Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22274

Colorectal Cancer Prediction Based on Random Forest Approach

Yutong Chen

Beijing Normal University, Beijing, China
202211998194@mail.bnu.edu.cn

Abstract: In this study, a predictive model for colorectal cancer was constructed using the random forest method to improve early diagnosis and treatment. The study was based on 89,945 sample data provided by Kaggle, covering patients' basic information, medical history, diagnosis, lifestyle and other multi-dimensional characteristics. Through data preprocessing and feature engineering, key predictors such as BMI, age, and relapse time were identified. The model significantly improved the prediction accuracy of death samples by parameter tuning and class imbalance treatment (e.g. SMOTE oversampling). The experimental results show that the random forest model performs well in the prediction of survival samples with an accuracy of 75.16%, but there is still room for improvement in the prediction of death samples. Compared to other models such as logistic regression and XGBoost, random Forest performs better on key metrics such as accuracy, recall, and F1 scores, but its model complexity and computational resource requirements are high. Through cross-validation, the study confirmed the influence of characteristics such as BMI, age, and recurrence time on the stability of model predictions. The clinical significance of the model is also discussed, indicating that it can assist physicians in optimizing treatment plans and personalized treatment decisions. Finally, the study suggests the use of larger data sets in the future to improve the generalization ability of the model.

Keywords: Colorectal cancer, Random forest model, Machine learning, Feature importance

1. Introduction

1.1. Background and Significance of the Study

Colorectal cancer (CRC), a common malignant tumor, has high incidence and mortality, challenging global public health. Early diagnosis is vital for patients' survival, yet many are diagnosed late due to hidden early symptoms.

Machine Learning (ML), especially Random Forest (RF), can aid accurate diagnosis and personalized treatment by analyzing medical data. While RF has been effective in some cancer predictions, its use in CRC prediction needs more validation. So, this study aims to assess the RF model's effectiveness in CRC prediction and compare it with other ML models.

Multiple factors like lifestyle and genetics influence CRC. Analyzing them helps identify high-risk groups for targeted strategies. The study's essence is using ML for early CRC screening, better

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ22274

treatment, and resource allocation. Validating RF's effectiveness provides a new clinical tool and boosts AI in medicine, hopefully offering new ways for CRC management and improving patients' lives.

1.2. Thesis Structure

In this thesis, the random forest method constructs a colorectal cancer prediction model to improve early diagnosis and treatment via machine learning. The introduction presents colorectal cancer's global morbidity, mortality, early - diagnosis importance, and machine - learning prospects to show the model's value. The literature review compares traditional and machine - learning methods and examines relevant research and applications. The model - construction part details dataset processing, random forest principles, parameter tuning, and evaluation metrics. Experiments display the model's performance, compare it with others, and analyze key factors. The discussion analyzes results, notes limitations, and suggests future directions. The conclusion summarizes findings, emphasizes applica- tion prospects, offers clinical advice, and provides references and appendices for study transparency and reproducibility.

2. Literature Review

Random Forest is an ensemble learning algorithm based on machine learning, first proposed by Leo Breiman[1] and Adele Cutler in 2001. Since it was proposed in 2001, random forest has made great progress in both theoretical research and practical application [2]. Over time, the random forest algorithm has been improved and optimized, and many variants have been derived to adapt to the data processing and analysis needs in different scenarios [3].

Colorectal cancer, a leading global malignancy, has seen growing use of machine learning (ML) for prediction and diagnosis. Traditional methods like Logistic Regression and Cox models struggle with complex, non-linear data, while ML techniques like Random Forest (RF) and Support Vector Machines (SVM) excel in handling high-dimensional data and capturing non-linear relationships. However, challenges like data quality and interpretability remain for clinical adoption.

