Contents lists available at ScienceDirect

European Journal of Surgical Oncology

journal homepage: www.ejso.com

Machine learning for predicting liver and/or lung metastasis in colorectal cancer: A retrospective study based on the SEER database

Zhentian Guo [a,b], Zongming Zhang [a,b,*], Limin Liu [a,b], Yue Zhao [a,b], Zhuo Liu [a,b], Chong Zhang [a,b], Hui Qi [a,b], Jinqiu Feng [b,c], Chunmin Yang [d], Weiping Tai [e], Filippo Banchini [f], Riccardo Inchingolo [g]

[a] *Department of General Surgery, Beijing Electric Power Hospital, State Grid Corporation of China, Capital Medical University, Beijing, 100073, China*
[b] *Key Laboratory of Geriatrics (Hepatobiliary Diseases) of China General Technology Group, Beijing, 100073, China*
[c] *Department of Immunology, Peking University School of Basic Medical Sciences, Peking University, Beijing, 100191, China*
[d] *Department of Gastroenterology, Beijing Electric Power Hospital, State Grid Corporation of China, Capital Medical University, Beijing, 100073, China*
[e] *Department of Gastroenterology, Beijing Shijitan Hospital, Capital Medical University, Beijing, 100038, China*
[f] *General Surgery Unit, Guglielmo da Saliceto Hospital, Piacenza, Italy*
[g] *Interventional Radiology Unit, "F. Miulli" Regional General Hospital, Acquaviva delle Fonti, 70021, Italy*

ARTICLE INFO

Keywords:
Machine learning
Colorectal cancer
Liver and/or lung metastasis

ABSTRACT

Objective: This study aims to establish a machine learning (ML) model for predicting the risk of liver and/or lung metastasis in colorectal cancer (CRC).

Methods: Using the National Institutes of Health (NIH)'s Surveillance, Epidemiology, and End Results (SEER) database, a total of 51265 patients with pathological diagnosis of colorectal cancer from 2010 to 2015 were extracted for model development. On this basis, We have established 7 machine learning algorithm models. Evaluate the model based on accuracy, and AUC of receiver operating characteristics (ROC) and explain the relationship between clinical pathological features and target variables based on the best model. We validated the model among 196 colorectal cancer patients in Beijing Electric Power Hospital of Capital Medical University of China to evaluate its performance and universality. Finally, we have developed a network-based calculator using the best model to predict the risk of liver and/or lung metastasis in colorectal cancer patients.

Results: 51265 patients were enrolled in the study, of which 7864 (15.3 %) had distant liver and/or lung metastasis. RF had the best predictive ability, In the internal test set, with an accuracy of 0.895, AUC of 0.956, and AUPR of 0.896. In addition, the RF model was evaluated in the external validation set with an accuracy of 0.913, AUC of 0.912, and AUPR of 0.611.

Conclusion: In this study, we constructed an RF algorithm mode to predict the risk of colorectal liver and/or lung metastasis, to assist doctors in making clinical decisions.

1. Introduction

Colorectal cancer (CRC), as one of the most common malignant tumors in the world, is likewise the second leading cause of cancer death worldwide [1]. Metastasis is considered a key clinical feature of refractory CRC and a high-risk factor for high mortality [2]. The 5-year survival rate of CRC patients is about 56 %, but the survival rate may significantly shorten when patients are complicated with metastasis [3–5]. The liver and lung are the most common distant metastasis sites of CRC patients [6]. Unlike many other types of cancer, metastatic CRC can still receive treatment [7].

With the advancement of treatment methods, surgical techniques, and perioperative care, the prognosis of patients with metastatic colorectal cancer has dramatically improved. For patients with early-stage CRC combined with liver or lung metastasis, neoadjuvant chemotherapy can effectively improve the survival rate of metastatic patients by reducing the tumor and implementing surgical resection [2,8,9]. In addition, it was suggested by Bailey et al. that although the incidence of CRC has been decreasing in older persons, the incidence was increasing dramatically in young adults [10]. Early detection of high-risk CRC patients prone to liver and/or lung metastasis will help doctors carry out

[*] Corresponding author. Key Laboratory of Geriatrics (Hepatobiliary Diseases) of China General Technology Group, Beijing, 100073, China.
E-mail address: zhangzongming@mail.tsinghua.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.ejso.2024.108362
Received 3 December 2023; Received in revised form 11 April 2024; Accepted 20 April 2024
Available online 27 April 2024

Important abbreviation in this text

Abbr	original text
ML	machine learning
CRC	colorectal cancer
SEER	Surveillance, Epidemiology, and End Results
LR	logistic regression
RF	random forest
DT	decision tree
SVM	support vector machine
NB	naïve Bayes
KNN	k-nearest neighbor
XGBoost	extreme gradient boosting
GBM	gradient boosting machine
NOS	not otherwise specified
OS	overall survival
AUC	area under the curve
ROC	receiver operating characteristics
CEA	carcinoembryonic antigen
AUPR	area under the precision-recall curve

early intervention and individualized treatment and further improve the survival rate of patients. Therefore, it is meaningful to define the risk factors of metastasis in CRC patients and establish an efficient prediction model.

Machine learning (ML) has been extensively applied in different fields of clinical research [11,12]. Machine learning algorithms are based on statistics; however, ML has numerous advantages over traditional statistical methods in establishing data models and estimating characteristic coefficients [13]. In practical applications, machine learning can help analyze helpful information hidden in a large amount of data, reveal the relationships between data, and apply various machine learning algorithms to build prediction models. This study aims to establish an ML model for predicting the risk of liver and/or lung metastasis in CRC.

2. Materials and methods

2.1. Data sources and study population

The data comes from the SEER database. The data collection adopts SEER * stat 8.4.1 software. The subjects of this study were patients diagnosed with CRC in the United States from 2010 to 2015, and we chose patients using the procedure depicted in Fig. 1. CRC patients with missing data, unclear clinical and pathological conditions, histological uncertainty, or incomplete survival information of other types were excluded. Demographic and clinicopathological information included age, gender, race, Hispanic, marital status, and histological types, including adenocarcinoma (8140/3, 8210/3, 8261/3, 8263/3), mucinous adenocarcinoma (8480/3) and signet-ring cell carcinoma (8490/3), marital status, Grade, primary tumor site, T stage, N stage, tumor size, CEA, tumor deposits; Refer to the ICD-*O*-3 manual for histological type codes and identify site codes (C18.0, C18.1, C18.2, C18.3,

