Understanding Predictors of Early-Onset Colorectal Cancer in Personal Health Data with Machine Learning

Ashis Das, Melinda Rossi, Michael Broder, Caitlin Sheetz
ADVI Health, Washington, DC

Poster Code CO92
Acceptance Code: 4488
ISPOR 2025
Montreal, QC, Canada
May 13-16, 2025

Background

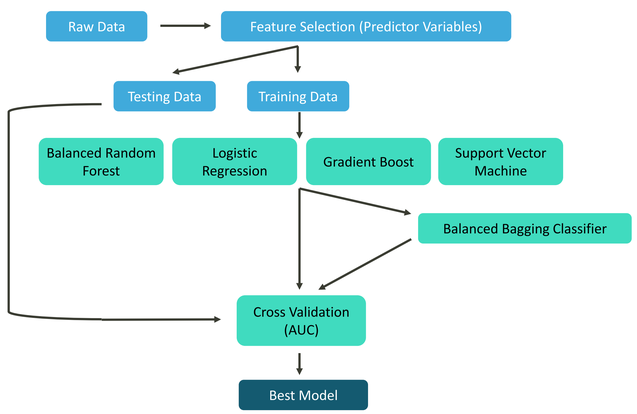
- Colorectal cancer (CRC) is the second leading cause of cancer-related mortality in the United States (US), Although its incidence is stabilizing, the incidence of early-onset CRC (EOCRC, diagnosed younger than 50 years) is increasing.
- EOCRC diagnosis is often delayed, and current screening processes lead to many false positives and are invasive, highlighting a need for a non-invasive and cost-effective method to estimate EOCRC risk.
- Previous studies predicting EOCRC with machine learning (ML) primarily used electronic health record data and were limited to single centers or small geographic areas.
- This study aims to identify key factors that predict the likelihood of EOCRC using ML and self-reported personal health data in a nationally representative US sample.

Methods

- **Data Source**: National Health Interview Survey (NHIS) data from 2019-2023.
- **Analytic Sample**: Adults aged less than 50 years with or without CRC.
- **Binary Outcome**: Whether individuals reported a CRC diagnosis.
- **Predictors**: Self-reported sociodemographic factors (age, sex, ethnicity, urbanicity, household income-to-poverty ratio, smoking) and clinical parameters (body mass index [BMI], hypertension, diabetes, hyperlipidemia).
- **ML Models**: Four ML models, with and without balanced bagging classifier (i.e., balanced random forest, logistic regression, gradient boost, and support vector machine) (Fig. 1). Shapley Additive exPlanations (SHAP) was used for leading predictor identification.
- **Model Tuning**: 70%/30% train/test split, followed by hyperparameter tuning using Python 3. The primary performance metric was area under the receiver operating characteristic curve (ROC-AUC).

Ashis Das
Director, RWE | ADVI Health | ashis.das@advi.com

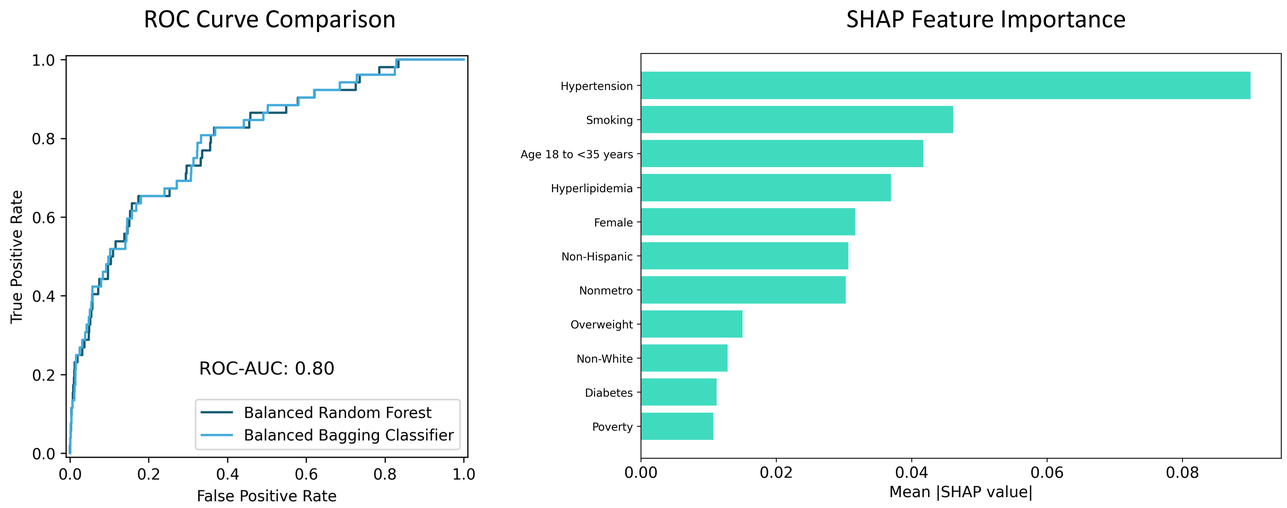
Figure 1. Model Selection and Validation



Results

- The balanced random forest model with balanced bagging classifier achieved the highest performance in terms of discriminative ability in predicting EOCRC, with a ROC-AUC of 0.80. AUC for other models were 0.79 (logistic regression) and 0.77 (support vector machine and gradient boost) (Fig. 2).
- Based on SHAP feature importance using balanced random forest with balanced bagging classifier, top five predictors were hypertension, smoking, age, hyperlipidemia, and sex (Fig. 2).

Figure 2. Best Model Fit and Feature Importance



Conclusions

- This study demonstrates the potential of using ML techniques to predict EOCRC using publicly available data in large US populations.
- Early identification of patients at risk for developing EOCRC could facilitate improved screening methods and address the current challenge of delayed diagnosis and treatment.