Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/166/2025.TJ23598

Machine Learning Prediction Models for Colorectal Cancer Based on the Novel Ensemble Framework

Qing Mi

SHU-UTS SILC Business School, Shanghai University, Shanghai, China
miqing0822@shu.edu.cn

Abstract: Colorectal Cancer (CRC) is a highly prevalent malignancy globally, and early prediction is crucial for improving prognosis. This study used a multidimensional CRC dataset (n=1000) provided by the Kaggle platform, which contains 14 clinical and lifestyle characteristics. First, data imbalance was mitigated through Random Oversampling (ROM) and standardization. Subsequently, a comprehensive evaluation was performed on seven baseline machine learning models, including Gradient Boosting Decision Tree (GBDT), eXtreme Gradient Boosting (XGBoost) and so on. Based on performance metrics such as accuracy and F1 score, GBDT and XGBoost were subsequently selected as the optimal base learners. Finally, the predictive probability features generated by the base learners are fed into the meta-learners such as Random Forest (RF), K Nearest Neighbor (KNN) and Multi-Layer Perceptron (MLP) for secondary modeling. The interpretability of the model is achieved through the Shapley Additive exPlanations (SHAP) value, which quantifies the marginal contribution of each feature to the prediction. Experiments show that the RF integration architecture based on GBDT and XGBoost baseline models has the best performance (accuracy of 0.9527 and AUC of 0.9923). SHAP analysis showed that Activity_Level and BMI were core predictors with limited contribution from gender, confirming the prioritization of exercise and weight management in CRC prevention. The framework demonstrated excellent robustness and maintained its predictive advantage even when inefficient base models e.g., Logistic Regression (LR) were introduced. This study provides an interpretable machine learning paradigm for CRC risk stratification with potential for clinical translation.

Keywords: Ensemble machine learning, colorectal cancer, SHAP interpretability

1. Introduction

Colorectal Cancer (CRC), a malignant tumor that occurs in the colon and rectum, is the third most common cancer worldwide, accounting for approximately 10% of all cancer cases and the second leading cause of cancer-related deaths worldwide. It is projected that by 2040, the number of new cases of CRC will increase to 3.2 million and the number of deaths will reach 1.6 million [1]. Current diagnostic methods for CRC contain imaging, collection of tissue samples (biopsy), and colonoscopy, which are effective but suffer from low patient compliance and low accuracy of non-invasive tests.

DOI: 10.54254/2755-2721/166/2025.TJ23598

In recent years, the development of machine learning technology has provided new solutions for CRC prediction. By analyzing clinical data and imaging features, machine learning models can achieve efficient risk prediction.

Machine learning originated from the perceptual machine algorithm in the mid-20th century, and has formed a rich technical system including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF) and other classical algorithms. Among them, deep learning has gained rapid development in recent years, and with the powerful feature extraction ability of neural network, it has made breakthrough progress in artificial intelligence, automatic driving, medical image analysis and other fields, and landmark innovative algorithms such as Transformer and Diffusion Model have emerged one after another [2], which continue to push the boundaries of artificial intelligence technology. The current application of machine learning in the field of cancer prediction has also made more significant progress. For example, in [3], Fang et al. compare the performance of machine learning models such as RF, Lightweight Gradient Boosting Machine (LightBGM), and Extreme Gradient Boosting (XGBoost) with traditional multiple linear regression methods, demonstrating that LightBGM has higher accuracy in predicting the fear of recurrence among breast cancer patients, highlighting the machine learning's advantages in psychological risk stratification. In [4], Li et al. used multifactorial logistic regression analysis to identify independent influences on survival prognosis after endoscopic treatment in patients with early-stage CRC. They also identified the major risk factors contributing to patient mortality based on the importance analysis of random forest features with Shapley Additive exPlanations (SHAP) value ranking. In addition, deep learning models such as Recurrent Neural Networks (RNN) and Long Short-term Memory networks (LSTM) have shown excellent performance in blood cancer and breast cancer prediction, respectively [5]. Convolutional Neural Networks (CNNs) have been widely used for image classification, lesion segmentation and pathology recognition in CRC due to their powerful feature extraction capabilities [6]. However, current machine learning-based cancer prediction research is mostly limited to the development and application of a single model, and has not yet fully explored the potential of the multi-model integration strategy in improving the prediction performance, which may be an important development direction for future researches.