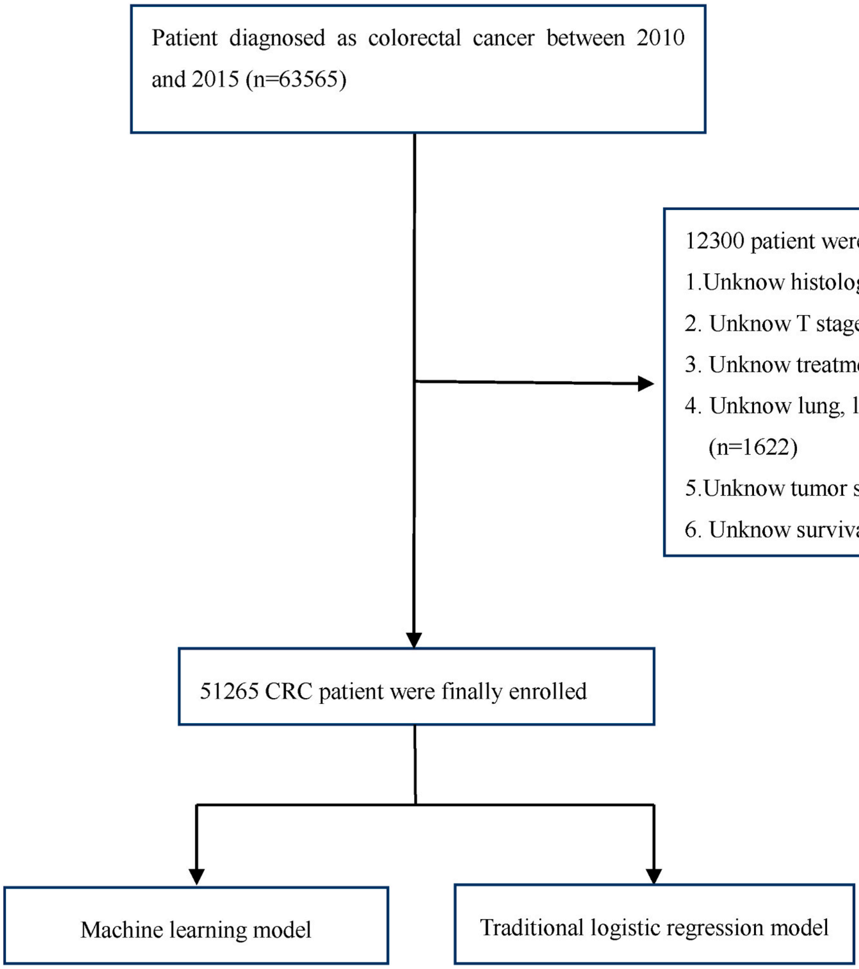
Patient diagnosed as colorectal cancer between 2010 and 2015 (n=63565)
12300 patient were excluded:
1.Unknow histological type (n=8158)
2. Unknow T stage, N stage (n=1732)
3. Unknow treatment methods (n=185)
4. Unknow lung, liver and bone metastasis status (n=1622)
5.Unknow tumor size (n=88)
6. Unknow survival information (n=515)
51265 CRC patient were finally enrolled
Machine learning model
Traditional logistic regression model

Fig. 1. The fow diagram of the selection process for the study.

C18.4, C18.5, C18.6, C18.7, C18.8, C18.9, C19.9, C20.9) and Cancer staging scheme (version 0204). Adopting AJCC 7th edition TNM stage. As the SEER database contains public data, informed consent from relevant patients for using the SEER database for research purposes was not required, nor was ethical approval. The National Cancer Institute, USA (reference number 19238-Nov2021) approved our request for access to the SEER data.

For external validation, we used data from 196 patients at Beijing Electric Power Hospital of Capital Medical University of China, with additional criteria of no neoadjuvant radiotherapy before surgery. The study was retrospective and did not involve patient safety or privacy, and an ethical exemption was granted.

2.2. Risk factor screening and model construction

Statistical analysis was conducted using SPSS software (version 26.0; IBM Corporation). In the univariable analysis, we employed Pearson's correlation analysis to examine the association between predictor variables, with results being presented in the form of heat maps. The Categorical variable is expressed in numbers and percentages and compared using the Chi-squared or Fisher's exact test. The predictive factors related to liver and/or lung metastasis were initially screened through univariable analysis ($p < 0.05$), and the variables that met the criteria were incorporated into a multivariable logistic regression (LR) analysis. The receiver operating characteristic (ROC) curve was plotted and analyzed based on the results. An area under the ROC curve (AUC) greater than 0.5 was considered meaningful. All computed p values were two-sided, and statistical significance was accepted at <0.05.

Use Python software (version 3.9.12, Python Software Foundation). Incorporate all variables into the ML model and build a prediction model. The data after the sampling process is randomly divided into a training set and test set at a ratio of 8:2. The training set uses seven standard machine learning algorithms, including random forest (RF), decision tree (DT), support vector machine (SVM), naive Bayes (NB), k nearest neighbor (KNN), eXtreme gradient boosting (XGBoost) and gradient boosting machine (GBM). Model evaluation is based primarily on accuracy, precision, recall, F1 score, AUC value and area under the precision-recall curve (AUPR), and the model with the highest ROC value is the optimal model. RF is a machine learning algorithm that processes classification and regression problems by constructing multiple decision trees; XGBoost is a classic decision tree algorithm applied to classification or regression prediction models; KNN algorithm is considered a critical classification algorithm in the field of supervised machine learning, widely used in pattern recognition data mining, etc [14]. SVM is a binary classifier, which usually divides things with multidimensional attributes into two categories accurately. DT model can accurately identify seven kinds of tumor histopathology, with high classification accuracy [15]. The NB algorithm has superior accuracy and has achieved good results in some clinical studies [16–18]. The GBM algorithm is a supervised learning algorithm with a vital function of identifying the importance of predictive variables [19]. We evaluated the model using two indicators, AUC and AUPR, and used the area under the curve as the basic indicator for evaluating model performance. The following formula was used to calculate model performance in this study: accuracy = (TP + TN)/(TP + TN + FP + FN); precision = (TP)/(TP + FP); recall = (TP)/(TP + FN); F1 score = (2*precision*recall)/(precision + recall). Based on the optimal machine learning algorithm model, evaluate the importance of each feature in predicting liver and/or lung metastasis using the principle of importance ranking.

The original data set is unbalanced since fewer colorectal cancer patients have liver and/or lung metastasis in the SEER database. We process the original data with the techniques of Under-sampling and Over-sampling and use the correlation matrix to analyze the changes in the data after sampling. The minority Over-sampling technique (SMOTE) or Under-sampling is a standard method for balancing classes on unbalanced data sets and is utilized to optimize models [15]. The correlation between variables becomes clearer after sampling, as shown in Fig. 2.

3. Results

3.1. Analysis of patient information

A total of 51265 patients were included in the present study, of which 7864 (15.3 %) had liver and/or lung metastasis, and 43401 (84.6 %) had no liver and/or lung metastasis. This also includes an external test set of 196 patients who were first diagnosed with colorectal cancer in our hospital from 2010 to 2017. Detailed information regarding the SEER database set and outer validation set can be found in Table 1.

In univariable analysis, liver and/or lung metastasis in colorectal patients was significantly correlated with age, histological type, Grade, primary tumor site, T stage, N stage, tumor size, CEA, and tumor deposition ($p < 0.05$), see Table 2—incorporation of the above characteristic variables into multivariable LR.

In the multivariable LR analysis, age, histological type, Grade, primary tumor site, T stage, N stage, tumor size, CEA, and tumor deposition are all independent predictors of metastasis to liver and/or lung of colorectal cancer; details are shown in Table 3. The ROC curve was drawn according to the traditional multivariable regression results (AUC = 0.833, 95%CI0.828-0.838, $p < 0.001$).

Analysis of machine-learning algorithm results: Seven machine-learning models are developed and compared based on accuracy, precision, recall, F1 score, and AUC value. The machine-learning model trained by over-sampling data is better than that trained by under-sampling data; see Table 4 for the details of seven machine-learning models constructed by over-sampling data. By using over-sampling and under-sampling to build seven machine-learning models, the performance of the training set and test set is shown in Fig. 3. In the over-sampling data-building model, the AUC of all models is higher than 0.700. Among them, the performance of the RF model is preferable to other models, In the internal test set, with an AUC of 0.956, and an area under the precision-recall curve (AUPR) of 0.896. In the external validation set, the RF algorithm model also achieved excellent results with an accuracy of 0.913, AUC of 0.912, and AUPR of 0.611, as shown in Fig. 5. Comparing the AUC values of the RF model and the traditional LR model, it is shown that the machine-learning RF algorithm has higher diagnostic efficiency than the traditional LR model and has good predictive performance. Utilizing RF for feature selection, tumor deposits are a vital predictor for determining whether colorectal cancer patients have liver and/or lung metastasis, as shown in Fig. 4.

