56.00–78.10)	−4.230	< 0.001
Uric acid (μmol/L)	326 (267–403)	332 (271–399)	−0.527	0.598
Fasting plasma glucose (mmol/L)	5.52 (4.91–6.33)	5.56 (5.05–6.64)	−1.025	0.305
Total cholesterol (mmol/L)	4.98 (4.12–5.77)	4.75 (4.13–5.62)	−1.623	0.105
Triglycerides (mmol/L)	1.50 (0.97–2.35)	1.53 (1.02–2.22)	−0.468	0.640
HDL-C (mmol/L)	1.30 (1.06–1.57)	1.27 (1.06–1.51)	−0.706	0.480
LDL-C (mmol/L)	2.92 (2.19–3.53)	2.68 (2.13–3.33)	−1.936	0.053
Homocysteine (μmol/L)	11.50 (10.00–14.40)	13.00 (10.90–15.80)	−4.842	< 0.001
C-reactive protein (mg/L)	1.30 (0.60–2.70)	1.50 (0.80–2.70)	−1.876	0.061
Derived indices				
NLR	1.87 (1.40–2.53)	1.94 (1.46–2.53)	−1.152	0.249

Variable	Normal colonoscopy group (n = 307)	CRA group (n = 467)	Statistic (Z/χ^2)	P value
Demographic and anthropometric characteristics				
PLR	118.45 (97.96–156.00)	116.57 (94.00–151.85)	−0.658	0.510
LMR	5.51 (4.23–6.88)	5.25 (4.17–6.53)	−1.463	0.143
SII	394.86 (274.13–574.30)	408.59 (305.86–592.98)	−1.618	0.106
PNI	52.47 (49.49–55.15)	53.65 (50.28–58.58)	−4.663	< 0.001
AST/ALT ratio	0.94 (0.71–1.23)	0.89 (0.70–1.13)	−2.226	0.026
TG/HDL-C ratio	1.12 (0.66–2.11)	1.21 (0.74–2.00)	−0.775	0.438
TC/HDL-C ratio	3.91 (2.99–4.72)	3.77 (2.96–4.70)	−0.912	0.362
Non-HDL-C (mmol/L)	3.69 (2.78–4.38)	3.47 (2.74–4.33)	−1.655	0.098
TyG index	8.87 (8.35–9.33)	8.89 (8.44–9.30)	−0.646	0.518
Categorical variables				
History of constipation	13 (4.2)	65 (13.9)	19.170	< 0.001
History of diarrhea	17 (5.5)	74 (15.8)	18.972	< 0.001
Type 2 diabetes mellitus	71 (23.1)	144 (30.8)	5.486	0.019
Dyslipidemia	206 (67.1)	352 (75.4)	6.303	0.012
Non-alcoholic fatty liver disease	68 (22.1)	112 (24.0)	0.349	0.555
Hypertension	244 (79.5)	353 (75.6)	1.599	0.207
Smoking	50 (16.3)	150 (32.1)	24.232	< 0.001
History of alcohol consumption	48 (15.6)	143 (30.6)	22.380	< 0.001

Variable	Normal colonoscopy group (n = 307)	CRA group (n = 467)	Statistic (Z/χ^2)	P value
Demographic and anthropometric characteristics				
First-degree family history of colorectal or other digestive cancers	20 (6.5)	49 (10.5)	3.610	0.057

Feature Selection Results

After mutual information prescreening and GBDT-based five-fold cross-validated permutation importance analysis, 10 of the 48 candidate features were retained. PNI, hemoglobin, and age had the highest mean permutation importance values, at 0.0779, 0.0423, and 0.0239, respectively. The remaining selected features were history of constipation, history of diarrhea, sex, LDL-C, history of alcohol consumption, serum creatinine, and HDL-C. The mean permutation importance of HDL-C was close to zero (0.00036), but it was positive in 60% of the cross-validation folds and therefore met the prespecified selection criteria (Fig. 2; Tables S2 and S3 in Additional file 1).

Model Selection and Performance in the Temporal Test Set

Based on out-of-fold predictions from five-fold stratified cross-validation in the training set, the Youden index was used to determine the classification threshold for each model. ExtraTrees_Isotonic had the highest sensitivity (0.8016), with an AUC of 0.752 (95% CI, 0.712–0.790) and a specificity of 0.6016. According to the prespecified selection rule, ExtraTrees_Isotonic and its threshold of 0.5563 were locked as the final model (Table 2).

Table 2
Ranking of candidate models based on out-of-fold predictions from five-fold cross-validation in the training set

Model	Threshold	AUC (95% CI)	Sensitivity	Specificity	Balanced accuracy
ExtraTrees_Isotonic	0.5563	0.752 (0.712–0.790)	0.8016	0.6016	0.7016
Stacking	0.4923	0.743 (0.701–0.784)	0.7989	0.6016	0.7003
ElasticNet_LR_Isotonic	0.4966	0.717 (0.675–0.758)	0.7882	0.5691	0.6787
Voting	0.5066	0.747 (0.705–0.786)	0.7668	0.6220	0.6944
ExtraTrees_Platt	0.5459	0.750 (0.708–0.789)	0.7587	0.6463	0.7025
ExtraTrees	0.4586	0.748 (0.707–0.787)	0.7587	0.6301	0.6944
XGBoost_Isotonic	0.5863	0.717 (0.675–0.757)	0.6997	0.6179	0.6588
ElasticNet_LR_Platt	0.5804	0.723 (0.682–0.764)	0.6890	0.6748	0.6819
XGBoost_Platt	0.5699	0.730 (0.689–0.769)	0.6783	0.6626	0.6704
XGBoost	0.4567	0.723 (0.684–0.763)	0.6649	0.6463	0.6556
ElasticNet_LR	0.5473	0.724 (0.682–0.765)	0.5737	0.7846	0.6791

The temporal test set included 155 patients: 94 with CRA and 61 with normal colonoscopy findings. The final ExtraTrees_Isotonic model achieved an AUC of 0.811 (95% CI, 0.740–0.875), a sensitivity of 0.819, a specificity of 0.705, and a Brier score of 0.176. Core performance results for the six top-ranked candidate models are shown in Table 3. Full performance metrics are provided in Table S4, and the multi-metric comparison is shown in Figure S2 in Additional file 1.

Table 3
Performance of the six top-ranked candidate model configurations in the temporal test set

Model	Threshold	AUC (95% CI)	Sensitivity	Specificity	Brier score
ExtraTrees_Isotonic	0.5563	0.811 (0.740–0.875)	0.8191	0.7049	0.1760
Stacking	0.4923	0.811 (0.736–0.877)	0.8191	0.6393	0.2062
ElasticNet_LR_Isotonic	0.4966	0.796 (0.721–0.864)	0.7979	0.6557	0.1787
Voting	0.5066	0.809 (0.736–0.873)	0.8191	0.6557	0.1809
ExtraTrees_Platt	0.5459	0.808 (0.734–0.873)	0.7872	0.7049	0.1751
ExtraTrees	0.4586	0.817 (0.743–0.883)	0.7979	0.7213	0.1926

The six candidate models had AUCs ranging from 0.796 to 0.817 in the temporal test set (Fig. 3). ExtraTrees had the highest AUC (0.817), while the final ExtraTrees_Isotonic model had an AUC of 0.811. The ROC curves showed substantial overlap. Temporal test-set results were used only for performance evaluation and not for model reselection.

AUCs of the six models were compared using the standard DeLong test, with Holm adjustment for multiple comparisons. All raw pairwise P values were greater than 0.05 (P = 0.2957–0.9777), and all Holm-adjusted P values were 1.000 (Figure S1 in Additional file 1). No statistically significant differences in AUC were detected between the models.

As a supplementary analysis, the area under the precision-recall curve (AUPRC) ranged from 0.852 to 0.874 across the six models. The final ExtraTrees_Isotonic model had an AUPRC of 0.871, above the temporal test-set prevalence baseline of 0.61 (Figure S3 in Additional file 1).

At the Youden threshold locked during training, ExtraTrees_Isotonic correctly identified 77 of 94 patients with CRA, with 17 false negatives. Among 61 patients with normal colonoscopy findings, 43 were correctly classified and 18 were false positives. This corresponded to a sensitivity of 0.819 and a specificity of 0.705 (Fig. 4). ExtraTrees had the highest specificity (0.721) but lower sensitivity (0.798), showing the trade-off between sensitivity and specificity across models.

The calibration curve for the final ExtraTrees_Isotonic model is shown in Fig. 5. Predicted probabilities were generally close to the observed proportions in the mid-to-high probability range, while larger deviations were seen at lower probabilities. The Brier score in the temporal test set was 0.176.

