Machine Learning-Based Classifiers To Predict Metastasis In Colorectal Cancer Patients

Nasrin Borumandnia

Shahid Beheshti University of Medical Sciences

Abolfazl Akbari

Iran University of Medical Sciences

Mohammad Gramifar

Shahid Beheshti University

Atefeh Talebi (✉ a_talebi5855@yahoo.com)

Iran University of Medical Sciences

Mohamad Amin Pourhoseingholi

Shahid Beheshti University of Medical Sciences

Mazaher Azizpour

Skane University, Ystad Hospital, Department of surgery, orthopaedics, intensive and perioperative care, Ystad, Sweden

Seidamir Pasha Tabaeian

Iran University of Medical Sciences

Meghdad Sedaghat

Shahid Beheshti University of Medical Sciences

Research Article

Keywords: Colorectal cancer, machine learning, metastasis

Posted Date: February 21st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1265319/v2>

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

[Read Full License](#)

Abstract

Background: The increasing prevalence of colorectal cancer (CRC) in the past three decades in Iran has made it as a key public health burden. The goal of the study was to predict the metastasis in CRC patients using machine learning approach.

Methods: This study is focused on 1127 CRC patients, who underwent proper treatments in tertiary Taleghani Hospital. The machine learning approach including Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Regression Tree (RT) and Logistic Regression (LR) were used for the prediction of metastasis in CRC. Receiver operating characteristic (ROC) curve, sensitivity, specificity and AUC were carried out to evaluate the performance of the approaches

Results: Out of 1127 patients, 183(16%) experienced metastasis. In prediction of metastasis, the NN and RF had the most sensitivity. Also the specificity of NN and DT were greater in comparison to other methods. The NN had the highest AUC (0.99%), topmost sensitivity (98.5%) and accuracy (99.6%).

Conclusion: Both tumor stage and size of tumor were regarded as indispensable variables in RF model, whereas the NN generally outperformed other ML-based approaches.

Background

Colorectal cancer (CRC) has been regarded as four most common types of cancer as well as the second leading cause of cancer deaths [1]. Though enhancements have been promising at lowering the incidence of CRC, both morbidity and mortality are still in high level [2]. The CRC has raised globally steady from the 1960s but according to geographical variation, the mortality rates is diversity throughout the world [3]. Machine learning (ML) model becomes an essential tool to identify people at both raised risk of developing CRC and uncover the risk factors for the disease [4].

Machine learning algorithms have revealed magnificent performance in predicting survival cancers and metastasis history. The ML-based approaches are innovative and practical models to effectively predict the Overall Survival (OS) of the CRC patients [5, 6]. ML methods prevent the obstacles in estimating the coefficients and accurate data modelling. Also, ML models can automatically operate noise in a dataset, non-linearity, complex interaction, large sample size and features. In general, ML approaches have been reveal to enhance treatment outcome in cancer research [7, 8].

The machine learning approaches were performed to predict of metastatic relapse in breast cancer in several studies [9, 10]. Also, a large number of epidemiologic studies have been performed to investigate specific hypotheses about risk factors of colorectal cancer [11-13]. These studies include survival analysis such as Cox proportional hazard, time-dependent Cox, cure and other types of survival approaches in clinical dataset. Other investigations have done machine learning models such as decision tree (DT), support vector machine (SVM), artificial neural network (ANN), naïve Bayes (NB) methods.

The goal of historical cohort study is to predict the survival of CRC using supervised machine learning methods, RF and LR methods. Evaluating of both models was performed by AUC (CI), sensitivity, specificity, accuracy.

Methods

Clinical data of 1127 patients who underwent medical treatment for rectal cancer in tertiary Taleghani hospital from 2013 to 2019, were scrutinized to make prediction models through machine learning classifier.

Metastasis was regarded as dependent variable, and also demographic and clinical were as independent variables. Those demographical characteristics included sex, age, education level, smoking, marital status, BMI and clinical factors comprised of the number of involved lymph nodes, grade of tumor, tumor stage, type of treatment, diabetes mellitus. The Ethics Committee of the Iran University of Medical Sciences approved the project (IR.IUMS.REC.1400.459).