This study developed an online network calculator for evaluating the risk of liver and/or lung metastasis in colorectal cancer patients, which can be applied to clinical patients. (http://121.43.117.60:8001/).

4. Discussion

CRC is one of the malignant tumors with the highest incidence in the world, and it is also the leading cause of death in cancer patients [1], and metastasis is the main cause of death in CRC patients [20]. About 25 % of patients experience metastasis at the initial diagnosis, nearly half of patients experience metastasis as the tumor progresses, and about 21 % are diagnosed with stage IV. Liver and lung metastasis are the most common metastasis in CRC patients [21,22]. As a common site of metastasis, population-based studies have demonstrated that 25%–30 % of CRC patients experience liver metastasis during the disease process [23]. The lung is the second most common site of CRC metastasis, second only to the liver, accounting for about 10%–15 % of metastasis [24], and the prognosis of patients with lung metastasis is generally better than that of CRC patients with other metastasis [25]. When CRC patients develop metastasis, the survival time is considerably shortened. Early identification of metastasis is conducive to quick intervention and delay

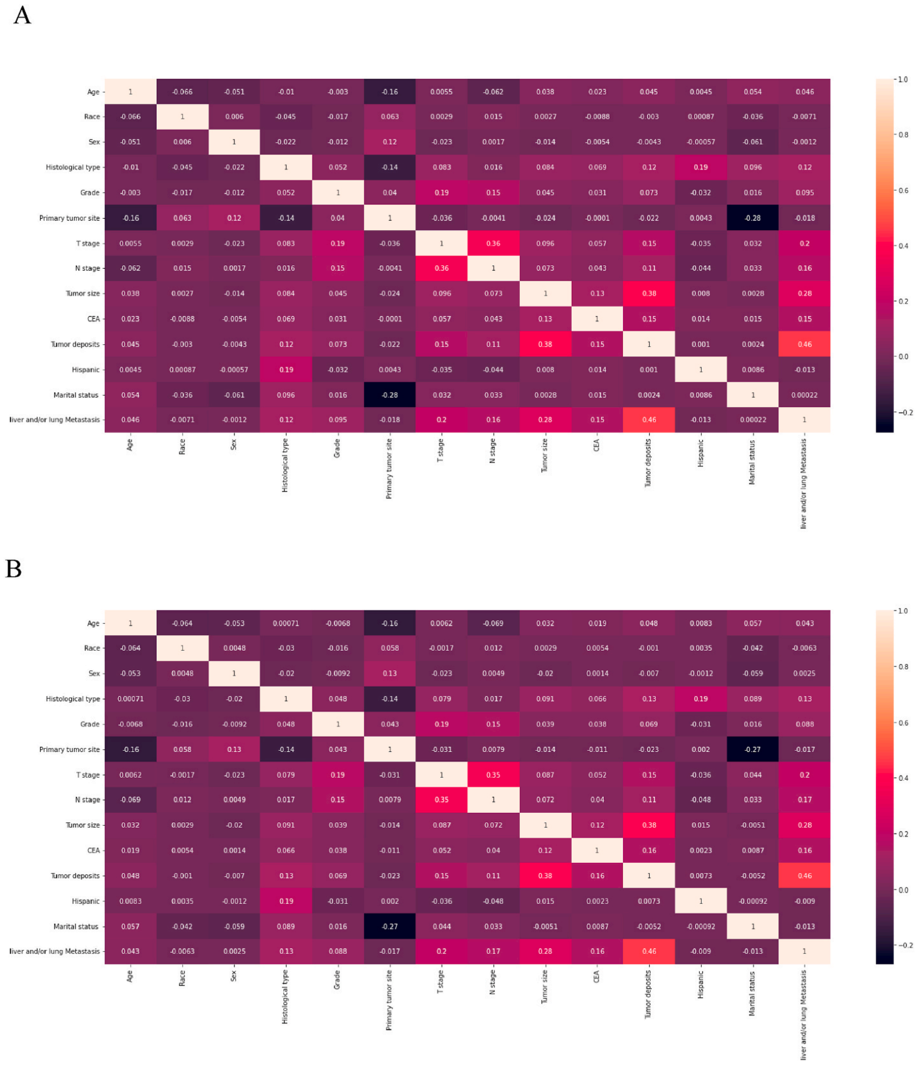

Fig. 2. Correlation heatmaps of patient characteristics features in different datasets (A) Over-sampling data (B) Under-sampling data.

of disease progression during the initial diagnosis of CRC. Especially in the treatment of early CRC patients, metastasis is of great importance to the selection of treatment methods [26]. At present, there are few studies on liver and/or lung metastasis in CRC patients, and most of them focus on the prognostic factors of CRC patients with liver and/or lung metastasis [27–29]. Liver and/or lung metastasis is the most common metastasis. Establishing a reasonable prediction model can effectively help clinicians screen high-risk groups as soon as possible and improve the patients' survival rates and quality of life. To solve this problem, we use a population-based database and its clinical and tumor characteristics to construct a risk prediction model, which shows excellent performance and reliability in identifying liver and/or lung metastasis in CRC patients. To our knowledge, this study is the first to combine machine learning algorithms to predict lung metastasis in CRC patients. The AUC performance of these models is mostly greater than 0.700 (Fig. 3; Table 3). Therefore, we believe that the developed model is robust and reliable, while allowing for more significant clinical benefits. By comparing the predictive performance of seven machine learning algorithms, we found that the model based on the RF algorithm performed the best.

Previous studies have reported that the risk factors affecting the prognosis of CRC patients with liver and/or lung metastasis have been analyzed, and a risk model for the prognosis of CRC patients with liver and/or lung metastasis has been developed [30]. Another study revealed that the expression of seven essential genes could predict the metastasis of colorectal cancer to the liver or lungs, providing clues for exploring the mechanism of target organ selection during the metastasis process of colorectal cancer [31]. Zhao et al. identified risk factors for CRC liver metastasis, including male, black, uninsured, and left colon, and established corresponding predictive models [32].

This study aims to predict liver and/or lung metastasis in CRC patients by establishing a machine-learning model and a traditional LR model. Multivariable analysis showed that the age, histological type, Grade, primary tumor site, T stage, N stage, tumor size, CEA, and tumor deposits were significantly correlated, and all were independent deposits of liver and/or lung metastasis of colorectal cancer. Similar to the results of multivariable analysis, the Feature selection of the RF model suggested that tumor deposition was the critical predictor of metastasis, followed by the location of the primary tumor and CEA level. Tumor deposits refer to the accumulation of tumor cells in the surrounding intestine fat, which is noted in 20%–25 % of colon cancer patients [33]. The AJCC 8th edition (2017) clarifies that TDs should not have any identifiable lymph node, vascular, or neural tissue on histopathological examination [34]. National Comprehensive Cancer Network (NCCN) guidelines suggest that tumor deposits should not be considered metastatic lymph nodes [35], and some studies suggest that tumor deposits should be considered metastatic disease [36]. At present, there are limited studies on tumor deposits, and most of them focus on their

Table 1
Clinical and pathological characteristics of the SEER database set and outer validation set.