Decision curve analysis showed that the final ExtraTrees_Isotonic model had a higher net benefit than the treat-all and treat-none strategies across most of the evaluated threshold probabilities (Fig. 6).

Around the Youden threshold of 0.556 locked during training, the model maintained a positive net benefit and remained above both reference strategies.

SHAP Interpretation Results

SHAP analysis was performed on the base ExtraTrees classifier within the final ExtraTrees_Isotonic model using the temporal test set. The analysis assessed the contribution and direction of each feature to model output. Ranked by mean absolute SHAP value, sex (male), history of constipation, and PNI were the top three features. They were followed by hemoglobin, history of diarrhea, age, history of alcohol consumption, serum creatinine, LDL-C, and HDL-C (Fig. 7).

The SHAP beeswarm plot showed the direction of each feature's contribution to model output (Fig. 8). Male sex was more often associated with positive SHAP values, while female sex was more often associated with negative values. History of constipation, history of diarrhea, and history of alcohol consumption were also mainly associated with positive SHAP values. Higher PNI, age, and hemoglobin values generally corresponded to higher SHAP values. SHAP values for LDL-C and HDL-C were mostly concentrated near zero, indicating smaller overall contributions.

SHAP dependence plots were generated for the three most important features (Fig. 9). Male sex and history of constipation were generally associated with positive SHAP values, consistent with the beeswarm plot. PNI showed a nonlinear relationship with SHAP values. At lower and intermediate input values, SHAP values were mostly near zero or negative. As PNI increased, SHAP values became positive and increased. Point colors indicate the values of a second feature; statistical interactions were not formally tested.

Overall, sex, history of constipation, PNI, and hemoglobin made relatively large contributions to predictions from the base ExtraTrees classifier, with a nonlinear contribution pattern for PNI.

Discussion

In this study, we developed a model to identify CRA in patients with MetS using routine clinical and laboratory data available before colonoscopy. Ten features were retained from 48 candidate predictors. ExtraTrees_Isotonic was selected as the final model according to the prespecified selection rule. In the temporal test set, the model achieved an AUC of 0.811, a sensitivity of 0.819, a specificity of 0.705, and a Brier score of 0.176. DCA showed greater net benefit than the treat-all and treat-none strategies across most evaluated threshold probabilities. SHAP analysis revealed sex, history of constipation, and PNI as the three largest contributors to model predictions. These results indicate satisfactory discrimination and relatively high sensitivity in the study population.

Previous meta-analyses have shown that MetS is associated with an increased risk of CRA [13], but few studies have examined individualized CRA prediction for patients with MetS. Similar studies have been conducted in other specific high-risk populations. Liu et al. established a model via automated machine learning (AutoML) among 506 patients with MASLD. Their best GBM model achieved an AUC of 0.791 in

the validation set [19]. Mu et al. developed a logistic regression model in 521 patients with T2DM, with an AUC of 0.878 [18]. Our study focused on a different question: among patients who already have MetS, which individuals are more likely to have CRA? The 10 features retained in the final model are not a simple combination of MetS diagnostic components. Only HDL-C belongs to the MetS diagnostic components, and the other features include LDL-C, age, sex, history of constipation and diarrhea, PNI, hemoglobin, history of alcohol consumption and serum creatinine. BMI, fasting plasma glucose, and triglycerides were not retained in the final model. This suggests that differences in demographic characteristics, gastrointestinal symptoms, nutritional and hematologic status, and lifestyle may provide additional information for distinguishing CRA from normal colonoscopy findings in patients with MetS.

SHAP analysis showed that sex, history of constipation, and PNI made the largest contributions to model predictions, followed by hemoglobin, history of diarrhea, age and history of alcohol consumption. A 2025 meta-analysis of 31 population-based studies found that male sex and older age were associated with higher adenoma detection rates [32]. In 2026, a screening study further revealed that after adjusting for age and sex, the association between metabolic syndrome (MetS) and advanced colorectal lesions was significantly attenuated, indicating that age and sex may partially explain the observed associations [33]. In our study, male sex was more often associated with positive SHAP values, while higher age also tended to correspond to higher SHAP values. This suggests that even in a population with pre-existing MetS, age and sex still provide discriminative information. History of alcohol consumption was also mainly associated with positive SHAP values. Previous meta-analyses have generally supported a link between higher alcohol intake and an elevated risk of colorectal adenoma (CRA) [34-35], whereas a recent analysis of the PLCO cohort found no significant association between higher lifetime alcohol intake and incident CRA [36].

Among gastrointestinal symptoms, history of constipation had the second largest SHAP contribution. History of diarrhea was also retained in the final model. Previous studies on constipation, diarrhea, and colorectal neoplastic lesions have produced inconsistent findings. Current evidence is insufficient to establish a consistent association between these symptoms and colorectal adenoma. A 2024 study involving 50,578 colonoscopy procedures found that constipation was not significantly associated with polyp detection in multivariable analysis [37]. In contrast, Guérin et al. reported a higher incidence of benign colorectal neoplasms among patients with chronic constipation in a large database study [38]. For diarrhea, Li et al. found that chronic diarrhea was associated with colorectal adenomatous polyps in a prospective study of 3,576 participants. Chronic diarrhea was also retained as an independent predictor in the multivariable model [39]. Current evidence is limited by differences in study populations, symptom definitions, and outcome measures. Therefore, neither history of constipation nor history of diarrhea can yet be considered an independent risk factor for CRA.

Among laboratory indicators, the contributions of PNI and hemoglobin are relatively prominent. PNI ranked first in permutation importance and third in SHAP analysis. Its contribution showed a nonlinear pattern. At lower and intermediate PNI values, SHAP values were mostly close to zero or negative. At

higher PNI values, SHAP values became positive and increased markedly. PNI is derived from serum albumin and peripheral blood lymphocyte count and reflects nutritional and immune status [40]. MetS is characterized by chronic low-grade inflammation, insulin resistance, and immunometabolic dysfunction [41]. In patients with MetS, PNI may capture nutritional and immune-inflammatory status not reflected by the conventional diagnostic components of MetS. A previous study found lower PNI in patients with recurrent CRA after resection, although this association was not significant after multivariable adjustment [42]. Another study involving 1,330 subjects found that after adjusting for factors such as diabetes, cardiovascular disease and BMI, PNI was still negatively associated with advanced colorectal neoplasia (OR=0.958, 95% CI: 0.930–0.987, P=0.005) [43]. This discrepancy may partly reflect differences in study populations and outcome definitions, as well as the nonlinear relationship between PNI and model output observed in our study.

Hemoglobin ranked second in feature selection and also showed a relatively large contribution in the SHAP analysis. A colonoscopy study involving 1,633 asymptomatic Chinese adults found that higher hemoglobin levels are associated with the detection of colorectal polyps [44]. Hemoglobin is also related to several metabolic features of MetS. Previous studies have reported positive correlations between hemoglobin and triglycerides, waist circumference, fasting plasma glucose, insulin, and HOMA-IR. Higher hemoglobin levels have also been associated with incident MetS [45]. Meta-analysis also showed that hemoglobin levels in MetS patients are higher than those in non-MetS populations [46]. These findings suggest that hemoglobin may capture part of the metabolic profile associated with MetS. Evidence directly linking hemoglobin to CRA remains limited.

Based on the research purpose of identifying CRA patients as much as possible before colonoscopy, candidate models were compared according to preset rules, and model selection placed more emphasis on sensitivity performance. For the final model, the classification threshold corresponding to the maximum Youden's index in the out-of-fold (OOF) prediction of the training set was 0.5563. At this threshold, 77 of 94 patients with CRA were correctly identified in the temporal test set, while 17 were missed. Among 61 patients with normal colonoscopy findings, 43 were correctly classified and 18 were classified as CRA. Such classification results reflect the trade-off between a higher CRA recognition rate and false positives. Under this classification threshold, some individuals with normal colonoscopy results may be classified as CRA, and a negative model result cannot completely exclude CRA. The final model had a Brier score of 0.176. The calibration curve showed that predicted probabilities were relatively close to the observed proportions in the mid-to-high probability range, while larger deviations were seen at lower probabilities. This indicates that the calibration is relatively good in the medium-to-high prediction probability range, but caution is warranted when interpreting predictions in the low-probability range [47]. DCA showed that the net benefit of the model was higher than that of the treat-all and treat-none strategies across most of the investigated threshold probability ranges. This finding suggests potential clinical utility in the study population [48]. Overall, the model may assist with pre-colonoscopy identification of CRA, but it cannot be used to rule out CRA or replace colonoscopy.