3. Model Construction

This study utilizes the Colorectal Cancer Prediction dataset from Kaggle, comprising 89,945 pa- tient records uniquely identified by Patient ID. The dataset includes 29 variables covering a wide range of information: basic patient details (age, gender, ethnicity, region, residence, socioeconomic status), medical history (family and personal cancer history), cancer diagnosis (stage, tumor aggressiveness, diagnosis time), screening and exams (colonoscopy access, screening regularity), lifestyle and health indicators (diet, body weight, physical activity, smoking, alcohol consumption, red meat and fiber intake), healthcare coverage and treatment (insurance, treatment routes, chemotherapy, radiation ther- apy, surgery), and follow-up information (adherence to follow-ups, survival status, recurrence status, and recurrence time). The dataset contains three 'int64', one 'float64', and 26 'object' type fields, with two continuous variables: body mass index and recurrence time.

3.1. Data Preprocessing

First, observe the three numerical variables BMI, Age and Time_to _ Recurrence, draw their box-and-line plots to observe the distribution of the variables:

DOI: 10.54254/2755-2721/2025.TJ22274

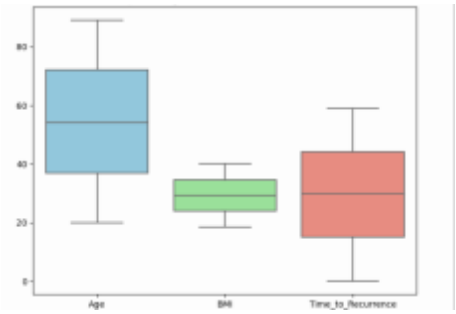


Figure 1: Boxplot of Age,BMl, and Time to Recurrence

As can be seen from the graph, there are very few outliers in the data, especially in the BMI variable, and the data itself is of high quality. Although there are some outliers in Age and Time_to Recu- rrence, these outliers do not affect the overall data too much, so no special transformations of the variables are needed. Additionally, the scales of the three variables varied widely, with Age and Time_to _ Recurrence having a significantly wider range of values, indicating greater variance, while BMI was more centralized. While these variables do not span multiple orders of magnitude, and therefore it is desirable to not normalize the data, we can still consider normalization when manipulating the data for dimensionality reduction.

In addition, histograms of Patient_ID, Age, BMI, and Time_to _ Recurrence were plotted as follows:

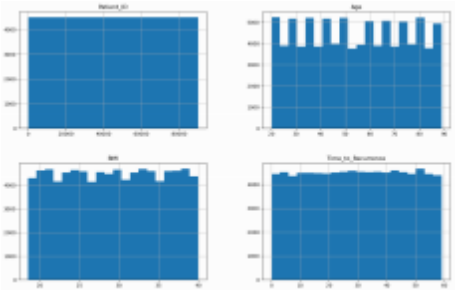


Figure 2: Histograms of Patient_ID, Age, BMI, and Time_to_Recurrence

The Patient_ID plot shows unique and consistent counts for each patient, indicating no duplicates. The Age plot reveals a uniform age distribution between 20-90 years, suggesting broad age coverage without concentration. The BMI graph indicates that values are concentrated within 20-40, with minimal fluctuations and no extreme outliers. The Time_to_Recurrence graph shows an even distribution of recurrence times within 0-60 intervals, with no concentrated recurrence period.

For Age and Time_to _ Recurrence, the Box_Cox transformation was used to bring the data closer to a normal distribution.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases} \quad (1)$$

where y is the original data and λ is the transformation parameter, determined by maximum likelihood estimation.

DOI: 10.54254/2755-2721/2025.TJ22274

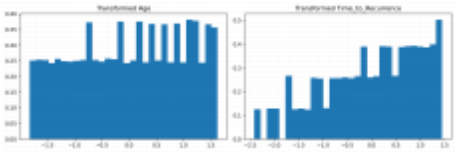


Figure 3: Histograms of post-normal distribution

The histogram of“Transformed Age” shows an approximately normal distribution centered near 0, with moderate dispersion and no extreme values. The histogram of “Transformed Time_to Recurrence” is right-skewed, with peaks near 0 and 1.5, indicating significant data variability and a non-standard single-peak pattern.

At the same time, the distribution of the categorical variables was visualized.