Categories	Outer validation (N = 196)	SEER database(N = 51265)	*P* value
Age (years)			0.164
<60	51 (26.0 %)	15692 (30.6 %)	
≥60	145 (74.0 %)	35573 (69.4 %)	
Gender			0.500
Female	88 (44.9 %)	24253 (47.3 %)	
male	108 (55.1 %)	27012 (52.7 %)	
Race			<0.001
white	0 (0 %)	39901 (77.8 %)	
black	0 (0 %)	4601 (9.0 %)	
other	196 (100 %)	6763 (13.2 %)	
Hispanic			<0.001
NO	196 (100 %)	47081 (91.8 %)	
YES	0 (0 %)	4184 (8.2 %)	
Marital status			<0.001
Married	196 (100 %)	21488 (41.9 %)	
Single	0 (0 %)	27176 (53.0 %)	
Unknown	0 (0 %)	2601 (5.1 %)	
Histological type			
Adenocarcinoma	163 (83.2 %)	47077 (91.8 %)	
Mucinous adenocarcinoma	28 (14.3 %)	3570 (7.0 %)	
Signet ring cell carcinoma	5 (2.5 %)	618 (1.2 %)	
Grade			<0.001
Grade I	26 (13.3 %)	3818 (7.5 %)	
Grade II	142 (72.5 %)	33642 (65.6 %)	
Grade III	24 (12.2 %)	7297 (14.2 %)	
Grade IV	1 (0.5 %)	1424 (2.8 %)	
Unknown	3 (1.5 %)	5084 (9.9 %)	
Primary tumor site			0.424
Cecum	12 (6.1 %)	8358 (16.3 %)	
Appendix	1 (0.5 %)	833 (1.6 %)	
Ascending colon	59 (30.1 %)	7304 (14.2 %)	
Hepatic flexure of colon	4 (2.0 %)	1829 (3.5 %)	
Transverse colon	12 (6.1 %)	3779 (7.4 %)	
Splenic flexure of colon	1 (0.5 %)	1165 (2.3 %)	
Descending colon	9 (4.6 %)	2241 (4.4 %)	
Sigmoid colon	38 (19.5 %)	10236 (20.0 %)	
Overlapping lesion of colon	0 (0 %)	423 (0.8 %)	
Colon, NOS	9 (4.6 %)	695 (1.4 %)	
Rectosigmoid junction	3 (1.5 %)	3879 (7.6 %)	
Rectum, NOS	48 (24.5 %)	10526 (20.5 %)	
T stage			<0.001
T1	6 (3.1 %)	9910 (19.3 %)	
T2	17 (8.7 %)	6426 (12.5 %)	
T3	129 (65.8 %)	23265 (45.4 %)	
T4	44 (22.4 %)	8151 (15.9 %)	
TX	0 (0 %)	3513 (6.9 %)	
N stage			0.849
N0	112 (57.1 %)	29691 (57.9 %)	
N1	56 (28.6 %)	13193 (25.7 %)	
N2	28 (14.3 %)	6671 (13.0 %)	
NX	0 (0 %)	1710 (3.4 %)	
Tumor size			0.018
< 5 cm	99 (50.5 %)	25436 (49.6 %)	
≥5 cm	97 (49.5 %)	17526 (34.2 %)	
Unknown	0 (0 %)	8303 (16.2 %)	
CEA			<0.001
Negative	105 (53.6 %)	15928 (31.0 %)	
Borderline	0 (0 %)	93 (0.2 %)	
Positive	58 (29.6 %)	14381 (28.1 %)	
Unknown	33 (16.8 %)	20863 (40.7 %)	
Tumor deposits			<0.001
NO	178 (90.8 %)	35694 (69.6 %)	
Yes	18 (9.2 %)	2839 (5.6 %)	
Unknown	0 (0 %)	12732 (24.8 %)	
liver and/or lung metastasis			0.001
NO	180 (91.8 %)	42303 (82.5 %)	
Yes	16 (8.2 %)	8962 (17.5 %)	

Table 2
The clinical and tumor characteristics of patients with liver and/or lung metastasis and patients without liver and/or lung metastasis.

Categories	With liver and/or lung metastasis (n = 7864)	Without liver and/or lung metastasis (n = 43401)	*P* value
Age (years)			<0.001
<60	2137(27.2 %)	13553(31.2 %)	
≥60	5727(72.8 %)	29848(68.8 %)	
Gender			0.830
Female	3729(47.4 %)	20523(47.3 %)	
male	4135(52.6 %)	22878(52.7 %)	
Race			0.731
white	6124(77.9 %)	33770(77.8 %)	
black	738(9.4 %)	3866(8.9 %)	
other	1002(12.7 %)	5765(13.3 %)	
Hispanic			0.083
NO	7262(92.3 %)	39826(91.8 %)	
YES	602(7.7 %)	3575(8.2 %)	
Marital status			0.391
Married	3389(43.1 %)	18096(41.7 %)	
Single	3985(50.7 %)	23194(53.4 %)	
Unknown	490(6.2 %)	2111(4.9 %)	
Histological type			<0.001
Adenocarcinoma	6678(84.9 %)	40412(93.1 %)	
Mucinous adenocarcinoma	1022(13.0 %)	2537(5.8 %)	
Signet ring cell carcinoma	164(2.1 %)	452(1.1 %)	
Grade			<0.001
Grade I	468(5.9 %)	3350(7.7 %)	
Grade II	4616(58.7 %)	29013(66.8 %)	
Grade III	1437(18.3 %)	5861(13.5 %)	
Grade IV	258(3.6 %)	1134(2.6 %)	
Unknown	1058(13.5 %)	4043(9.4 %)	
Primary tumor site			0.007
Cecum	1431(18.2 %)	6923(16.0 %)	
Appendix	199(2.5 %)	631(1.5 %)	
Ascending colon	1026(13.1 %)	6275(14.5 %)	
Hepatic flexure of colon	297(3.8 %)	1531(3.5 %)	
Transverse colon	596(7.6 %)	3178(7.3 %)	
Splenic flexure of colon	181(2.3 %)	984(2.3 %)	
Descending colon	272(3.5 %)	1969(4.5 %)	
Sigmoid colon	1390(17.7 %)	8843(20.4 %)	
Overlapping lesion of colon	64(0.8 %)	358(0.8 %)	
Colon, NOS	187(2.4 %)	523(1.2 %)	
Rectosigmoid junction	601(7.6 %)	3275(7.5 %)	
Rectum, NOS	1620(20.5 %)	8911(20.5 %)	
T stage			<0.001
T1	962(12.2 %)	8944(20.6 %)	
T2	626(8.0 %)	5796(13.3 %)	
T3	3373(42.9 %)	19858(45.8 %)	
T4	1891(24.0 %)	6260(14.4 %)	
TX	1012(12.9 %)	2543(5.9 %)	
N stage			<0.001
N0	3609(45.8 %)	26060(60.0 %)	
N1	2230(28.4 %)	10950(25.2 %)	
N2	1524(19.4 %)	5146(11.9 %)	
NX	501(6.4 %)	1245(2.9 %)	
Tumor size			<0.001
< 5 cm	2292(29.1 %)	23162(53.3 %)	
≥5 cm	3009(38.3 %)	14526(33.5 %)	
Unknown	2563(32.6 %)	5713(13.2 %)	
CEA			<0.001
Negative	836(10.6 %)	15103(34.8 %)	
Borderline	7(0.1 %)	86(0.2 %)	
Positive	4772(60.7 %)	9617(22.2 %)	
Unknown	2249(28.6 %)	18959(42.8 %)	
Tumor deposits			<0.001
NO	2461(31.3 %)	33262(76.6 %)	
Yes	597(7.6 %)	2245(5.2 %)	
Unknown	4806(61.1 %)	7894(18.2 %)	

Table 3
Risk factors for liver and/or lung metastasis of colorectal cancer patients in multivariable logistic regression.