Statistical analysis

The characteristics of participants were given by reporting the mean $\pm$ SD for continuous variables and frequency with percentage for categorical ones. The missing data was imputed using model-based imputer methods. The chi-square test was used to assess the association between qualitative variables. The continuous variables were compared using independent sample t test. The significance level for the statistical analysis was considered 0.05. The machine learning approach including Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Regression Tree (RT) and Logistic Regression (LR) were used for the prediction of metastasis in CRC. By 5-fold cross validation strategy, we evaluate the performance of methods. In this regard, the data were randomly split into 80% for the training and 20% for the testing. The Area Under Curve (AUC) of receiver operating characteristic, sensitivity, specificity, and precision was reported for train and test sets. All analyses were implemented using Orange3 software version 3.21.0 and SPSS version 23.

Results

The total number of 1127 registered CRC patients in which 183 (16.2%) patients experience metastasis, were included in this retrospective study. The mean $\pm$ SD age of patients was 53.59 ± 14.35 years, with ranges from 14 to 94 years. Totally, 437 (38.8%) of patients were female. Demographical and clinical information are given in Table 1. According the results in Table 1, BMI, smoking, the number of involved lymph nodes, stage of tumor, tumor size and treatment were the influence factors on metastasis in CRC patients ($P < .05$).

Subsequently, various ML algorithms were applied for the metastasis prediction in CRC patients. Table 2 shows the performances of different ML algorithms in terms of sensitivity, specificity, precision and AUC indices, which were evaluated using 5-fold cross-validation. It can be explained that the performance of

all methods were acceptable to be considered for the clinical research. The most sensitivity and specificity was estimated for the NN algorithm (98.5% and 99.6% respectively). The NN and RF had the most AUC and SVM was in the third rank. In addition, evaluation of the results showed that all methods had good performance in test datasets, too.

The ROC curves are plotted for determining the diagnostic ability of the ML algorithms in Figure 1, separately for test and train datasets. The calibration plot in Figure 2 illustrates the class probabilities against those predicted by the 6 machine learning approaches. These curves reveal the match between the probability predicted by algorithms and actual class probabilities.

Table 1
Demographic and clinical attributes of 1127 patients with CRC

Variables		Total (n=1127)		Metastasis				P- value
				No (n=944)		Yes (n=183)		
Age, mean (SD)		53.59(14.35)		53.67(14.45)		53.20(13.82)		.686
Size, mean (SD)		5.34(3.18)		5.33(3.28)		5.40(2.61)		.805
Sex	Female	437	38.8%	376	86.0%	61	14.0%	.099
	Male	690	61.2%	568	82.3%	122	17.7%	
Marital Status	Married	1050	93.2%	880	83.8%	170	16.2%	.874
	Single	77	6.8%	64	83.1%	13	16.9%	
Education	Illiterate	309	27.4%	251	81.2%	58	18.8%	.334
	Primary school	373	33.1%	311	83.4%	62	16.6%	
	High school	266	23.6%	231	86.8%	35	13.2%	
	University	179	15.9%	151	84.4%	28	15.6%	
BMI	<18	79	7.0%	65	82.3%	14	17.7%	.016
	18-25	641	56.9%	521	81.3%	120	18.7%	
	>25	407	36.1%	358	88.0%	49	12.0%	
Smoking	No	841	74.6%	719	85.5%	122	14.5%	.007
	Yes	286	25.4%	225	78.7%	61	21.3%	
Diabetes	No	1047	92.9%	873	83.4%	174	16.6%	.209
	Yes	80	7.1%	71	88.8%	9	11.3%	
Family history	No	719	63.8%	599	83.3%	120	16.7%	.585
	Yes	408	36.2%	345	84.6%	63	15.4%	
Grade of tumor	Poorly	78	6.9%	62	79.5%	16	20.5%	.116
	Moderately	412	36.6%	336	81.6%	76	18.4%	
	Well	637	56.5%	546	85.7%	91	14.3%	
Number of involved lymph node	N0	531	47.1%	481	90.6%	50	9.4%	<.001
	N1	596	52.9%	463	77.7%	133	22.3%	

Variables		Total (n=1127)		Metastasis				P-value
				No (n=944)		Yes		
						(n=183)		
Tumor stage	I	84	7.5%	82	97.6%	2	2.4%	<.001
	II	427	37.9%	423	99.1%	4	0.9%	
	III	412	36.6%	406	98.5%	6	1.5%	
	IV	204	18.1%	33	16.2%	171	83.8%	
Type of treatment	Other treatments	119	10.6%	80	67.2%	39	32.8%	<.001
	Surgery	1008	89.4%	864	85.7%	144	14.3%	

Table 2

Performance criteria by 5-fold cross validation strategy for ML methods in predicting metastasis