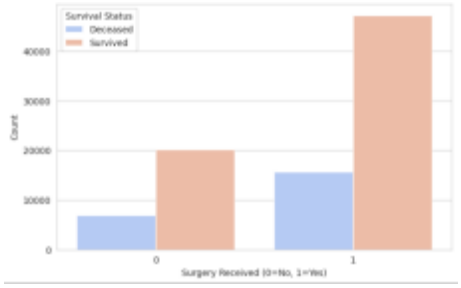


Figure 4: Surgery Received vs. Survival Status

As can be seen in the graph: when not undergoing surgery (0), the height of the blue bar is low, indicating that fewer people have died; the height of the pink bar is medium, indicating that more people have survived. When undergoing surgery (1), the height of the blue bar increases but remains lower than the pink bar, indicating that more people survived.Overall, undergoing surgery appears to be associated with a higher number of survivors.

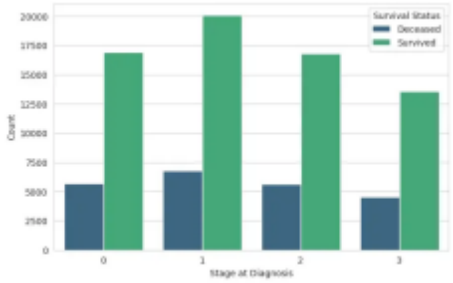


Figure 5: Stage at Diagnosis vs.Survival Status

The figure shows the relationship between diagnosis stage and survival. For Stage 0, survivors outnumber deaths slightly (over 15,000 survivors, 5,000 - 7,500 deaths). Stage 1 has the most survivors (over 18,000) and fewer deaths (6,000 - 9,000). Stage 2 has over 16,000 survivors and a death count like Stage 0 (5,500 - 8,500). Stage 3 has the fewest survivors (above 12,000) and the lowest death count (3,500 - 6,500). Overall, as the stage increases, survivors decrease while deaths fluctuate, suggesting better survival at early stages.

DOI: 10.54254/2755-2721/2025.TJ22274

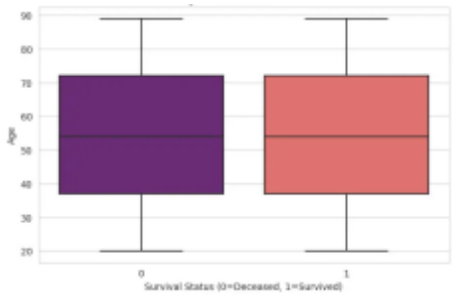


Figure 6: Age vs.Survival Status

This box-and-line graph shows the distribution of age in different survival states.This implies that age is not a major factor influencing survival status in this particular sample.

3.2. Random Forest Model

3.2.1. Fundamentals of Random Forests

Random Forest belongs to the ”Bagging (Bootstrap Aggregating)” framework of integrated learning methods. The core idea of integrated learning is to reduce the variance of the model and improve the generalization ability by combining the prediction results of multiple base learners (e.g., decision trees). Specifically, Random Forest achieves integration in the following two ways: sample random- ness, where multiple subsample sets are extracted from the training set through Bootstrap sampling, and each decision tree is trained based on a different subsample set; and feature randomness, where a portion of the features are randomly selected for optimal segmentation when the nodes of each decision tree are split, instead of using all the features.

In terms of feature randomization, let the dataset have a total of M features, denoted as F = {f1 , f2,..., fM }. Instead of using all M features to find the optimal split when splitting the nodes of each decision tree, m(m < M) features are randomly selected, denoted as Fsubset = {fs1, fs2,..., fsm }, fsi ∈ F, and then based on these m features, we compute the information gain , Gini index and other indicators, and then determine the optimal splitting features and splitting points. For example, when calculating the information gain, for a feature f, its information gain IG(D, f) is calculated as:

$$IG(D,f) = H(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} H \tag{2}$$

Where $H(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$ is the information entropy of the dataset D,K is the number of kinds of category labels in the dataset, and p_k is the proportion of samples in the dataset belonging to the kth class; V is the number of values of the feature f, and Dv is the subset of the samples when the value of the feature f takes the vth value and |D| and |Dv | are the number of samples in the dataset D and subset Dv , respectively. The diversity among base learners is further enhanced by this feature randomization.