In order to solve the above limitations, this study used Colorectal Cancer Dietary and Lifestyle Dataset from Kaggle platform to systematically investigate the effects of multidimensional factors such as Body Mass Index (BMI), lifestyle, and ethnicity on the risk of CRC based on SHAP values. Subsequently, eight machine learning algorithms based on DT and K Nearest Neighbor (KNN) were used as the initial prediction models. Based on the evaluation metrics such as Accuracy, Precision, XGBoost, GBDT and Logistic Regression (LR) were selected as baseline models, and the output probabilities of these three models will be used as higher-order features to input models such as RF and Multi-Layer Perceptron (MLP) network for integrated learning to further improve the performance of the models for CRC prediction.

2. Method

2.1. Dataset preparation

The dataset used in this study comes from the Kaggle public platform and contains a total of 1,000 sample data [7]. Each piece of data consists of 14 feature dimensions, and the specific features include base identifier (Participant_ID), demographic features (Age, Gender, Ethnicity), clinical features (BMI, Family_History_CRC, Pre-existing_Conditions), behavioral features (Lifestyle), and nutritional metabolic characteristics (Carbohydrates(g), Proteins(g), Fats(g), Vitamin_A(IU), Vitamin_C(mg), Iron(mg)). The target variable was CRC_Risk, a dichotomous label used to identify

DOI: 10.54254/2755-2721/166/2025.TJ23598

the risk of CRC development (0 for low risk and 1 for high risk), and the sample size of the two categories was counted to be 845 versus 155 entries, respectively. This dataset provides a clinically interpretable data base for CRC risk prediction studies through multidimensional biomedical indicators and dietary factors.

In the data preprocessing stage, the missing values in the categorical variables Pre-existing Conditions were firstly filled in and uniformly labeled as 'None' to clearly identify the samples with no past medical history. For the multicategorical features Pre-existing Conditions and Ethnicity, they were converted to binary dummy variables using the One-Hot Encoding technique to avoid erroneous numerical associations of unordered categorical variables by the model. For the binary categorical variables Gender ("Male"→0, "Female"→1) and Family_History_CRC ("No" → 0, "Yes" → 1), binarized coding is used to preserve their Boolean logic features. In particular, for the ordered categorical variable Lifestyle, the potential progressive effect of behavioral patterns on disease risk was quantified based on the mapping of the gradient of exercise to the numeric variable Activity_Level (Smoker=0, Sedentary=1, Moderate Exercise=2, Active=3). To alleviate the problem of category imbalance of the target variables, the sample of the high-risk group was expanded by using the Random Oversampling (ROM) technique, so that the sample size of both categories reached 845 cases to ensure that each category receives a balanced attention when the model is trained. On this basis, eight continuous variables such as Age and BMI were standardized by StandardScaler, and each feature was converted to a distribution with mean 0 and standard deviation 1, which effectively eliminated the influence of the difference in the scale between different features on the model training. Finally, the preprocessed dataset is divided into training set and test set according to the ratio of 8:2, in which the test set accounts for 20%, and the fixed random seed (random_state=42) is set to ensure the reproducibility of the experiment.

2.2. Benchmark machine learning models

1) RF: RF is a supervised algorithm based on integrated learning that integrates multiple decision trees using Bootstrap Aggregating and reduces model variance through feature random sampling with sample self-sampling [8]. It predicts results by majority voting or mean aggregation and is suitable for classification and regression tasks.

2) DT: DT is a supervised learning algorithm based on tree structure, which constructs classification or regression models by recursively dividing the feature space [9]. It uses criteria such as information gain and Gini coefficient to select the optimal splitting node, which is intuitively interpretable, but has the limitation of easy overfitting.

3) KNN: KNN is an instance-based inert learning algorithm that measures sample proximity by Euclidean distance or cosine similarity and makes predictions based on the majority class labels of the K nearest neighbors [10]. Its computational complexity grows linearly with data size and is sensitive to feature scale.

4) XGBoost: XGBoost is an efficient integrated learning algorithm based on the gradient boosting framework, which iteratively optimizes the decision tree model and introduces regularization terms to improve the generalization performance [11]. The algorithm uses a second-order Taylor expansion to approximate the loss function, supports parallel computation and feature importance evaluation, and achieves a remarkable balance between accuracy and efficiency.