This study has several limitations. First, this study is a single-center retrospective study. Although the temporal test set was used to evaluate the model performance, both the training set and the temporal test set came from the same hospital. Patient characteristics and the proportion of CRA may differ across institutions or from those in the general MetS population. The current results mainly reflect single-center temporal validation within our study population. Further validation in independent external cohorts is needed. Second, waist circumference was unavailable in this retrospective cohort. We used BMI ≥25.0 kg/m² to define the obesity component of MetS. BMI cannot directly reflect abdominal fat distribution, and Asian populations may have higher metabolic risk even at lower BMI levels [49-50]. Substituting BMI for waist circumference may have affected MetS classification, the composition of the study population, and model performance. Third, the control group included only patients with normal colonoscopy findings. Patients with non-adenomatous polyps were not included. Therefore, it remains unclear whether the model can effectively distinguish CRA from other types of colorectal polyps. Future multicenter prospective studies are needed to externally validate the model, including its discrimination, calibration, classification threshold, and clinical utility. A broader range of colonoscopic outcomes should also be included to develop and evaluate multiclass models that distinguish CRA from normal colonoscopy findings and other types of colorectal polyps.

Conclusions

In conclusion, the ExtraTrees_Isotonic model used routine clinical and laboratory data available before colonoscopy to estimate the likelihood of CRA in patients with MetS. The model showed good discrimination and relatively high sensitivity in temporal validation. The information used by the model is not limited to MetS diagnostic components, but also includes demographic characteristics, gastrointestinal symptoms, nutritional and hematological measures, lifestyle and other aspects. Currently, the model has undergone only single-center temporal validation, and its external applicability, probability calibration and clinical decision-making value still need further verification in independent populations.

Abbreviations

A/G
albumin-to-globulin ratio
ADASYN
Adaptive Synthetic Sampling
ALP
alkaline phosphatase
ALT
alanine aminotransferase
AST
aspartate aminotransferase

AUC
area under the receiver operating characteristic curve
AUPRC
area under the precision-recall curve
AutoML
automated machine learning
BMI
body mass index
CI
confidence interval
CRA
colorectal adenoma
CRC
colorectal cancer
DCA
decision curve analysis
DBIL
direct bilirubin
GBDT
gradient boosting decision tree
GGT
gamma-glutamyl transferase
HDL-C
high-density lipoprotein cholesterol
LDL-C
low-density lipoprotein cholesterol
LMR
lymphocyte-to-monocyte ratio
MASLD
metabolic dysfunction-associated steatotic liver disease
MCC
Matthews correlation coefficient
MetS
metabolic syndrome
MI
mutual information
NCEP ATP III
National Cholesterol Education Program Adult Treatment Panel III
NLR
neutrophil-to-lymphocyte ratio

OOF
out-of-fold
PLR
platelet-to-lymphocyte ratio
PNI
prognostic nutritional index
RBC
red blood cell count
ROC
receiver operating characteristic
SHAP
SHapley Additive exPlanations
SII
systemic immune-inflammation index
TBIL
total bilirubin
TC
total cholesterol
TG
triglycerides
TPE
tree-structured Parzen estimator
TyG
triglyceride-glucose index
WBC
white blood cell count
XGBoost
Extreme Gradient Boosting.

Declarations

Ethics approval and consent to participate

The study protocol was approved by the Ethics Committee of the Third Affiliated Hospital of Beijing University of Chinese Medicine(Approval No. BZYSY-2025KYKTPJ-78) and was conducted in accordance with the Declaration of Helsinki. Given the retrospective study design, the requirement for informed consent was waived by the Ethics Committee. All data were de-identified before analysis.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding authors on reasonable request, subject to institutional and ethical requirements.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Scientific Research Program of Hebei Provincial Administration of Traditional Chinese Medicine (Project No. B2025031).

Authors' contributions

QK contributed to methodology, data curation, and manuscript writing. WG contributed to data curation and visualization. YF contributed to conceptualization, methodology, and supervision. SR and BD contributed to data curation and manuscript writing. JW, ZW, AH, and HL contributed to methodology, supervision, and manuscript writing. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024 May-Jun;74(3):229-263. doi:10.3322/caac.21834. Epub 2024 Apr 4. PMID:38572751.
2. Morgan E, Arnold M, Gini A, Lorenzoni V, Cabasag CJ, Laversanne M, Vignat J, Ferlay J, Murphy N, Bray F. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. Gut. 2023 Feb;72(2):338-344. doi:10.1136/gutjnl-2022-327736. Epub 2022 Sep 8. PMID:36604116.
3. Li M, Cui M, Zhou X, Song Y. Construction and Validation of a Novel Forecasting Nomogram to the Risk of Colorectal Adenomas: Preventing Colorectal Cancer at Its Origin. Cancer Prev Res (Phila). 2024 Dec 3;17(12):557-564. doi:10.1158/1940-6207.CAPR-24-0066. PMID:39285680.
4. US Preventive Services Task Force; Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Krist AH, Kubik M, Li L, Ogedegbe G, Owens DK, Pbert L, Silverstein M, Stevermer J, Tseng CW, Wong JB. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2021 May 18;325(19):1965-1977.

doi:10.1001/jama.2021.6238. Erratum in: JAMA. 2021 Aug 24;326(8):773. doi:10.1001/jama.2021.12404. PMID:34003218.

5. Issaka RB, Chan AT, Gupta S. AGA Clinical Practice Update on Risk Stratification for Colorectal Cancer Screening and Post-Polypectomy Surveillance: Expert Review. Gastroenterology. 2023 Nov;165(5):1280-1291. doi:10.1053/j.gastro.2023.06.033. Epub 2023 Sep 21. PMID:37737817; PMCID:PMC10591903.
6. Expert Group on Early Diagnosis and Treatment of Cancer, Chinese Society of Oncology, Chinese Medical Association. [Expert consensus on the early diagnosis and treatment of colorectal cancer in China (2023 edition)]. Zhonghua Yi Xue Za Zhi. 2023 Dec 26;103(48):3896-3908. Chinese. doi:10.3760/cma.j.cn112137-20230804-00164. PMID:38129166.
7. Saito Y, Oka S, Kawamura T, Shimoda R, Sekiguchi M, Tamai N, Hotta K, Matsuda T, Misawa M, Tanaka S, Iriguchi Y, Nozaki R, Yamamoto H, Yoshida M, Fujimoto K, Inoue H. Colonoscopy screening and surveillance guidelines. Dig Endosc. 2021 May;33(4):486-519. doi:10.1111/den.13972. PMID:33713493.
8. Zhou YL, Deng JW, Liu ZH, Ma XY, Zhu CQ, Xie YH, Zhou CB, Fang JY. Derivation and validation of lifestyle-based and microbiota-based models for colorectal adenoma risk evaluation and self-prediction. BMJ Open Gastroenterol. 2025 Apr 2;12(1):e001597. doi:10.1136/bmjgast-2024-001597. PMID:40175093; PMCID:PMC12001358.
9. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, Fruchart JC, James WP, Loria CM, Smith SC Jr; International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; International Association for the Study of Obesity. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation. 2009 Oct 20;120(16):1640-5. doi:10.1161/CIRCULATIONAHA.109.192644. Epub 2009 Oct 5. PMID:19805654.
10. Noubiap JJ, Nansseu JR, Lontchi-Yimagou E, Nkeck JR, Nyaga UF, Ngouo AT, Tounouga DN, Tianyi FL, Foka AJ, Ndoadoumgue AL, Bigna JJ. Geographic distribution of metabolic syndrome and its components in the general adult population: A meta-analysis of global data from 28 million individuals. Diabetes Res Clin Pract. 2022 Jun;188:109924. doi:10.1016/j.diabres.2022.109924. Epub 2022 May 15. PMID:35584716.
11. Noubiap JJ, Nansseu JR, Nyaga UF, Ndoadoumgue AL, Ngouo AT, Tounouga DN, Tianyi FL, Foka AJ, Lontchi-Yimagou E, Nkeck JR, Bigna JJ. Worldwide trends in metabolic syndrome from 2000 to 2023: a systematic review and modelling analysis. Nat Commun. 2026;17:573. doi:10.1038/s41467-025-67268-5. PMID:41350289; PMCID:PMC12808104.
12. Milano A, Bianco MA, Buri L, Cipolletta L, Grossi E, Rotondano G, Tessari F, Efthymakis K, Neri M. Metabolic syndrome is a risk factor for colorectal adenoma and cancer: a study in a White

population using the harmonized criteria. Ther Adv Gastroenterol. 2019 Sep 3;12:1756284819867839. doi:10.1177/1756284819867839. PMID:31523276; PMCID:PMC6727097.