DOI: 10.54254/2755-2721/2025.TJ22274

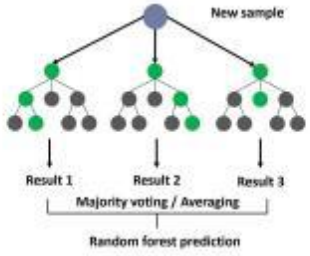


Figure 7: basic principles of decision trees

In addition to the information gain described above, the Gini index is a common measure of purity.

For a dataset D, its Gini index Gini(D) is calculated as:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3}$$

In the decision tree construction process, the features and splitting points that make the purity metrics such as Gini index or information gain of the split subset optimal will be selected for node splitting.For this problem, the prediction result of random forest can be expressed as:

$$y\hat{} = \text{mode}(\{T1(x), T2(x), ..., Tk(x)\}) \tag{4}$$

where Ti (x) denotes the prediction result of the ith decision tree, and mode denotes the plural.

3.2.2. Parameterization and Tuning of Models

In this study, we used a Random Forest Classifier. By default, it allows free tree growth until certain conditions, risking overfitting, especially with complex data. To fix this, we tuned the model. We set the number of trees to 200 for balance, limited max depth to 30 to prevent overfitting, set min sample splits to 5 for better generalization, min samples per leaf to 2 to avoid overfitting, and max features to log2 for more randomness and robustness.

3.2.3. Resolution of Category Imbalances

In the dataset of this study, there is a significant category imbalance where the number of survival (category 1) samples is much larger than the number of death (category 0) samples. This imbalance may cause the model to be overly inclined to predict the dominant category (category 1), thus ignoring a few categories (category 0). To mitigate this problem, we used the category weight(class weight $='$ balanced$'$) parameter when training the model, which adjusts the model's focus on different categories based on their frequency. By doing this, the model will give higher weight to samples from a few classes (deaths) during the training process, thus improving its predictive power on the test set.

3.3. Model Evaluation Indicators

Precision, recall and F1 score [4] are a few key metrics when evaluating colon cancer diagnostic models, and they are described in detail below.

Precision is used to measure the degree of precision of the model's determination results. It is calculated as

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

DOI: 10.54254/2755-2721/2025.TJ22274

where TP represents the number of cases that the model accurately determines as having colon cancer, and FP represents the number of cases where the model mistakenly determines a person who does not have colon cancer as having colon cancer in the context of colon cancer diagnosis.

Recall is concerned with the model's ability to identify true positive samples and is calculated as

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

where FN refers to the number of cases where the model incorrectly determined that a patient with colon cancer had colon cancer as not having colon cancer.

The F1 score[5] is the reconciled mean of the precision and recall rates and is calculated as

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{7}$$

4. Experimental Results and Analysis

4.1. Predictive Performance of Random Forest Model

Table 1: Classification Report

	precision	recall	f1-score	support
0	0.67	0.00	0.00	4471
1	0.75	1.00	0.86	13518
accuracy			0.75	17989
macro avg	0.71	0.50	0.43	17989
weighted avg	0.73	0.75	0.65	17989

Without addressing category imbalance, the model achieves% accuracy for category 0 (death).

After parameter tuning, overall accuracy improves to 75.16%, with 75% of survival samples (category 1) correctly classified. For category 1, precision is 0.75, recall is 1.00, and the F1 score is 0.86, indicating excellent performance in identifying survival samples. However, for category 0, precision is 0.67, but recall and F1 score are 0, meaning the model fails to identify any true death samples, misclassifying all as survival.

To improve accuracy for category 0, SMOTE is applied to oversample death samples, balancing the dataset. This significantly enhances the model's ability to predict category 0, as demonstrated in the final results.