Factors	OR	95%CI	P-value
Age (years)			
<60	0.877	0.824–0.934	<0.001
≥60	Ref		
Histological type			
Adenocarcinoma	0.909	0.734–1.126	0.383
Mucinous adenocarcinoma	1.318	1.051–1.651	0.017
Signet ring cell carcinoma	Ref		
Grade			
Grade I	0.779	0.676–0.896	<0.001
Grade II	0.830	0.754–0.913	<0.001
Grade III	0.950	0.848–1.064	0.378
Grade IV	0.949	0.793–1.136	0.572
Unknown	Ref		
Primary tumor site			
Cecum	0.936	0.853–1.026	0.158
Appendix	0.968	0.788–1.190	0.760
Ascending colon	0.840	0.762–0.927	0.001
Hepatic flexure of colon	0.938	0.802–1.097	0.425
Transverse colon	0.927	0.823–1.044	0.211
Splenic flexure of colon	0.912	0.751–1.107	0.352
Descending colon	0.752	0.644–0.879	<0.001
Sigmoid colon	0.841	0.769–0.919	<0.001
Overlapping lesion of colon	0.667	0.488–0.910	0.011
Colon, NOS	0.907	0.730–1.126	0.375
Rectosigmoid junction	0.988	0.878–1.111	0.836
Rectum, NOS	Ref		
T stage			
T1	0.505	0.446–0.572	<0.001
T2	0.503	0.437–0.579	<0.001
T3	0.613	0.547–0.688	<0.001
T4	0.841	0.744–0.950	0.006
TX	Ref		
N stage			
N0	0.711	0.615–0.823	<0.001
N1	0.879	0.755–1.023	0.096
N2	1.090	0.927–1.282	0.298
NX	Ref		
Tumor size			
< 5 cm	0.501	0.465–0.541	<0.001
≥5 cm	0.752	0.698–0.810	<0.001
Unknown	Ref		
CEA			
Negative	0.648	0.594–0.706	<0.001
Borderline	0.889	0.399–1.981	0.773
Positive	3.834	3.604–4.078	<0.001
Unknown	Ref		
Tumor deposits			
NO	0.183	0.172–0.194	<0.001
Yes	0.517	0.464–0.576	<0.001
Unknown	Ref		

Table 4
Comparison prediction performances of different models for Over-sampling.

Model	Accuracy	AUC	Precision	Recall rate	F1-score
SVM	0.785	0.857	0.771	0.807	0.788
KNN	0.852	0.890	0.818	0.902	0.819
DT	0.741	0.920	0.698	0.842	0.763
NB	0.737	0.800	0.750	0.705	0.727
RF	0.895	0.956	0.859	0.944	0.899
XGBoost	0.809	0.883	0.795	0.828	0.811
GBM	0.773	0.849	0.770	0.774	0.772

influence on prognosis [37–39]. A study found that tumor deposits have more impact on prognosis in CRC patients than lymph node metastasis [40]. A systematic review and meta-analysis by Nagtegaal et al. [33] found that the combination of TD and LN strongly predicted peritoneal and liver metastasis. Previous studies have shown that TD is an independent predictor that increases the likelihood of metastasis and reduces survival rat [33,41,42]. This study found that tumor deposits have a crucial predictive role in the occurrence of metastasis in CRC patients. Similar to previous studies on liver metastasis, the rectum is more likely to merge with metastasis than the colon [32]. However, other studies have reached the opposite conclusion [43]. It was also believed that the distant metastatic site was related to the primary site. Abdominal metastasis were more frequent in patients with colon cancer, and extra-abdominal metastasis were more common in rectal patients [44]. This might be due to their different embryonic origins and impact on biological habits, resulting in substantially different epidemiological and clinical manifestations [45–48]. More research is needed to verify these views. CEA is a tumor marker on the membrane of colorectal cancer cells. More and more studies have found that the level of CEA before treatment is related to the stage and metastasis of the tumor. Studies have shown that 76 % of metastatic patients are accompanied by elevated CEA [49–52]. Both experimental and clinical evidence suggest that soluble CEA in serum might have an instrumental role in CRC liver metastasis [53–55], which is likewise the same as the results of this study. Patients with high CEA levels are more likely to be associated with metastasis. T staging is an essential criterion of tumor progression, which is positively correlated with most metastasis [14]. Our study also suggested that the higher the T stage, the greater the possibility of metastasis. The lymphatic system is an essential pathway for the metastasis of colorectal cancer. Some studies have shown that the N stage is significantly correlated with lung metastasis in CRC patients [56]. Studies have also found that the liver is one of the organs with the most abundant lymphatic system. The liver is quickly involved when there is regional lymph node metastasis [14], the same result as this study. Patients with poorly differentiated CRC are more prone to metastasis [32], which is consistent with the results of this study. As reported in a study [57], the larger the tumor, the greater the trend of distant metastasis. The age and histological type of tumor do not perform well in the RF model, but they are still significant predictors. This study found that gender is unrelated to distant metastasis in CRC patients, which is consistent with some research findings [27]. However, some studies [58] have raised different opinions, and further research is needed to verify the results.

Currently, most of the traditional statistical methods we use are regression models that assume a linear relationship between variables and results [59]. However, only a small number of variables and results have a linear relationship in the model. In recent years, considerable data processing technologies represented by artificial intelligence algorithms have provided new ways to analyze clinically relevant issues and establish predictive models. At present, it has been extensively used to construct clinical prediction models of different research types. Machine learning algorithm has many advantages, including preventing over-fitting and processing unbalanced data [60]. Machine learning can extract useful information from a large number of data, use it to establish a mathematical model, and further verify the performance of the model in a new data set [61]. Machine learning is widely used in the medical field, including all aspects of disease diagnosis, treatment, and prognosis evaluation of diseases. It is believed that machine learning will play a more crucial role in the medical field with the continuous increase of data and the improvement of algorithms.

Based on the SEER database, we built seven prediction models to predict the risk of liver and/or lung metastasis in CRC patients. We evaluated the seven algorithm models through accuracy, precision, recall rate, F1 score, and AUC value, among which RF has a good prediction (AUC = 0.953), higher than the traditional Logistic regression model (AUC = 0.833). RF was the better model for predicting the risk of liver and/or lung metastasis in CRC patients using the SEER database.

This study also has a few limitations: 1) the external validation cohort was single-center data with a small number of patients who were all Asian. Therefore, more patient data from multiple hospitals will be needed to validate our model's diagnostic efficacy and extrapolation. 2) The accuracy of the model is expected to be further improved, and more risk factors related to metastasis can be included in the future. 3) The