13. Wu H, Zhang J, Zhou B. Metabolic syndrome and colorectal adenoma risk: A systematic review and meta-analysis. Clin Res Hepatol Gastroenterol. 2021 Sep;45(5):101749. doi:10.1016/j.clinre.2021.101749. Epub 2021 Jun 29. PMID:34214702.
14. Ku MS, Chiu SY, Chien KL, Lee YC, Chen SL, Chen CD. Gender difference in metabolic syndrome and incident colorectal adenoma: A prospective observational study (KCIS No.42). Medicine (Baltimore). 2021 Jun 4;100(22):e26121. doi:10.1097/MD.0000000000026121. PMID:34087861; PMCID:PMC8183717.
15. Molla MD, Wassie MM, Symonds EL, Abebe Z, Winter JM. Effect of Metabolic Factors on the Risk of Colorectal Precancerous Polyps Among Individuals at Above-Average Risk for Colorectal Cancer: A Systematic Review and Meta-Analysis. Obes Rev. 2026 Jun;27(6):e70086. doi:10.1111/obr.70086. Epub 2026 Jan 15. PMID:41540821.
16. Molla MD, Symonds EL, Winter JM, Cock C, Wassie MM. Metabolic factors are associated with a higher risk of conventional adenoma in males during surveillance colonoscopy: findings from a South Australian cohort. Sci Rep. 2026;16:1532. doi:10.1038/s41598-025-31239-z. Erratum in: Sci Rep. 2026 Jun 18;16(1):19014. doi:10.1038/s41598-026-58187-6. PMID:41361413; PMCID:PMC12796162.
17. Nakai K, Watari J, Tozawa K, Tamura A, Hara K, Yamasaki T, Kondo T, Kono T, Tomita T, Ohda Y, Oshima T, Fukui H, Sakurai J, Kim Y, Hayakawa Y, Fujisawa T, Morimoto T, Miwa H. Sex differences in associations among metabolic syndrome, obesity, related biomarkers, and colorectal adenomatous polyp risk in a Japanese population. J Clin Biochem Nutr. 2018 Sep;63(2):154-163. doi:10.3164/jcbn.18-11. Epub 2018 Apr 3. PMID:30279628; PMCID:PMC6160732.
18. Mu S, Li H, Che X, Sun W, Zou L, Wei X. Targeting colorectal adenomatous polyps in type 2 diabetes: development of a risk assessment model. Scand J Gastroenterol. 2025 Sep;60(9):856-864. doi:10.1080/00365521.2025.2544311. Epub 2025 Aug 7. PMID:40773680.
19. Liu L, Zhang H, Zhao Z, Xu C, Qu S. Development of automated machine learning models in predicting colorectal adenoma in metabolic-associated steatotic liver disease. Front Med (Lausanne). 2026 May 28;13:1852441. doi:10.3389/fmed.2026.1852441. PMID:42292193; PMCID:PMC13253390.
20. Huang PD, Li ZL, Lin ZL, Cheng SY, Zhai LY, Wei SF, Xu YF, Zhang BP, Pan HF, Guo SJ, Zhang HY. [Development of a risk prediction model for colorectal adenoma in non-smoking women]. Zhonghua Liu Xing Bing Xue Za Zhi. 2026 Apr 10;47(4):759-766. Chinese. doi:10.3760/cma.j.cn112338-20250728-00533. PMID:42020179.
21. Li YQ, Kuai WT, Chen L, Zeng MH, Tao XM, Han JX, Wang YK, Xu LX, Ge LY, Liu YG, Li S, Xu L, Mi YQ. Predicting colorectal adenomatous polyps in patients with chronic liver disease: A novel nomogram. World J Gastroenterol. 2025 Jan 14;31(2):99082. doi:10.3748/wjg.v31.i2.99082. PMID:39811506; PMCID:PMC11684197.

22. Li DL, Ma LL, Guan ZA, Zhao YX, Jiang C. Establishment and validation of a clinical prediction model for colorectal adenoma risk factors. Oncol Lett. 2025 May 2;30(1):322. doi:10.3892/ol.2025.15068. PMID:40370646; PMCID:PMC12076052.
23. He B, Xiao Z, Zou L, Wei J, Xiang Z, Sang F, Guo X. Unveiling the unique gut microbial signatures in colorectal adenomas: establishment and validation of a cross-kingdom microbiome predictive model. Front Microbiol. 2026 Jun 16;17:1854806. doi:10.3389/fmicb.2026.1854806. PMID:42382346; PMCID:PMC13314950.
24. Ma J, Zhang Q, Su J, Wang Y, Li W, Zhou M, Shi H, Liu K, Wu J. Risk factors and prediction model for post-polypectomy metachronous colorectal adenoma. Int J Colorectal Dis. 2026 Apr 8;41(1):87. doi:10.1007/s00384-026-05126-6. PMID:41951907; PMCID:PMC13180775.
25. Tran TT, Lee J, Gunathilake M, Kim J, Kim SY, Cho H, Kim J. A comparison of machine learning models and Cox proportional hazards models regarding their ability to predict the risk of gastrointestinal cancer based on metabolic syndrome and its components. Front Oncol. 2023 Mar 2;13:1049787. doi:10.3389/fonc.2023.1049787. PMID:36937438; PMCID:PMC10018751.
26. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation. 2002 Dec 17;106(25):3143-421. doi:10.1161/circ.106.25.3143. PMID:12485966.
27. Expert Panel on Metabolic Syndrome of Chinese Diabetes Society. Recommendations on metabolic syndrome of Chinese diabetes society (Chinese). Chin J Diabetes. (2004) 12:156–61.
28. Liu W, Tang X, Cui T, Zhao H, Song G. Development and visualization of a risk prediction model for metabolic syndrome: a longitudinal cohort study based on health check-up data in China. Front Nutr. 2023 Nov 21;10:1286654. doi:10.3389/fnut.2023.1286654. PMID:38075230; PMCID:PMC10702500.
29. Hu Y, Zhang F, Zhang F, Xiang P. Association between metabolic syndrome and 90-day mortality in acute-on-chronic liver failure patients based on propensity score matching and prognostic model construction. BMC Gastroenterol. 2026 May 19;26(1):434. doi:10.1186/s12876-026-04943-x. PMID:42157040; PMCID:PMC13352876.
30. Fang Y, Rao S, Zhuo Y, Lin J, Zhang X, Wan J. Risk prediction model for progression of type 2 diabetic nephropathy with and without metabolic syndrome: a retrospective cohort study. Front Endocrinol (Lausanne). 2025 Jul 30;16:1592180. doi:10.3389/fendo.2025.1592180. PMID:40810070; PMCID:PMC12343258.
31. Tang X, Zhang L, Wu R, Liu J, Cheng Y. Association of metabolic syndrome and prognosis in elderly patients with community-acquired pneumonia. BMC Pulm Med. 2025 Oct 14;25(1):469. doi:10.1186/s12890-025-03914-4. PMID:41088316; PMCID:PMC12522349.
32. Fernandes C, Estevinho M, Marques Cruz M, Frazzoni L, Rodrigues PP, Fuccio L, Dinis-Ribeiro M. Adenoma detection rate by colonoscopy in real-world population-based studies: a systematic review

and meta-analysis. Endoscopy. 2025 Jan;57(1):49-61. doi:10.1055/a-2382-5795. Epub 2024 Sep 3. PMID:39227020.