5) GBDT: GBDT is an integrated learning algorithm based on the Boosting framework, which constructs strong prediction models by iteratively training weak classifiers and gradient optimizing the loss function [12]. The algorithm uses an additive model with a forward-stepping strategy to gradually improve the prediction accuracy by minimizing the residuals.

6) MLP: MLP is a feed-forward artificial neural network that achieves complex function approximation by nonlinear transformation of multiple hidden layers [13]. It relies on

DOI: 10.54254/2755-2721/166/2025.TJ23598

backpropagation and stochastic gradient descent to optimize the weights, but needs to suppress overfitting by Dropout or batch normalization.

7) LR: LR is a generalized linear model that maps linear combinations to probabilistic outputs via a Sigmoid function for binary classification tasks [14]. The algorithm uses great likelihood estimation to optimize the parameters, which is computationally efficient and interpretable, but the modeling ability is limited by the linear decision boundary.

2.3. Ensemble machine learning model

As shown in Fig. 1, this study proposes a two-stage integrated learning framework that aims to improve the generalization performance of CRC risk prediction models through feature fusion in probability space. First, the original dataset was fully trained and predicted based on seven baseline machine learning models (RF, DT, KNN, XGBoost, GBDT, MLP, and LR), and 2-3 baseline models that performed better on both the training set and the test set were selected by comprehensively evaluating the classification performance metrics such as the accuracy, precision, and F1-score. Subsequently, the predicted probabilities of the outputs of the preferred baseline models on the training and test sets are used as meta-features, which are inputted into three types of heterogeneous meta-learners, namely RF, KNN and MLP, respectively, for secondary modeling to construct a probability-driven feature enhancement space. In this process, the probabilities output from the baseline models can effectively characterize the uncertainty of the samples near the decision boundary, while the meta-learners explore the complementarity information among the multi-model predictions through the nonlinear combination mechanism. Finally, the performance of the three fusion models on independent test sets is systematically evaluated by comprehensive evaluation metrics such as accuracy, recall, and F1-score, and the integrated architecture with the best overall performance is selected as the final prediction system. The method provides a more robust integrated solution for medical classification tasks by co-optimizing the two-stage models, which retains the local discriminative advantages of the base model and reduces the overfitting risk of a single model through probabilistic feature reconstruction.

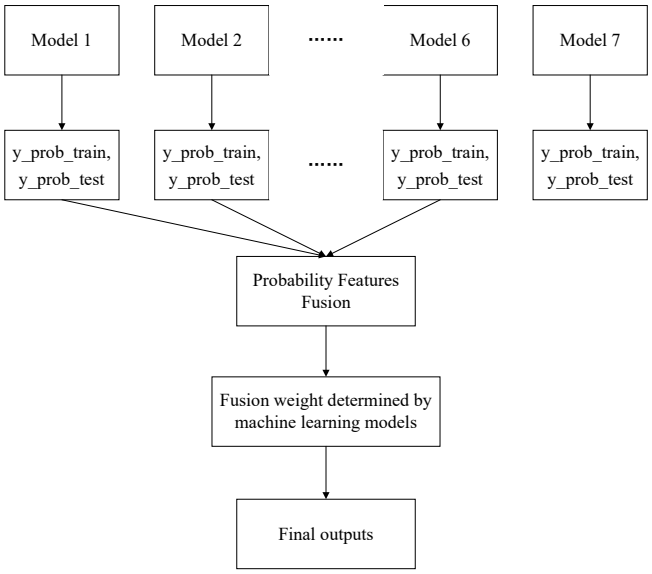


Figure 1: Flowchart of probability-based feature fusion framework

DOI: 10.54254/2755-2721/166/2025.TJ23598

2.4. Evaluation metrics

Models' performance was assessed using five core metrics to comprehensively measure classification effectiveness. Accuracy reflects the proportion of correct predictions of the model as a whole, but is sensitive to category imbalance data; Precision focuses on the reliability of the positive category predictions, reflecting the ability to control false positives; Recall focuses on the completeness of the positive category samples, reflecting the level of false-negative avoidance; F1 Score evaluates the model ranking effectiveness from a threshold-independent point of view by summing the average of precision and recall, which is suitable for scenarios that require balanced checking and accuracy. And the area under the ROC curve (AUC) evaluates the model sorting ability from a threshold-independent perspective, and the closer its value is to 1, the better the model's performance in distinguishing between positive and negative classes. These five metrics systematically assess the prediction quality, robustness and generalization ability of the classifier from different dimensions.