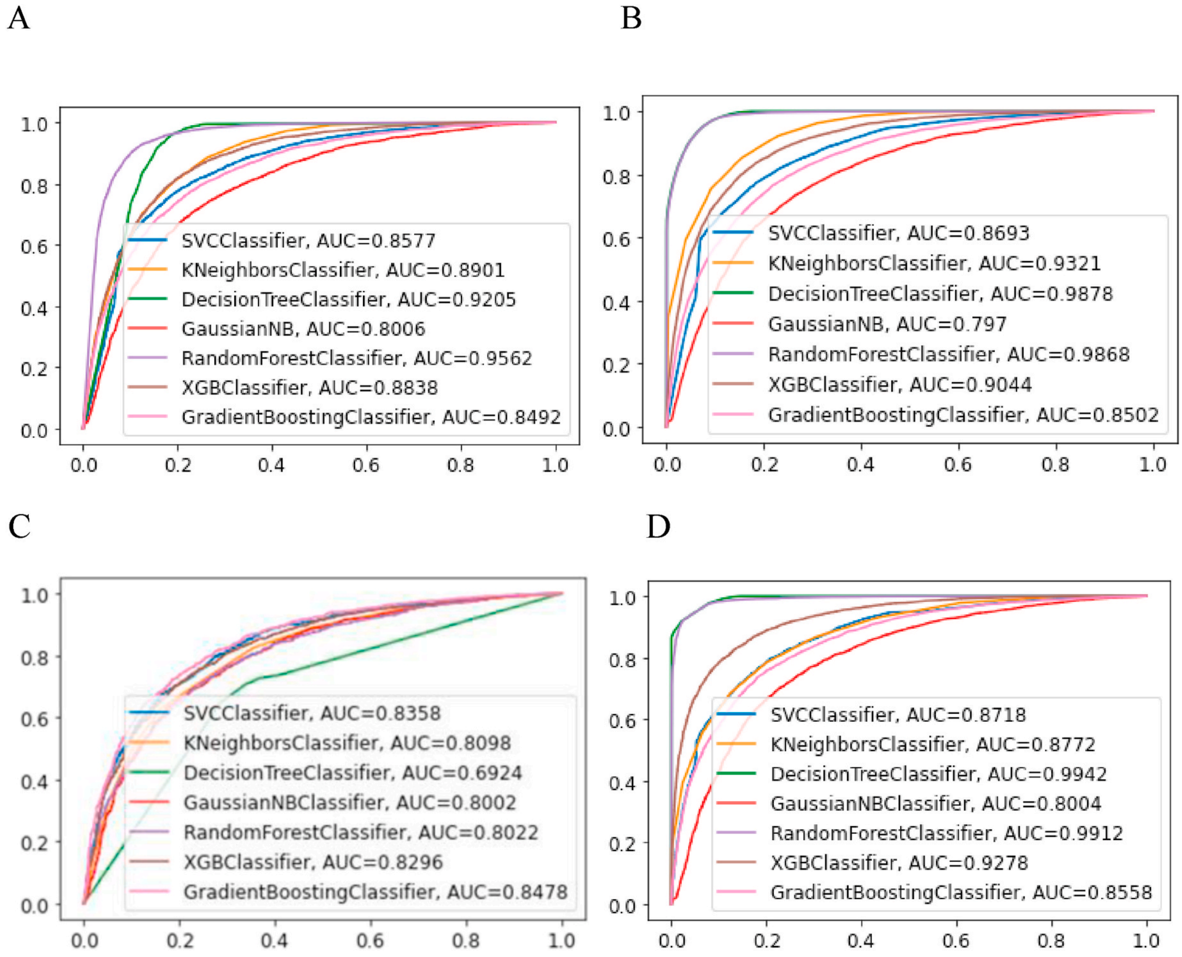

Fig. 3. ROC curves of 7 ML algorithms in different datasets
(A) The ROC curves of the 7 ML algorithms model in the test set with over-sampling. (B) The ROC curves of the 7 ML algorithms model in the training set with over-sampling. (C) The ROC curves of the 7 ML algorithms model in the test set with under-sampling. (D) The ROC curves of the 7 ML algorithms model in the training set with under-sampling.

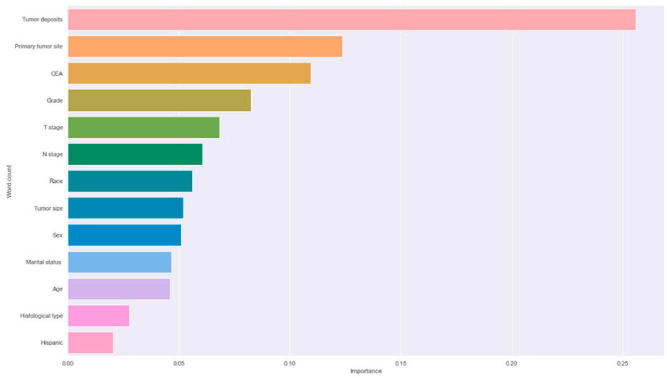

Fig. 4. Feature importance derived from random forest model. The plot shows relative importance of the variables in random forest model.

SEER database lacks vital information such as tumor family history and tumor markers other than CEA, which may also be significant predictors and prognostic indicators of distant metastasis in colorectal cancer, similarly, the SEER database also lacks information on specific treatment plans, such as adjuvant therapy and neoadjuvant therapy, and further analyzes their impact on patient prognosis, and the model does not include regional differentiation in decision making.

This study developed and validated a prediction model based on machine learning algorithms, which utilizes clinical features and quantifies the main factors leading to liver and/or lung metastasis to predict the occurrence of liver and/or lung metastasis in CRC patients. Among them, tumor deposits, primary tumor site, and CEA level are the top three most important factors for liver and/or lung metastasis in CRC patients. Compared with traditional LR models, the RF algorithm has better predictability; therefore, it can provide personalized treatment and more effective allocation of medical resources for patients. This study developed an online network calculator for evaluating the risk of liver and/or lung metastasis in colorectal cancer patients, which can be applied to clinical patients (http://121.43.117.60:8001/).

CRediT authorship contribution statement

Zhentian Guo: Conceptualization, Data curation, Investigation, Software, Writing – original draft, Writing – review & editing. **Zongming Zhang:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Limin Liu:** Data curation, Investigation. **Yue Zhao:** Data curation, Investigation. **Zhuo Liu:** Data curation, Investigation. **Chong Zhang:** Data curation, Investigation. **Hui Qi:** Data curation, Investigation. **Jinqiu Feng:** Data curation, Investigation. **Chunmin Yang:** Writing – review & editing. **Weiping Tai:** Writing – review & editing. **Filippo Banchini:** Supervision, Writing – review & editing. **Riccardo Inchingolo:** Supervision, Writing – review & editing.