Table 2: Classification Report

	precision	recall	f1-score	support
0	0.67	0.54	0.61	4471
1	0.75	1.00	0.86	13518

4.2. Comparison to Other Machine Learning Models

In order to comprehensively assess the performance of the Random Forest model, we compare and analyze it with the logistic regression and extreme gradient boosting models. These three models have their own characteristics in data processing and prediction mechanism, and the comparison results can provide a strong basis for model selection and optimization.

DOI: 10.54254/2755-2721/2025.TJ22274

Table 3: Classification Report

	accuracy	precision	recall	F1 score
Logistic regression	0.75	0.75	1.00	0.86
XGBoost	0.74	0.75	0.99	0.85
random forest	0.75	0.76	1.00	0.86

The comprehensive comparison results show that in this colorectal cancer prediction study, the random forest model performs well in key metrics such as accuracy, precision, recall, and F1 score, and demonstrates a better performance when dealing with data related to colorectal cancer prediction. However, the random forest model is not perfect. Its model structure is relatively complex, and its interpretability is slightly weaker than that of logistic regression. Although the decision basis of the model can be explained by feature importance analysis, it may still be difficult for some doctors to understand. In addition, the computational complexity of the random forest model is also higher, and the training and prediction process requires more computational resources and time.

4.3. Feature Importance Analysis

We obtained feature importance scores in the random forest model using model.feature_importances.

These scores reflect the contribution to node - splitting purity. We paired the scores with feature names from X.columns in a pandas DataFrame feature_importance_df. Then, we sorted features by impor- tance in descending order using sort_values. Eventually, we got the top 10 features and their scores as follows

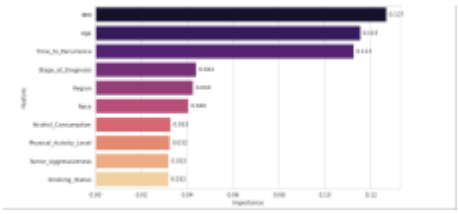


Figure 8: Top10 Feature Importance from Random ForestModel

According to the graph, the top ten features by importance are BMI, Age, Time_to_Recurrence, Stage_at_Diagnosis, Region, Race, Alcohol_Consumption, Physical_Activity_Level, Tumor_Aggressiveness, and Smoking_Status. BMI has the highest importance score, greatly influencing model predictions. It reflects the weight-height ratio and relates to disease risks. Age is the second most impor- tant. As people age, bodily functions decline, increasing disease likelihood and affecting predictions. Time_to_Recurrence is crucial for assessing disease progression and prognosis, directly impacting model results. Though some studies link age, BMI, and health risks, a correlation test showed no cor- relation between age and BMI, indicating their separate significance for health risks. Smoking_Status and Tumor_Aggressiveness have low importance scores, probably due to weak links with target vari- ables in the dataset.

4.4. Cross-validation

Cross-validation, which involves dividing the dataset multiple times for training, can mitigate result bias caused by the randomness of a single data split. To verify the reliability of feature importance results, 5-fold cross-validation is employed. The dataset is partitioned in different ways several times. Feature importance is calculated during each cross-validation, and the results are as follows

DOI: 10.54254/2755-2721/2025.TJ22274

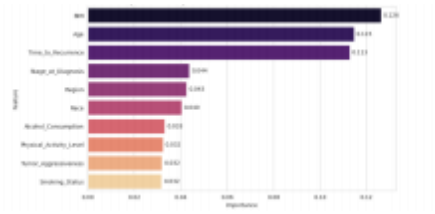


Figure 9: Top 10Feature lmportance from Random Forest Model(Cross-validated)

By comparing the results before and after the cross-test we found that the results of feature importance before and after the cross-test were relatively stable, indicating that the model's assessment of feature importance has a certain degree of reliability and stability, and that features such as body mass index, age, and recurrence time are the key factors affecting the model's prediction.