2.5. SHAP feature importance

Shapley Additive exPlanations is a feature attribution method based on game-theoretic Shapley values for explaining the predictive behavior of machine learning models. The core idea is to quantify the impact of features on the models' outputs by calculating their average marginal contributions across all possible subset combinations, thus satisfying the mathematical rigor required for interpretability. Compared to traditional methods (e.g., Gini importance), SHAP's advantage lies in its model-independence, which consistently explains the prediction mechanisms of different algorithms (e.g., GBDT, neural networks) and supports both global feature importance and local sample-level interpretation. For tree models, the efficient TreeSHAP algorithm reduces the computational complexity from exponential to polynomial level, making it suitable for high-dimensional biomedical data analysis. This feature makes SHAP an effective tool for parsing the decision logic of complex models.

3. Results and discussion

3.1. The performance comparison among models

Table 1: Performance comparison of baseline machine learning models in CRC prediction

Model name	Accuracy	Precision	Recall	F1 Score	ROC AUC
Gradient Boosting	**0.9201**	**0.8686**	0.9744	**0.9184**	**0.9897**
XGBoost	0.9083	0.8342	**1.0000**	0.9096	0.9761
RF	0.8935	0.8261	0.9744	0.8941	0.9646
DT	0.8876	0.8315	0.9487	0.8862	0.9218
KNN	0.8698	0.8043	0.9487	0.8706	0.9310
MLP	0.8580	0.8000	0.9231	0.8571	0.9428
LR	0.7692	0.7241	0.8077	0.7636	0.8609

DOI: 10.54254/2755-2721/166/2025.TJ23598

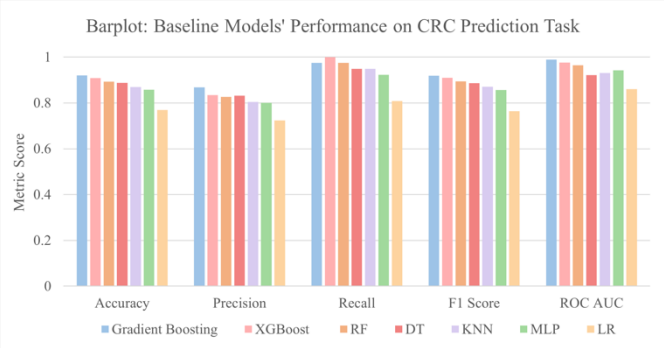
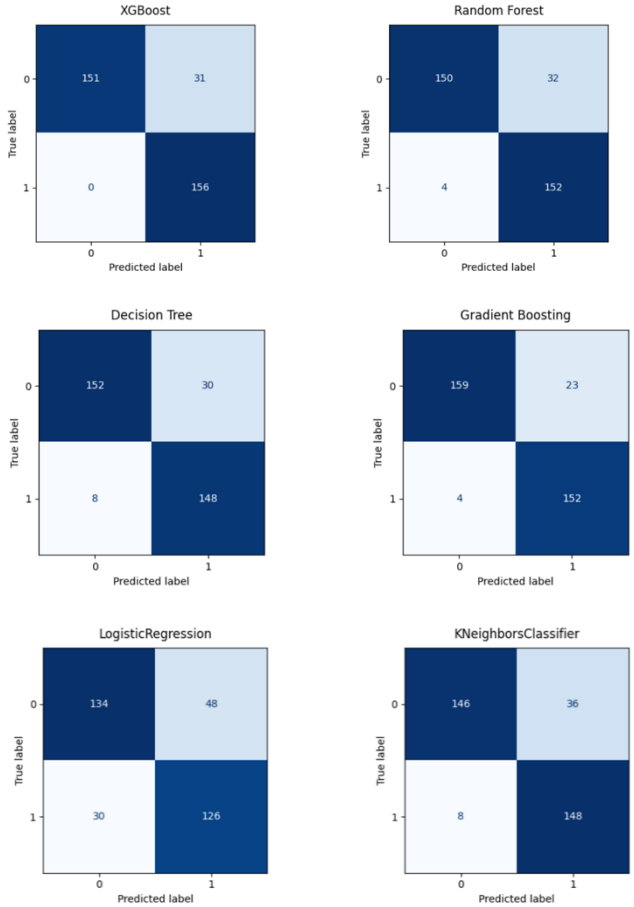