33. Singhartinger F, Semmler G, Paar V, Lichtenauer M, Völkerer A, Holzinger J, Ausserwinkler M, Flamm M, Aigner E, Datz C, Wernly B. Metabolic syndrome and risk of advanced colorectal adenomas in a screening population: Frequentist and Bayesian analyses. Colorectal Dis. 2026 Mar;28(3):e70424. doi:10.1111/codi.70424. PMID:41857687; PMCID:PMC13003009.
34. Zhu JZ, Wang YM, Zhou QY, Zhu KF, Yu CH, Li YM. Systematic review with meta-analysis: alcohol consumption and the risk of colorectal adenoma. Aliment Pharmacol Ther. 2014 Aug;40(4):325-37. doi:10.1111/apt.12841. Epub 2014 Jun 18. PMID:24943329.
35. Ben Q, Wang L, Liu J, Qian A, Wang Q, Yuan Y. Alcohol drinking and the risk of colorectal adenoma: a dose-response meta-analysis. Eur J Cancer Prev. 2015 Jul;24(4):286-95. doi:10.1097/CEJ.0000000000000077. PMID:25162969.
36. O'Connell CP, Berndt SI, Chudy-Onwugaje K, Kunzmann A, Huang WY, Barry KH, Loftfield E. Association of alcohol intake over the lifetime with colorectal adenoma and colorectal cancer risk in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Cancer. 2026 Feb 1;132(3):e70201. doi:10.1002/cncr.70201. Erratum in: Cancer. 2026 Feb 15;132(4):e70324. doi:10.1002/cncr.70324. PMID:41582658; PMCID:PMC12833583.
37. Abu Baker F, Mari A, Taher R, Nicola D, Gal O, Zeina AR. The Yield of Colonoscopy in the Evaluation of Constipation: An Age-Based Analysis of Outcome. J Clin Med. 2024 May 15;13(10):2910. doi:10.3390/jcm13102910. PMID:38792451; PMCID:PMC11122582.
38. Guérin A, Mody R, Fok B, Lasch KL, Zhou Z, Wu EQ, Zhou W, Talley NJ. Risk of developing colorectal cancer and benign colorectal neoplasm in patients with chronic constipation. Aliment Pharmacol Ther. 2014 Jul;40(1):83-92. doi:10.1111/apt.12789. Epub 2014 May 15. PMID:24832002.
39. Li W, Chen Z, Chen H, Han X, Zhang G, Zhou X. Establish a Novel Model for Predicting the Risk of Colorectal Ademomatous Polyps: a Prospective Cohort Study. J Cancer. 2022 Aug 15;13(10):3103-3112. doi:10.7150/jca.74772. PMID:36046645; PMCID:PMC9414019.
40. Keskinkilic M, Semiz HS, Ataca E, Yavuzsen T. The prognostic value of immune-nutritional status in metastatic colorectal cancer: Prognostic Nutritional Index (PNI). Support Care Cancer. 2024;32(6):374. doi:10.1007/s00520-024-08572-6. PMID:38777931; PMCID:PMC11111560.
41. Cho YH, Lee Y, Choi JI, Lee SR, Lee SY. Biomarkers in metabolic syndrome. Adv Clin Chem. 2022;111:101-156. doi:10.1016/bs.acc.2022.07.003. PMID:36427908.
42. He Q, Du S, Wang X, Liu J, Xu X, Liu W, Zhang J, Jiang K. Development and validation of a nomogram based on neutrophil-to-lymphocyte ratio and fibrinogen-to-lymphocyte ratio for predicting recurrence of colorectal adenoma. J Gastrointest Oncol. 2022 Oct;13(5):2269-2281. doi:10.21037/jgo-22-410. PMID:36388694.
43. Zhang C, Wu S, Lu L, Jin S. Systemic inflammatory biomarkers (NLR, SII, PNI and FPR) combined with CEA for predicting advanced colorectal neoplasms: development and temporal validation of a

machine learning model. Front Oncol. 2026 Jul 22;16:1832546. doi:10.3389/fonc.2026.1832546. PMID:42558186.

44. Shen C, Shi X, Al-Ashoor M, Xu C. Association between hemoglobin levels and colorectal polyps in asymptomatic Chinese adults. Gastroenterol Res Pract. 2020;2020:3808163. doi:10.1155/2020/3808163. PMID:32411196.
45. He S, Gu H, Yang J, Su Q, Li X, Qin L. Hemoglobin concentration is associated with the incidence of metabolic syndrome. BMC Endocr Disord. 2021 Mar 19;21(1):53. doi:10.1186/s12902-021-00719-4. PMID:33740939.
46. Mahmood A, Samad S, Haider H, Hassan W, Mushtaq R, Perwaiz A, Ali A, Azharuddin M. Association of red blood cell and platelet parameters with metabolic syndrome: a systematic review and meta-analysis of 170,000 patients. Horm Metab Res. 2024 Jul;56(7):517-525. doi:10.1055/a-2241-5361. PMID:38195797.
47. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec 16;17(1):230. doi:10.1186/s12916-019-1466-7. PMID:31842878.
48. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-574. doi:10.1177/0272989X06295361. PMID:17099194.
49. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet. 2004 Jan 10;363(9403):157-163. doi:10.1016/S0140-6736(03)15268-3. PMID:14726171.
50. Chen S, Chen Y, Liu X, Li M, Wu B, Li Y, Liang Y, Shao X, Holthöfer H, Zou H. Insulin resistance and metabolic syndrome in normal-weight individuals. Endocrine. 2014 Aug;46(3):496-504. doi:10.1007/s12020-013-0079-8. Epub 2013 Nov 5. PMID:24190050.

Figures

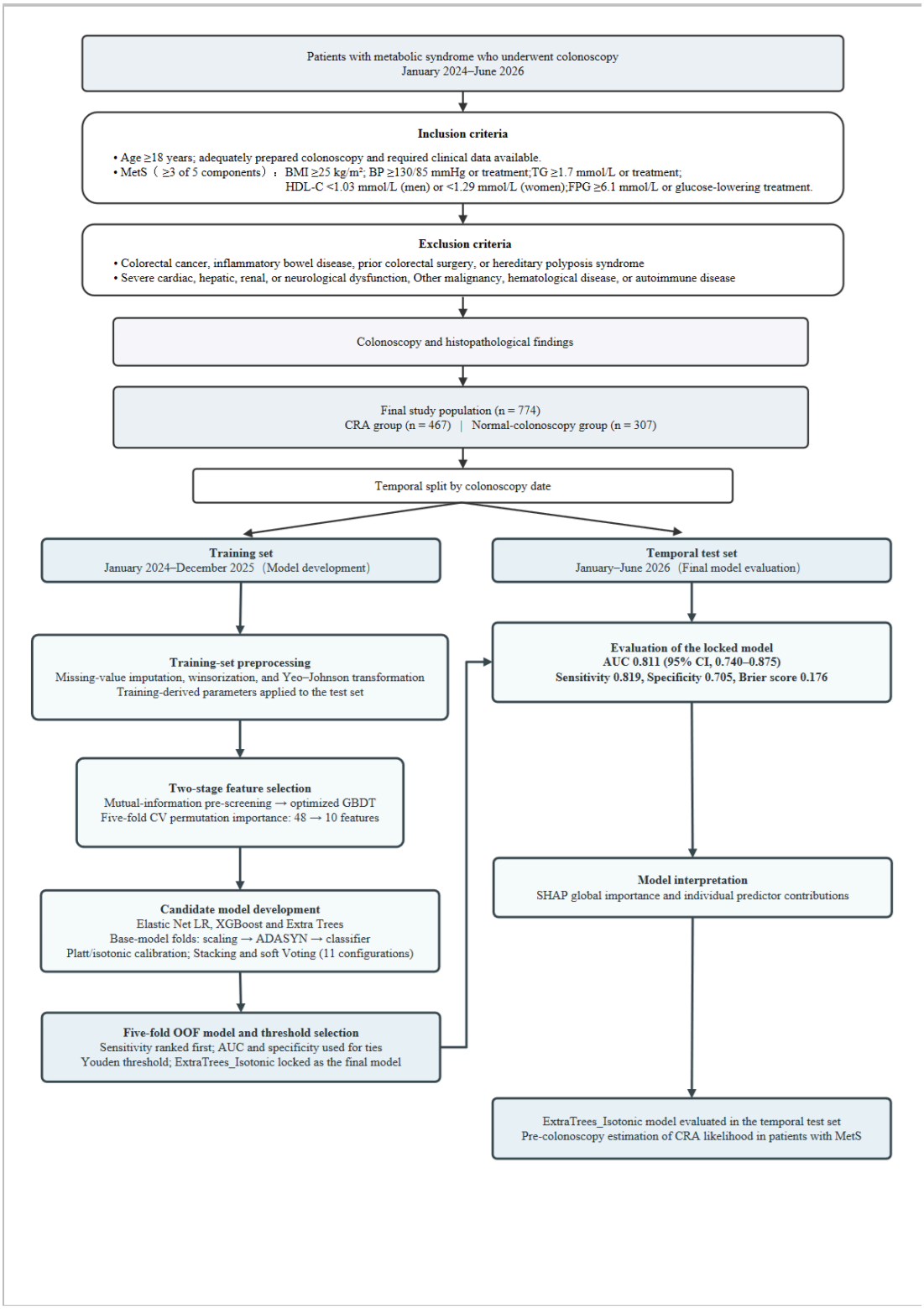


Figure 1

Study design and model development workflow.