5. Deliberations

Mechanisms of Influence of Key Characteristics on Colorectal Cancer and Clinical Significance of Model Predictions Random forest modeling identified key features in colorectal cancer prediction. BMI is top - ranked. High BMI, linked to obesity, causes chronic inflammation, increasing abnormal intestinal cell growth risk; low BMI weakens immunity. Age is second - most important. As age rises, DNA repair, immune function, and intestinal mucosa decline, upping cancer risk. Shorter Time_to_Recurrence means more aggressive tumors and a worse prognosis. Stage_at_Diagnosis is vital. Early diagnosis improves survival, while late diagnosis is tough to treat. Regional/ethnic factors, due to environment, lifestyle (e.g., high - fat, low - fiber diets), and genes (specific mutations), can raise disease risk. Alcohol, inactivity, smoking, and tumor aggressiveness, though less impactful, still affect the disease. Alcohol harms the intestinal mucosa, inactivity keeps harmful substances longer, smoking brings carcinogens, and tumor aggressiveness shows malignancy. The model's predictions are clinically useful. It helps doctors assess patient survival odds for survival samples, optimizing treatment and avoiding overtreatment. After data balancing, it better predicts death samples, helping doctors spot high - risk patients and adjust strategies. Also, it aids personalized treatment. For low - survival patients, aggressive treatments can be used; for high - survival ones, milder ones are chosen. It evaluates treatment effectiveness, guiding doctors to pick better options and allocate resources to high - risk patients.

6. Conclusion

6.1. Key Findings of the Study

The aim of this study is to construct a colorectal cancer prediction model with the help of random forest method. we conducted an in-depth analysis of the Colorectal Cancer Prediction dataset provided by Kaggle and achieved a series of important results. First, the study identified the key features that affect the prediction of colorectal cancer, including BMI, Age, Time_to_Recurrence and so on. These features play an important role in the process of colorectal cancer occurrence, development, and patient prognosis.

During the model training and evaluation phase, we found that the random forest model has some ability in predicting the survival and death of colorectal cancer patients. However, due to the problem of category imbalance in the dataset, it resulted in a situation of poor accuracy in the model's predic- tion of death samples. For this reason, we used methods such as setting category weights and SMOTE oversampling to process the data, and the accuracy of the model's prediction of death samples was significantly improved after processing.

DOI: 10.54254/2755-2721/2025.TJ22274

In the comparison test with other machine learning models, the advantages and limitations of the random forest model in processing the data of this study were further verified. Meanwhile, through cross-validation and testing on different datasets, the model was confirmed to be reliable and stable for feature importance assessment, which provides strong support for subsequent research and clinical applications.

6.2. Limitations and Future Research Direcions

The Colorectal Cancer Prediction dataset provided by Kaggle, which was used in this study, has some limitations, although it covers a wide range of patient information. The number of samples in the dataset is relatively limited, which may not adequately cover all possible colorectal cancer case scenarios, resulting in a model that lacks in generalization ability.

Moreover, the predictions of the random forest model may also be affected by data noise.

In order to improve the accuracy and generalization ability of the model, future studies need to be validated using larger datasets. Larger datasets can encompass a wider range of patient populations, including a diversity of different regions, races, ages, genders, and other factors, thus providing a more comprehensive picture of the various clinical features and incidence of colorectal cancer.

References

[1] *Breiman, L. Random Forests Machine Learning Volume 45, Number 1 (2001), 5 - 32.*

[2] *Tin Kam Ho, Random decision forests, In Proceedings of the 3rd international conference on document analysis and recognition, Volume 1, 1995, Pages 278 - 282.*

[3] *Andy Liaw and Matthew Wiener, Classification and regression by randomForest, R news Volume 2, Number 3, 2002, Pages 18 - 22.*

[4] *Yutaka Sasaki, The truth of the F - measure, Technical Report, Gmu - cs - tr - 2007 - 02, George Mason University, 2007.*

[5] *Tom M. Mitchell, Machine Learning, McGraw - Hill, 1997.*