Figure 2: Baseline models' performance on CRC prediction task

Table 2: Performance comparison of stacking ensemble models with different base and meta learners

Baseline Models	Meta Models	Accuracy	Precision	Recall	F1 Score	ROC AUC
GBDT+ XGBoost	RF	**0.9527**	**0.9268**	**0.9744**	**0.9500**	**0.9923**
	KNN	0.9438	0.9152	0.9679	0.9408	0.9708
	MLP	0.9320	0.8982	0.9615	0.9288	0.9876
LR+ GBDT+ XGBoost	RF	**0.9556**	**0.9325**	**0.9744**	**0.9530**	**0.9954**
	KNN	0.9467	0.9207	0.9679	0.9437	0.9772
	MLP	0.9379	0.9042	0.9679	0.9350	0.9881



DOI: 10.54254/2755-2721/166/2025.TJ23598

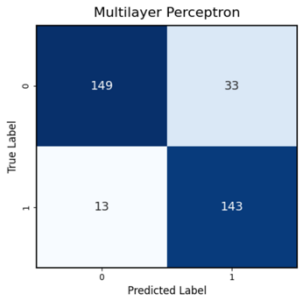


Figure 3: Confusion matrix of baseline models

The experiment compares the overall performance of seven machine learning models in the CRC prediction task. As shown in Table I, Fig. 2 and Fig. 3, GBDT exhibits the most superior overall performance, significantly outperforming the other models in terms of Accuracy (0.9201), Precision (0.8686), F1 score (0.9184) and ROC AUC (0.9897). This excellent performance is highly consistent with its confusion matrix results: the model produces only 4 false negatives (FN) and 23 false positives (FP), suggesting excellent classification robustness and balance. The XGBoost model achieves 100% recall (FN=0) but is accompanied by a higher number of false positives (FP=31), which explains its relatively low precision. While RF and DT perform similarly, RF has fewer false negatives, highlighting the advantages of the integrated learning approach. The KNN and MLP models are similar in terms of the number of false positives (FP=36), but the MLP has significantly more false negatives reflecting the fact that its deeper feature extraction capabilities have not been fully utilized. LR performed the least well, with significantly high numbers of false negatives (FN=30) and false positives (FP=48) in its confusion matrix, a result corroborated by its lowest ROC AUC (0.8609) and accuracy (0.7692), which amply demonstrates the inherent limitations of linear models in processing complex biomedical data.

From the analysis of the model's own characteristics, the overall performance of the integrated learning methods (Gradient Boosting, XGBoost, RF) is excellent, which is mainly attributed to their enhanced generalization ability by combining multiple weak learners and effectively capturing the complex interactions among features. The outstanding performance of Gradient Boosting may stem from its training mechanism of progressively optimizing the residuals, which reduces bias and variance, while XGBoost's extremely high recall (1.0) may be related to its regularization strategy and custom loss function optimization, but may also imply overfitting risk. In contrast, a single decision tree (DT) performs slightly worse due to the lack of robustness of integrated learning, KNN and MLP perform moderately well due to data distribution and feature scale sensitivity, and LR's linearity assumption makes it difficult to fit nonlinear patterns in CRC prediction, resulting in the worst performance.