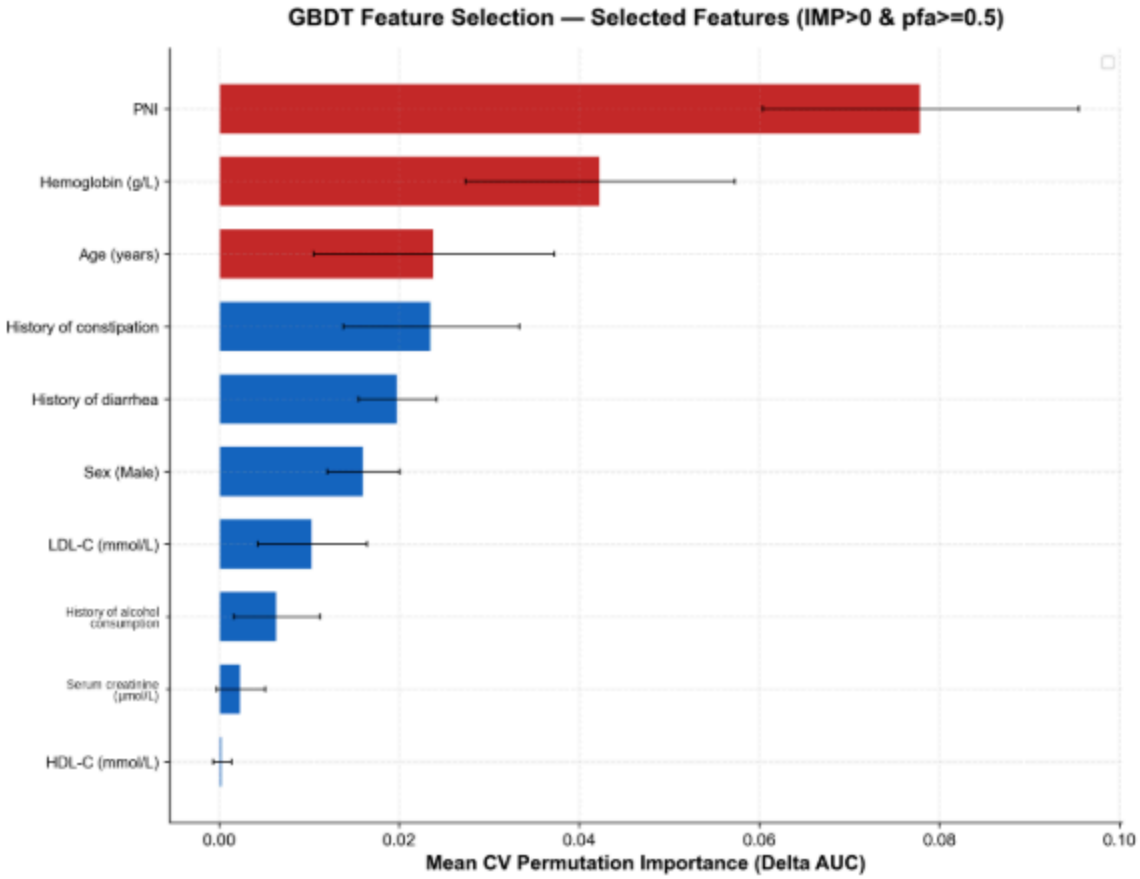


Figure 2

Cross-validated permutation importance of the 10 selected features.

Note: Bars show the mean decrease in AUC after feature permutation, and error bars show variation across the five cross-validation folds. The three features with the highest mean permutation importance are shown in red; the other selected features are shown in blue.

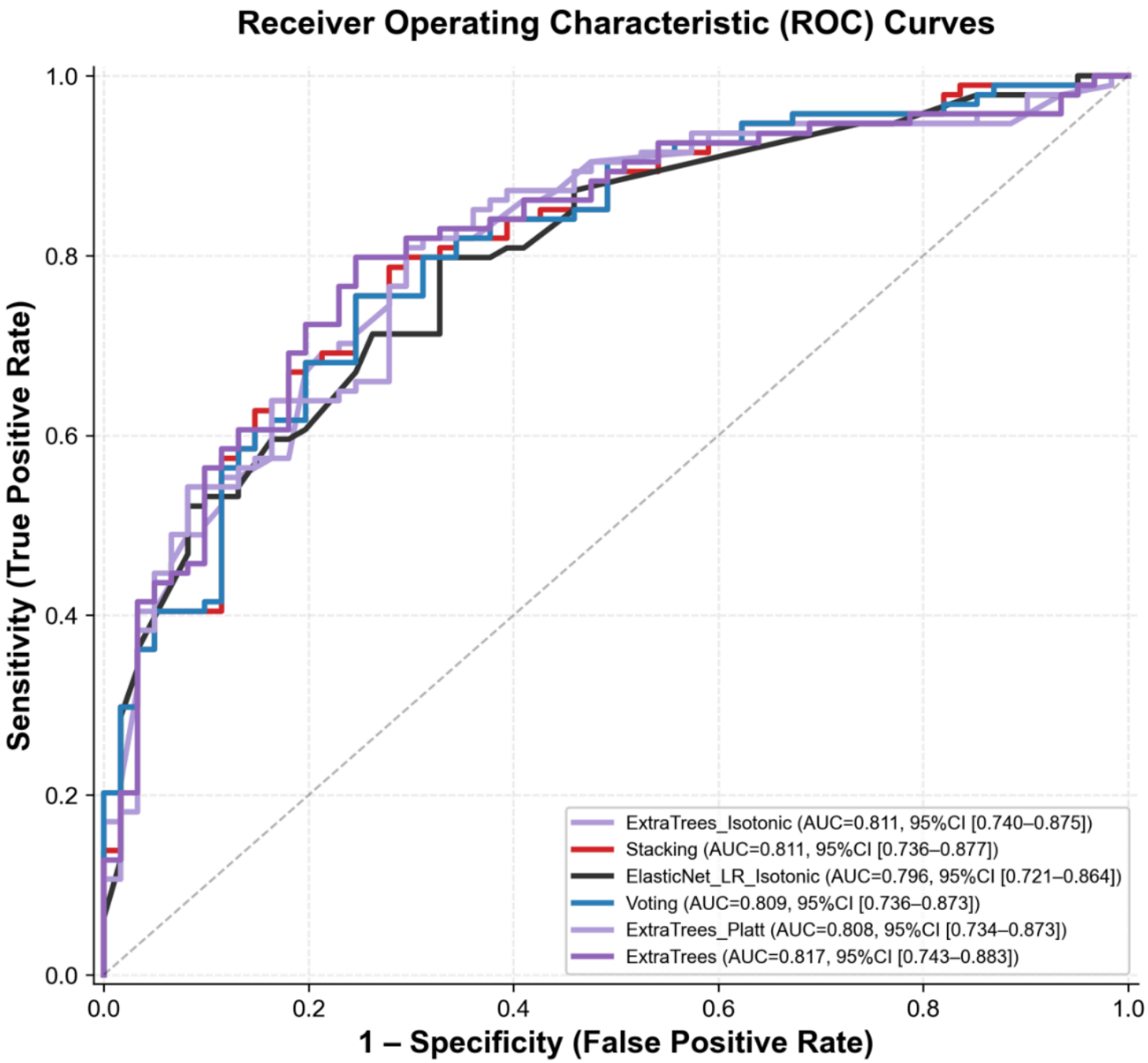


Figure 3

ROC curves of the six top-ranked candidate models in the temporal test set.

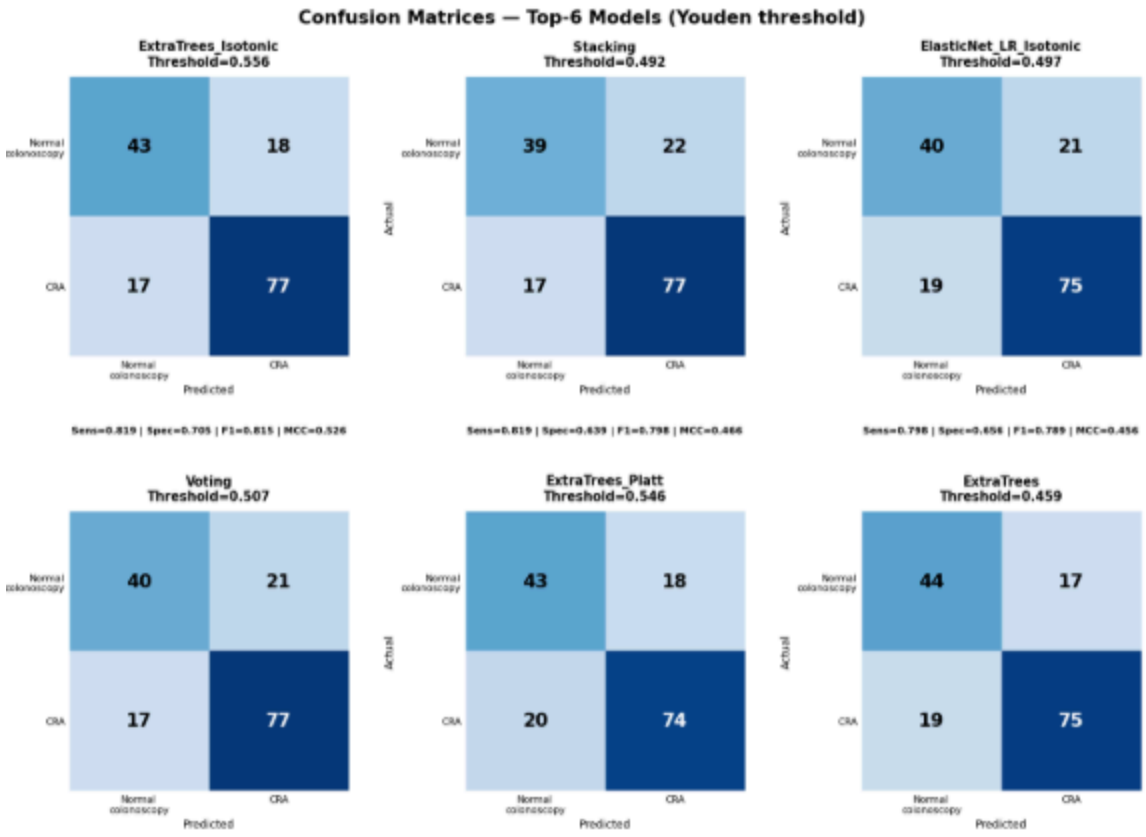


Figure 4

Confusion matrices of the six top-ranked models in the temporal test set.