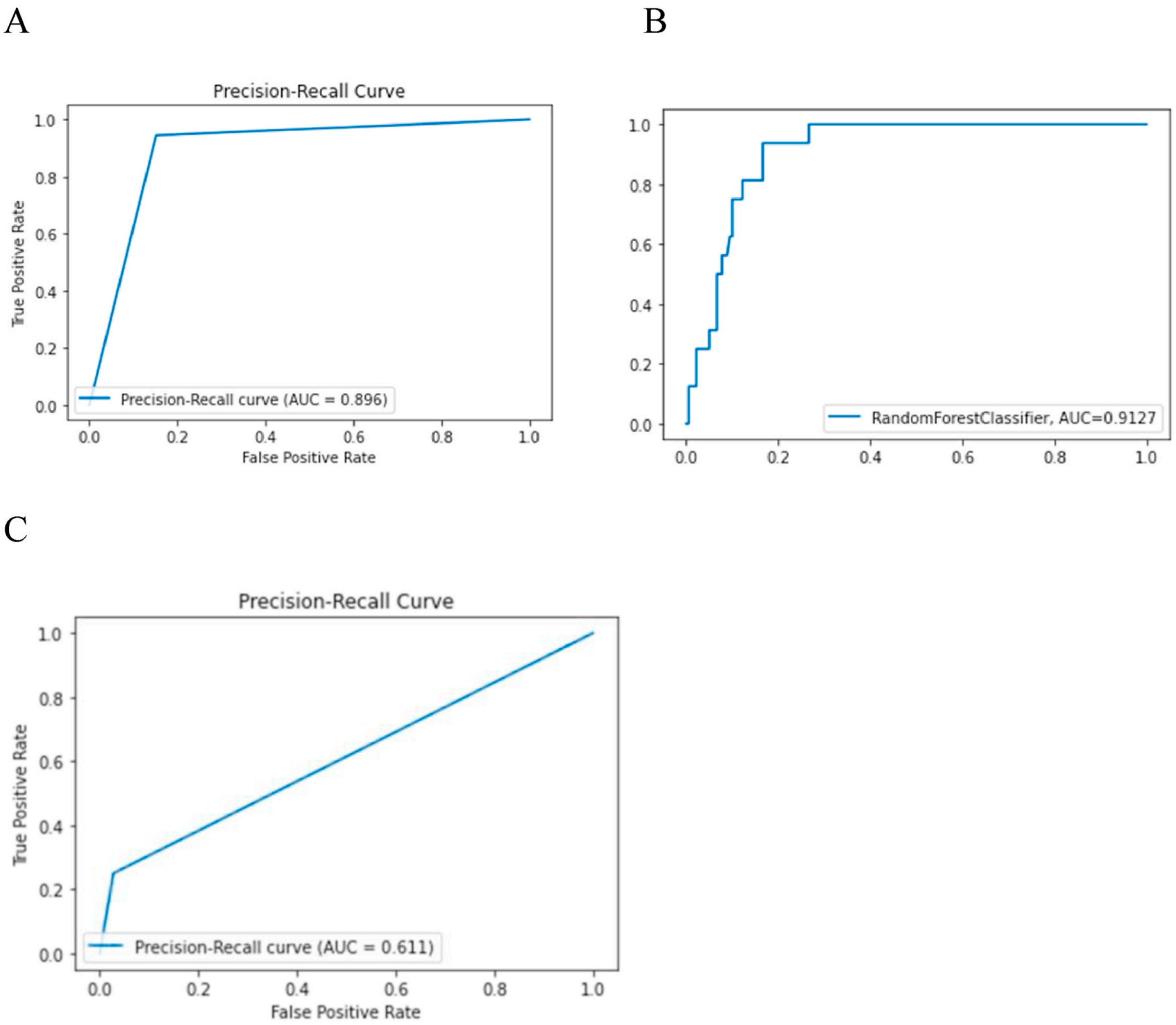

Fig. 5. The random forest algorithm predicts the AUPR curve in the test set and validation set, as well as the AUC curve in the validation set
(A) The PR curves of RF algorithm model in the internal test set. (B) The ROC curves of the RF algorithm model in the outer validation set. (C) The PR curves of RF algorithm model in the outer validation set.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgement

This study was supported by Beijing Municipal Science & Technology Commission (No. Z171100000417056) and the Key Support Project of Guo Zhong Health Care of China General Technology Group (GZKJ-KJXX-QTHT-20230626).

We are very grateful to Dr. Xiangxiang Cui, a medical computer scientist from the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, for evaluating the experimental design and analysis of this article and providing valuable feedback.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.

[2] Kow AWC. Hepatic metastasis from colorectal cancer. J Gastrointest Oncol 2019; 10:1274–98.

[3] Davini F, Ricciardi S, Zirafa CC, Romano G, Alì G, Fontanini G, et al. Lung metastasectomy after colorectal cancer: prognostic impact of resection margin on long term survival, a retrospective cohort study. Int J Colorectal Dis 2020;35:9–18.

[4] Panahi MH, Panahi H, Mahdavi Hezaveh A, Mansournia MA, Bidhendi Yarandi R. Survival rate of colon and rectum cancer in Iran: a systematic review and meta-analysis. Neoplasma 2019;66:988–94.

[5] Wang X, Mao M, Xu G, Lin F, Sun P, Baklaushev VP, et al. The incidence, associated factors, and predictive nomogram for early death in stage IV colorectal cancer. Int J Colorectal Dis 2019;34:1189–201.

[6] Disibio G, French SW. Metastatic patterns of cancers: results from a large autopsy study. Arch Pathol Lab Med 2008;132:931–9.

[7] Akgül Ö, Çetinkaya E, Ersöz Ş, Tez M. Role of surgery in colorectal cancer liver metastasis. World J Gastroenterol 2014;20:6113–22.

[8] Li J, Yuan Y, Yang F, Wang Y, Zhu X, Wang Z, et al. Expert consensus on multidisciplinary therapy of colorectal cancer with lung metastasis (2019 edition). J Hematol Oncol 2019;12:16.

[9] Van Cutsem E, Cervantes A, Adam R, Sobrero A, Van Krieken JH, Aderka D, et al. ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. Ann Oncol 2016;27:1386–422.

[10] Daly MC, Paquette IM. Surveillance, Epidemiology, and End results (SEER) and SEER-Medicare databases: use in clinical research for improving colorectal cancer outcomes. Clin Colon Rectal Surg 2019;32:61–8.

[11] Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 2020;395:350–60.

[12] Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smaïl-Tabbone M, et al. Application of artificial intelligence to Gastroenterology and Hepatology. Gastroenterology 2020;158:76–94.e2.

[13] Achilonu OJ, Fabian J, Bebington B, Singh E, Eijkemans MJC, Musenge E. Predicting colorectal cancer recurrence and patient survival using supervised machine learning Approach: a South African population-based study. Front Public Health 2021;9:694306.

[14] Qiu B, Su XH, Qin X, Wang Q. Application of machine learning techniques in real-world research to predict the risk of liver metastasis in rectal cancer. Front Oncol 2022;12:1065468.

[15] Liu W, Wang S, Ye Z, Xu P, Xia X, Guo M. Prediction of lung metastasis in thyroid cancer using machine learning based on SEER database. Cancer Med 2022;11:2503–15.

[16] Jiang W, Shen Y, Ding Y, Ye C, Zheng Y, Zhao P, et al. A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system. Int J Cancer 2018;142:357–68.

[17] Golpour P, Ghayour-Mobarhan M, Saki A, Esmaily H, Taghipour A, Tajfard M, et al. Comparison of support vector machine, naïve Bayes and logistic regression for Assessing the necessity for coronary angiography. Int J Environ Res Public Health 2020;17.

[18] Cui S, Zhao L, Wang Y, Dong Q, Ma J, Wang Y, et al. Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. Injury 2018;49:1865–70.

[19] Asadikia A, Rajabifard A, Kalantari M. Region-income-based prioritisation of sustainable development goals by gradient boosting machine. Sustain Sci 2022;17:1939–57.

[20] Miyoshi N, Ohue M, Yasui M, Noura S, Shingai T, Sugimura K, et al. Novel prognostic prediction models for patients with stage IV colorectal cancer after concurrent curative resection. ESMO Open 2016;1:e000052.

[21] Van Cutsem E, Cervantes A, Nordlinger B, Arnold D. Metastatic colorectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2014;25(Suppl 3):iii1–9.

[22] Kopetz S, Chang GJ, Overman MJ, Eng C, Sargent DJ, Larson DW, et al. Improved survival in metastatic colorectal cancer is associated with adoption of hepatic resection and improved chemotherapy. J Clin Oncol 2009;27:3677–83.

[23] Engstrand J, Nilsson H, Strömberg C, Jonas E, Freedman J. Colorectal cancer liver metastasis - a population-based study on incidence, management and survival. BMC Cancer 2018;18:78.

[24] Parnaby CN, Bailey W, Balasingam A, Beckert L, Eglinton T, Fife J, et al. Pulmonary staging in colorectal cancer: a review. Colorectal Dis 2012;14:660–70.

[25] Zhang GQ, Taylor JP, Stem M, Almaazmi H, Efron JE, Atallah C, et al. Aggressive multimodal treatment and metastatic colorectal cancer survival. J Am Coll Surg 2020;230:689–98.

[26] Han T, Zhu J, Chen X, Chen R, Jiang Y, Wang S, et al. Application of artificial intelligence in a real-world research for predicting the risk of liver metastasis in T1 colorectal cancer. Cancer Cell Int 2022;22:28.

[27] Liu C, Wang T, Yang J, Zhang J, Wei S, Guo Y, et al. Distant metastasis pattern and prognostic prediction model of colorectal cancer patients based on big data mining. Front Oncol 2022;12:878805.

[28] Liu LL, Sun JD, Xiang ZL. Survival nomograms for colorectal carcinoma patients with lung metastasis and lung-only metastasis, based on the SEER database and a single-center external validation cohort. BMC Gastroenterol 2022;22:446.