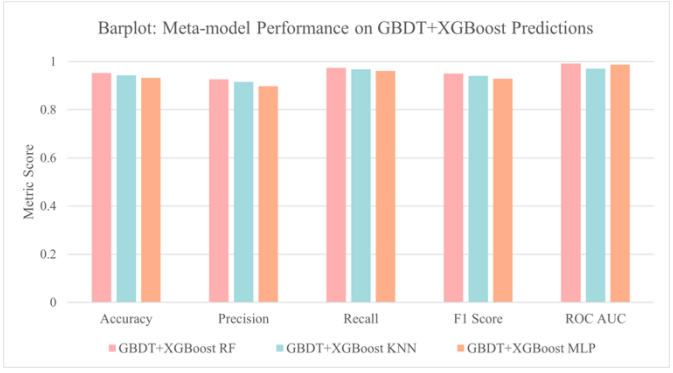


Figure 4: Meta-model performance on GBDT+XGBoost predictions

DOI: 10.54254/2755-2721/166/2025.TJ23598

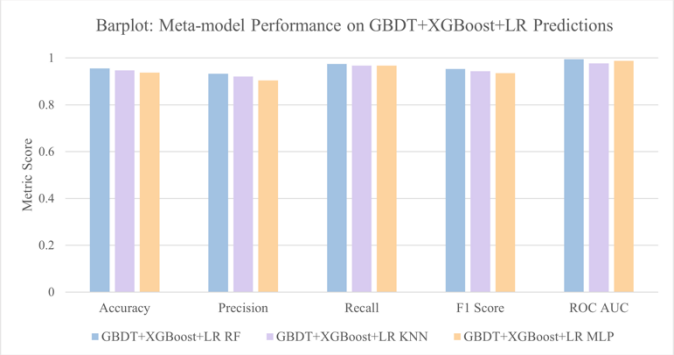


Figure 5: Meta-model performance on GBDT+XGBoost+LR predictions

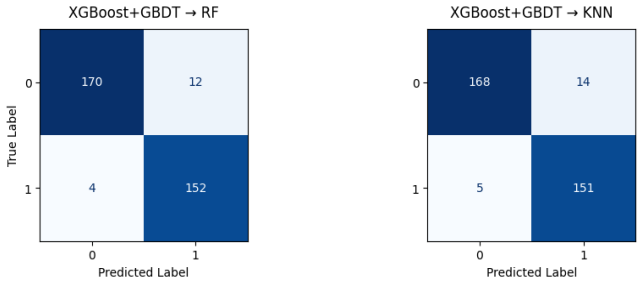


Figure 6: Confusion matrices: XGBoost+GBDT ensemble with RF/KNN meta-models

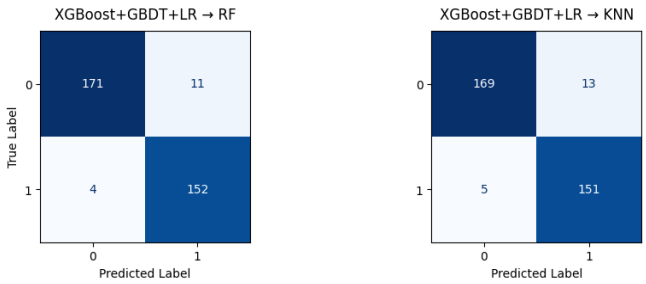


Figure 7: Confusion matrices: XGBoost+GBDT+LR ensemble with RF/KNN meta-models

As shown in Table II, Fig. 4, Fig. 5, Fig. 6 and Fig. 7, the stacked integration-based modeling framework significantly improves the performance of the CRC prediction task. By using GBDT and XGBoost as base learners and RF, KNN and MLP as meta-models for integrated modeling, respectively, all combinations show excellent prediction capabilities. Among them, the integrated RF model based on GBDT and XGBoost performs the most outstandingly, with an accuracy of 0.9556, an F1 score of 0.9530, and a near-perfect ROC AUC (0.9954). It is worth noting that even when the poorly performing LR is introduced into the base learner, its integrated performance is still significantly better than the prediction results of a single base model, e.g., the RF integrated model of GBDT, XGBoost and LR has an F1 score of 0.9500, which is even more than the integrated scheme that only uses GBDT and XGBoost as the base learner. This phenomenon indicates that the stacked integration approach can effectively integrate the advantages of different base learners and further optimize the prediction results through the nonlinear mapping ability of the metamodel, thus improving the robustness and generalization performance of the model in general.

From the perspective of algorithmic mechanism, the superiority of stacked integration mainly stems from two aspects: first, GBDT and XGBoost, as high-performance base learners, are able to

DOI: 10.54254/2755-2721/166/2025.TJ23598

adequately capture the complex nonlinear relationships and higher-order feature interactions in the data, which provide high-quality input features for the meta-model; second, RF, as a meta-model, further reduces the variance of the model by integrating the predictions of multiple decision trees, thus enhancing the overall stability. which enhances the overall stability. In contrast, KNN and MLP perform less well as metamodels, which may be related to their scale sensitivity to the input data (KNN) or dependence on the training sample size (MLP). In addition, although LR performs poorly when used as a base learner alone, its linear nature may play a role in complementing the diversity error in the integrated framework, thus indirectly improving the generalization ability of the overall model.