Note: Confusion matrices were generated using the Youden thresholds locked during training.

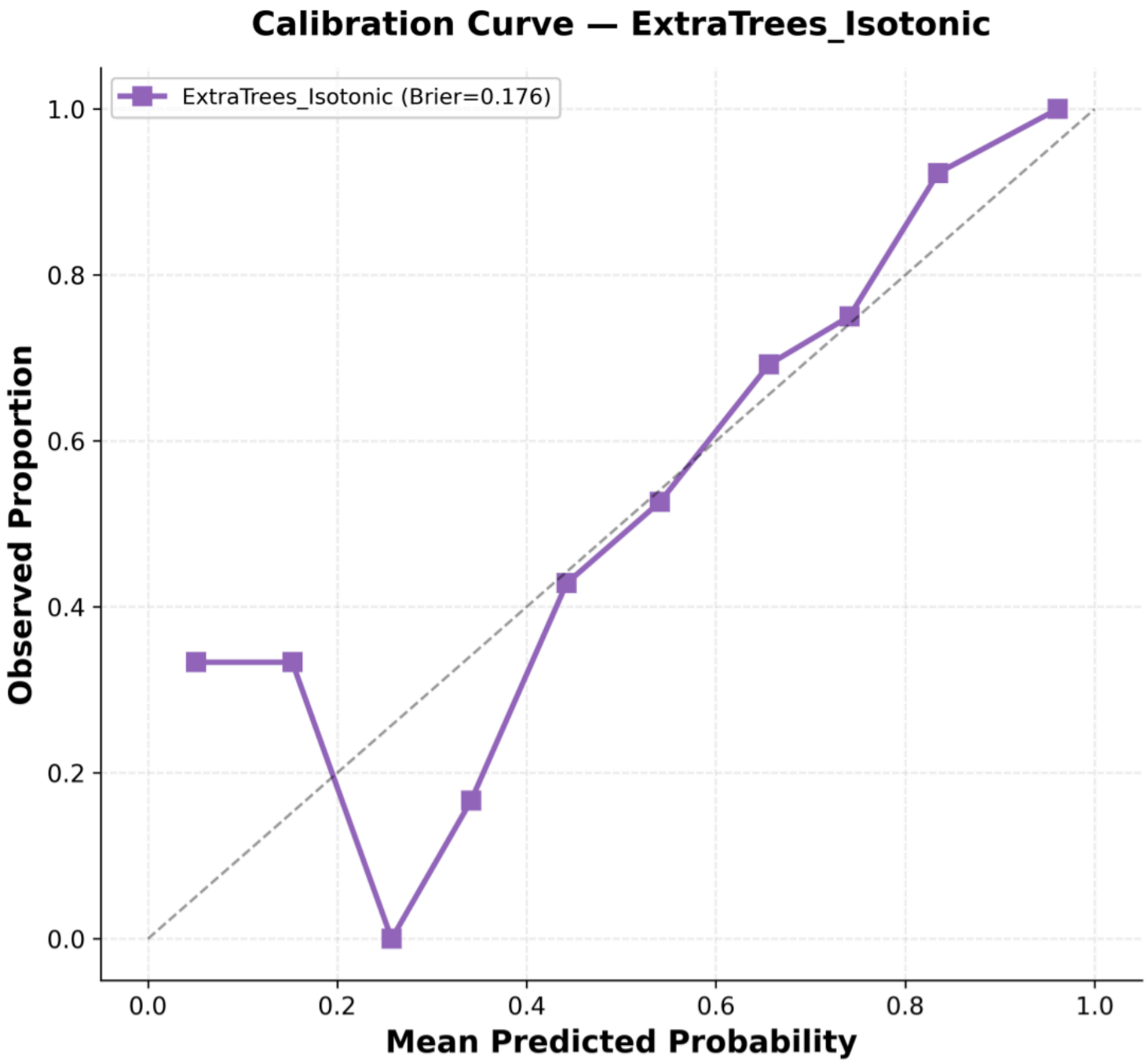


Figure 5

Calibration curve of the final ExtraTrees_Isotonic model in the temporal test set.

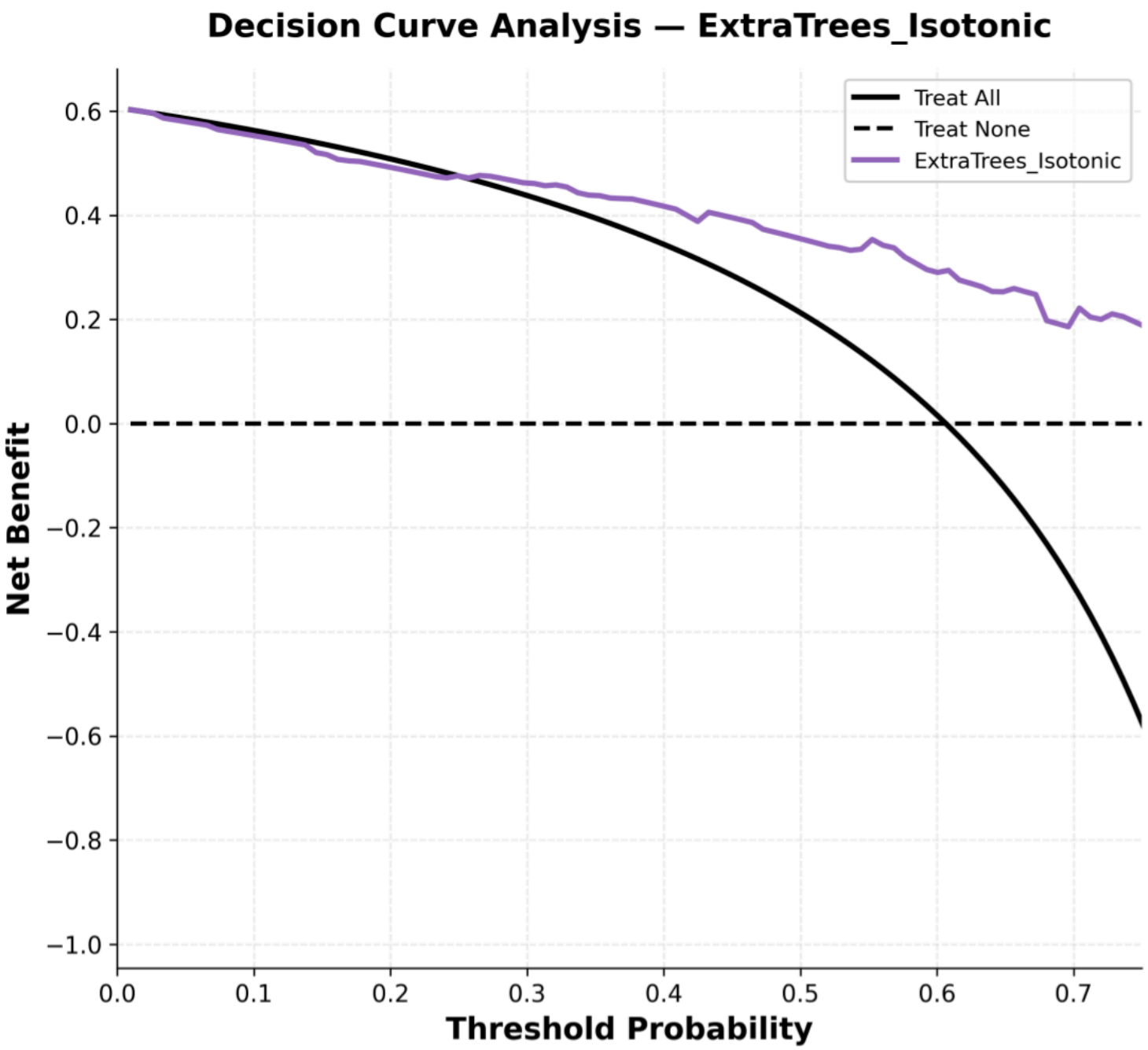


Figure 6

Decision curve analysis of the final ExtraTrees_Isotonic model in the temporal test set.