[29] Wu Q, Wang WJ, Huang YQ, Fang SY, Guan YJ. Nomograms for estimating survival in patients with liver-only colorectal metastasis: a retrospective study. Int J Surg 2018;60:1–8.

[30] Chen X, Hu W, Huang C, Liang W, Zhang J, Wu D, et al. Survival outcome of palliative primary tumor resection for colorectal cancer patients with synchronous liver and/or lung metastasis: a retrospective cohort study in the SEER database by propensity score matching analysis. Int J Surg 2020;80:135–52.

[31] Tang L, Lei YY, Liu YJ, Tang B, Yang SM. The expression of seven key genes can predict distant metastasis of colorectal cancer to the liver or lung. J Dig Dis 2020;21:639–49.

[32] Tang M, Wang H, Cao Y, Zeng Z, Shan X, Wang L. Nomogram for predicting occurrence and prognosis of liver metastasis in colorectal cancer: a population-based study. Int J Colorectal Dis 2021;36:271–82.

[33] Nagtegaal ID, Knijn N, Hugen N, Marshall HC, Sugihara K, Tot T, et al. Tumor deposits in colorectal cancer: improving the value of modern staging-A systematic review and meta-analysis. J Clin Oncol 2017;35:1119–27.

[34] Khan H, Radomski SN, Siddiqi A, Zhou N, Paneitz DC, Johnston FM, et al. Tumor deposits are associated with a higher risk of peritoneal disease in non-metastatic colorectal cancer patients. J Surg Oncol 2023;127:975–82.

[35] Anderson KC, Alsina M, Atanackovic D, Biermann JS, Chandler JC, Costello C, et al. Multiple myeloma, version 2.2016: clinical practice guidelines in oncology. J Natl ComprCancNetw 2015;13:1398–435.

[36] Mirkin KA, Kulaylat AS, Hollenbeak CS, Messaris E. Prognostic significance of tumor deposits in stage III colon cancer. Ann Surg Oncol 2018;25:3179–84.

[37] Ueno H, Nagtegaal ID, Quirke P, Sugihara K, Ajioka Y. Tumor deposits in colorectal cancer: refining their definition in the TNM system. Ann Gastroenterol Surg 2023;7:225–35.

[38] Pu H, Pang X, Fu J, Zheng R, Chen Y, Zhang D, et al. Significance of tumor deposits combined with lymph node metastasis in stage III colorectal cancer patients: a retrospective multi-center cohort study from China. Int J Colorectal Dis 2022;37:1411–20.

[39] Delattre JF, Cohen R, Henriques J, Falcoz A, Emile JF, Fratte S, et al. Prognostic value of tumor deposits for disease-free survival in patients with stage III colon cancer: a post hoc analysis of the IDEA France Phase III Trial (PRODIGE-GERCOR). J Clin Oncol 2020;38:1702–10.

[40] Brouwer NPM, Nagtegaal ID. Tumor deposits improve staging in colon cancer: what are the next steps? Ann Oncol 2021;32:1209–11.

[41] Moon JY, Lee MR, Ha GW. Prognostic value of tumor deposits for long-term oncologic outcomes in patients with stage III colorectal cancer: a systematic review and meta-analysis. Int J Colorectal Dis 2022;37:141–51.

[42] Basnet S, Lou QF, Liu N, Rana R, Shah A, Khadka M, et al. Tumor deposit is an independent prognostic indicator in patients who underwent radical resection for colorectal cancer. J Cancer 2018;9:3979–85.

[43] Qiu M, Hu J, Yang D, Cosgrove DP, Xu R. Pattern of distant metastasis in colorectal cancer: a SEER based study. Oncotarget 2015;6:38658–66.

[44] Hugen N, van de Velde CJH, de Wilt JHW, Nagtegaal ID. Metastatic pattern in colorectal cancer is strongly influenced by histological subtype. Ann Oncol 2014;25:651–7.

[45] Yahagi M, Okabayashi K, Hasegawa H, Tsuruta M, Kitagawa Y. The worse prognosis of right-sided compared with left-sided colon cancers: a systematic review and meta-analysis. J Gastrointest Surg 2016;20:648–55.

[46] Kalantzis I, Nonni A, Pavlakis K, Delicha EM, Miltiadou K, Kosmas C, et al. Clinicopathological differences and correlations between right and left colon cancer. World J Clin Cases 2020;8:1424–43.

[47] Labadie JD, Savas S, Harrison TA, Banbury B, Huang Y, Buchanan DD, et al. Genome-wide association study identifies tumor anatomical site-specific risk variants for colorectal cancer survival. Sci Rep 2022;12:127.

[48] Livingston AJ, Bailey CE. Invited editorial: does side really matter? Survival analysis among patients with right- versus left-sided colon cancer: a propensity score-adjusted analysis. Ann Surg Oncol 2022;29:9–10.

[49] Wu S, Gu W. Association of T Stage and serum CEA levels in determining survival of rectal cancer. Front Med 2019;6:270.

[50] Becerra AZ, Probst CP, Tejani MA, Aquina CT, González MG, Hensley BJ, et al. Evaluating the prognostic role of elevated preoperative carcinoembryonic antigen levels in colon cancer patients: results from the national cancer database. Ann Surg Oncol 2016;23:1554–61.

[51] Ogata Y, Murakami H, Sasatomi T, Ishibashi N, Mori S, Ushijima M, et al. Elevated preoperative serum carcinoembrionic antigen level may be an effective indicator for needing adjuvant chemotherapy after potentially curative resection of stage II colon cancer. J Surg Oncol 2009;99:65–70.

[52] Quah HM, Chou JF, Gonen M, Shia J, Schrag D, Landmann RG, et al. Identification of patients with high-risk stage II colon cancer for adjuvant therapy. Dis Colon Rectum 2008;51:503–7.

[53] Holt AD, Kim JT, Murrell Z, Huynh R, Stamos MJ, Kumar RR. The role of carcinoembryonic antigen as a predictor of the need for preoperative computed tomography in colon cancer patients. Am Surg 2006;72:897–901.

[54] Aarons CB, Bajenova O, Andrews C, Heydrick S, Bushell KN, Reed KL, et al. Carcinoembryonic antigen-stimulated THP-1 macrophages activate endothelial cells and increase cell-cell adhesion of colorectal cancer cells. Clin Exp Metastasis 2007;24:201–9.

[55] Jessup JM, Samara R, Battle P, Laguinge LM. Carcinoembryonic antigen promotes tumor cell survival in liver through an IL-10-dependent pathway. Clin Exp Metastasis 2004;21:709–17.

[56] Kato Y, Shigeta K, Okabayashi K, Tsuruta M, Seishima R, Matsui S, et al. Lymph node metastasis is strongly associated with lung metastasis as the first recurrence site in colorectal cancer. Surgery 2021;170:696–702.

[57] Chen CH, Hsieh MC, Hsiao PK, Lin EK, Lu YJ, Wu SY. A critical reappraisal for the value of tumor size as a prognostic variable in rectal adenocarcinoma. J Cancer 2017;8:1927–34.

[58] Weyant MJ, Carothers AM, Mahmoud NN, Bradlow HL, Remotti H, Bilinski RT, et al. Reciprocal expression of ERalpha and ERbeta is associated with estrogen-mediated modulation of intestinal tumorigenesis. Cancer Res 2001;61:2547–51.

[59] Bai BL, Wu ZY, Weng SJ, Yang Q. Application of interpretable machine learning algorithms to predict distant metastasis in osteosarcoma. Cancer Med 2023;12:5025–34.

[60] Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30.

[61] Scott IA. Demystifying machine learning: a primer for physicians. Intern Med J 2021;51:1388–400.