3.2. Feature importance represented by SHAP

As shown in Fig. 8, the SHAP analysis based on GBDT clarified the differences in the contribution of each clinical feature to CRC risk prediction. Feature importance ranking showed that Activity_Level and BMI had the most significant impact on model output, with a wide range of SHAP values and large absolute values. Specifically, Activity_Level showed a significant negative association, indicating that lower activity levels significantly increased the risk of disease, while BMI showed a positive association, suggesting that higher BMI was positively associated with disease risk. In addition, the distribution of positive SHAP values for Family_History_CRC and Age, as sub-important characteristics, further validated the clinical significance of having a family history of CRC and increasing age in the development of CRC.

Regarding metabolic indicators, the distribution of SHAP values for nutritional traits such as Fats(g) and Vitamin A(IU) was relatively symmetrical and concentrated around the zero value, suggesting that their contribution to risk prediction was more neutral. It is worth noting that the absolute SHAP values of some co-morbidity characteristics (e.g., Disease_Hypertension) and ethnicity categorical variables (e.g., Ethnicity_Asian, Ethnicity_Hispanic) were generally small, suggesting that these factors had limited discriminative value in the GBDT model.

BMI and Activity_Level had significant effects on CRC risk prediction, while the contribution of the characteristic Gender was relatively limited. In terms of pathophysiologic mechanisms, high BMI may promote CRC through mechanisms such as insulin resistance and chronic inflammatory response [15]. Higher Activity_Level, on the other hand, plays a protective role by improving insulin sensitivity, regulating intestinal flora and enhancing immune surveillance. In contrast, smoking in Activity_Level will alter the structure of the gut flora, leading to an increase in the relative abundance of pro-inflammatory bacteria and a decrease in the relative abundance of anti-inflammatory bacteria, which in turn promotes tumorigenesis [16]. In contrast, the effect of gender differences is relatively small, and although the gene KDM5D on the Y chromosome enables cancer cells in male CRC patients to evade detection and destruction by immune cells [17], promoting cancer development and metastasis, this risk may be statistically masked by other strong predictors such as BMI. These findings suggest that weight management and exercise interventions may have more public health significance in CRC prevention strategies than focusing on gender differences alone.

DOI: 10.54254/2755-2721/166/2025.TJ23598

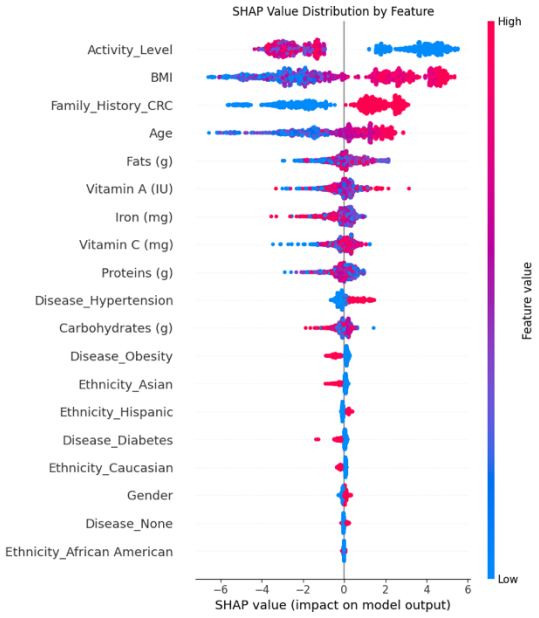


Figure 8: SHAP values of GBDT model predictions

4. Conclusion

In this study, a two-stage integrated learning framework is constructed based on the CRC multidimensional dataset from the Kaggle platform, aiming to optimize the performance of CRC risk prediction models through probabilistic feature fusion. The results show that the integrated learning approach significantly outperforms a single model in the CRC prediction task. Among them, the integrated architecture with GBDT, XGBoost as the baseline model and RF as the meta-model performs optimally and achieves a significant performance improvement compared to the single model (Accuracy=0.9527). Moreover, even with the incorporation of a poorly performing baseline model, such as LR, the integrated model still maintains excellent prediction ability (Accuracy=0.9556), which fully proves the robustness of the framework and the strong feature migration ability. However, the study still has some limitations, such as the small sample size of the dataset and limited geographic representation of data sources, which may limit the generalizability of the model. Future work could be extended to multicenter clinical data and introduce metabolomics or genomics features to enhance model interpretability.