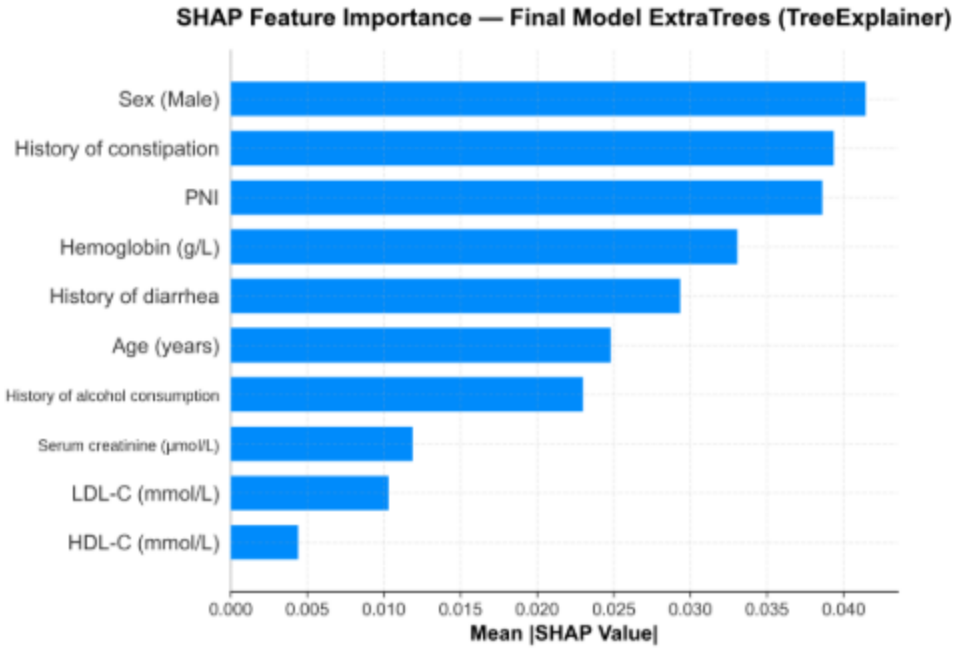


Figure 7

Global SHAP feature importance for the base ExtraTrees classifier.

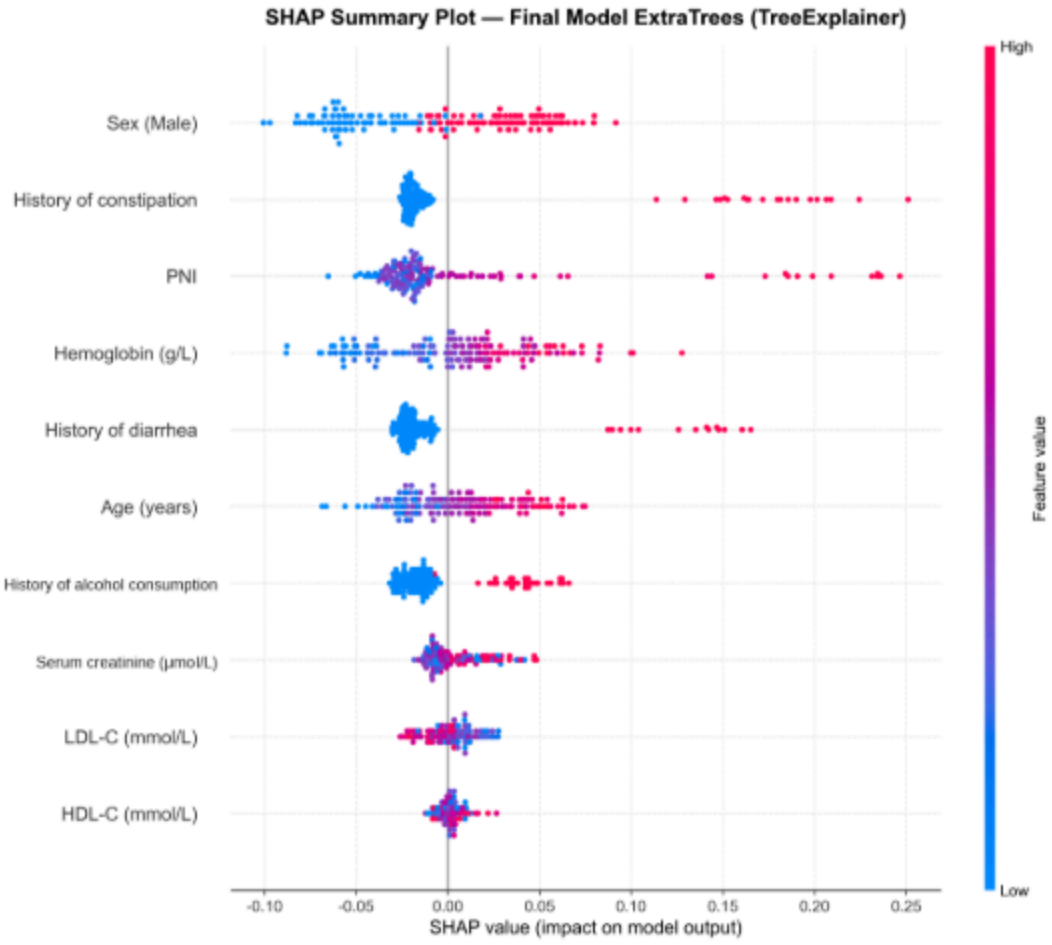


Figure 8

SHAP beeswarm plot for the base ExtraTrees classifier.

Note: Colors indicate preprocessed feature values.

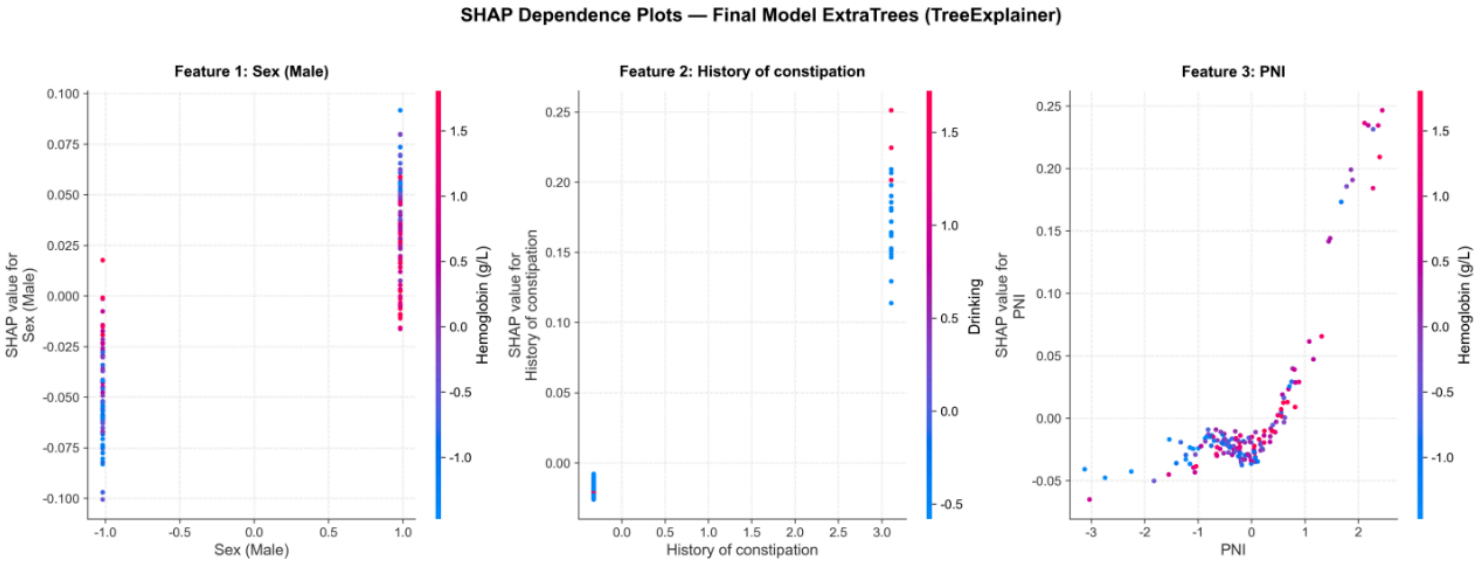


Figure 9

SHAP dependence plots for the three most important features.

Note: The x-axes show preprocessed and standardized model inputs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialsKongetal.docx