References

[1] *World Health Organization, "Colorectal cancer," Fact Sheet, Mar. 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer. [Accessed: Apr. 25, 2025].*

[2] *L. Li, "Research on image coding based on DCT and wavelet transform," M.S. thesis, Guangdong University of Technology, Guangzhou, China, 2008.*

[3] *C. Fang, Y. Li, M. Xiong, et al., "Comparison of multiple linear regression and machine learning in predicting fear of cancer recurrence in newly diagnosed breast cancer patients," Journal of Wuhan University (Medical Sciences), pp. 1–7, 2023.*

[4] *Z. Li, Y. Cai, Y. Wang, et al., "Machine learning-based prediction of cancer-specific survival after endoscopic treatment in early colorectal adenocarcinoma patients," Nursing Research, vol. 38, no. 14, pp. 2459–2467, 2024.*

[5] *E. Y. Abbasi, "Cancer prediction and diagnosis using integrated multi-omics approaches with machine learning and deep learning models," Ph.D. dissertation, Beijing University of Posts and Telecommunications, 2024.*

[6] *X. Pan, K. Tong, C. Yan, et al., "Research progress in colorectal cancer recognition using convolutional neural networks," Journal of Biomedical Engineering, vol. 41, no. 4, pp. 854–860, 2024.*

DOI: 10.54254/2755-2721/166/2025.TJ23598

[7] *Z. Y. A., "Colorectal cancer dietary and lifestyle dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/ziya07/colorectal-cancer-dietary-and-lifestyle-dataset. [Accessed: Apr. 25, 2025].*

[8] *J. Luan, C. Zhang, B. Xu, Y. Xue, and Y. Ren, "The predictive performances of random forest models with limited sample size and different species traits," Fisheries Research, vol. 227, 105534, 2020.*

[9] *S. Deng, W. Yuan, S. Guan, X. Lin, Z. Liao, and M. Li, "A decision tree algorithm based on adaptive entropy of feature value importance," Big Data Research, 100530, 2025.*

[10] *H. I. Abdalla and A. A. Amer, "Enhancing data classification using locally informed weighted k-nearest neighbor algorithm," Expert Systems with Applications, vol. 276, 126942, 2025.*

[11] *X. Zhao, P.-F. Zhang, D. Zhang, Q. Zhao, and Y. Tuerxunmaimaiti, "Prediction of interlaminar shear strength retention of FRP bars in marine concrete environments using XGBoost model," Journal of Building Engineering, vol. 105, 112466, 2025.*

[12] *J. Luo, Y. Yuan, and S. Xu, "Improving GBDT performance on imbalanced datasets: An empirical study of class-balanced loss functions," Neurocomputing, vol. 634, 129896, 2025.*

[13] *S. Sarakon, W. Massagram, and K. Tamee, "Multisource data fusion using MLP for human activity recognition," Computers, Materials and Continua, vol. 82, no. 2, pp. 2109–2136, 2025.*

[14] *Z. Zhang, M. Tantai, H. Ma, S. Yu, B. Chen, and Z. Lu, "Analysis of risk factors for lumbar spondylolisthesis: A logistic regression study," World Neurosurgery, vol. 197, 123931, 2025.*

[15] *X. Luo, Y. Ju, S. Meng, et al., "The relationship between smoking and gut microbiota and inflammation in patients with colorectal adenoma," Chinese Journal of Microecology, vol. 37, no. 1, pp. 70–77, 2025.*

[16] *X. Ma, N. Li, H. Zhao, et al., "Analysis of incidence and mortality of colorectal cancer and its risk factors in China in 1990 and 2021," Journal of Practical Oncology, vol. 40, no. 2, pp. 114–119, 2025.*

[17] *J. Li, Z. Lan, W. Liao, et al., "Histone demethylase KDM5D upregulation drives sex differences in colon cancer," Nature, vol. 619, pp. 632–639, 2023.*