Approach	Dataset	Sensitivity%	Specificity%	Precision%	AUC(CI)
Logistic regression	Train	93.1	97.7	87.1	.98(.97-.99)
	Test	94.2	92.0	77.8	.93(.90-.97)
Neural Network	Train	98.5	99.6	97.7	.99(.99-1.0)
	Test	90.4	92.0	77.0	.94(.91-.97)
Random Forest	Train	94.7	98.8	94.6	.99(.996-.999)
	Test	92.3	92.0	77.4	.94(.91-.98)
Naive Bayes	Train	93.1	97.5	86.5	.96(.94-.98)
	Test	94.2	92.0	77.8	.93(.90-.97)
Regression Tree	Train	89.3	99.5	96.7	.96(.94-.98)
	Test	88.5	93.1	79.3	.91(.86-.96)
Support Vector Machine	Train	93.1	97.7	87.1	.98(.97-1.0)
	Test	94.2	92.5	79.0	.95(.92-.98)

Table 3 demonstrated that the most considerable variable was the tumor stage. Then tumor size, age, type of treatment and number of lymph nodes played indispensable role in predicting metastasis.

Table 3
Variable importance for predicting metastasis in patients with CRC.

Importance of variables	Gini
Stage of tumor	0.214514
Tumor size	0.018229
Age	0.007231
Type of treatment	0.003025
The number of involved lymph nodes	0.002423
BMI	0.002095
Diabetes mellitus	0.001661
Marital status	0.001509
Education	0.000976
Grade of tumor	0.000852
Sex	0.000847
Smoking	0.000294
Family history	0.00024

Discussion

In the present study, the ML models were applied to predict the metastasis in CRC patients. The clinical efficacy of our models were determined through the ROC curve analysis and other indexes for evaluation the predictive performance including sensitivity, specificity and precision. Classifier performance was assessed using the six approaches of machine learning. In our study, however all models had acceptable performance, the NN model had greater predictive efficiency than the others.

A number of modeling techniques in ML-based have been suggested in CRC dataset. Alternative studies have applied DT, SVM, NN, RF and LR [14–17].

A study investigated the prediction of tumor in the TNM staging (tumor, node, and metastasis) stage in colon cancer patients [18]. In the survey, they applied ML techniques, such as random forest, logistic regression, support vector machine, artificial neural network, K-nearest neighbor and adaptive boosting, based on grouping Tumor Aggression Score (TAS) into two categories (>9.8 and <9.8). They concluded that when tumor size only was regarded as a prognostic factor, the random forest model can outperform over other approaches with an accuracy %84, 74% in training and test sets. In our study, we performed six of ML-based approaches in colorectal cancer data. Moreover, neural network and random forest have the

highest sensitivity and neural network and decision tree are considered as the highest values in specificity. In their study, both patients' age and stage of tumor are regarded as more important in neural network and other approaches, respectively. However, tumor stage was considered as essential variable, which was compatible in our study.

Boyne et.al in 2020 predicted early discontinuation of adjuvant chemotherapy among individuals Age > 17 Years, who have colon cancer patients with high stage, using LR and RF models [17]. Their results revealed that time from surgery to chemotherapy initiation and distance from treatment facility seemed to be the most considerable predictor factors. They concluded that RF algorithm may help predict early discontinuation of chemotherapy among colon cancer patients with stage III. In our study, neural network and random forest were of primary and secondary importance. The primary outcome of their study was chemotherapy discontinuation, defined as receipt of, <5 months and >5 months, while metastasis was determined as dependent variable in our study. Also, in their study RF was considered as the better model than LR approach, but all ML-based approaches had ideal performances in our study.

An investigation was done in South Africa based on LR, NB, C5.0, RF, SVM, and ANN algorithms for predictive analytics for recurrence and survival outcomes in CRC patients [16]. Three datasets were regarded as simulated, recurrent and survival data. They surveyed significant variables in all models and compared them using AUC, which evaluated the discriminatory power of the predictive models, which was supported with a threshold (accuracy) metric. Their results demonstrated that all models had AUC greater than 80%, but ANN model was considered as better method with AUC approximately equal to 100%, which was compatible with our study. However, an inconsistent result presented histology and CRC complication were priority in six methods of African study, however, tumor stage was the prime candidate in our study.

A survey was carried out on Indonesian population who suffer from CRC in 4 Hospitals from 2012 to 2015 [19]. Their predictor factors composed of comorbidity, stage of cancer, age, type of treatment, cancer location, gender, metastasis in CRC patients. In the survey, RF algorithm was used in data classification through tree merging by training on sample data. Moreover, the accuracy of those models was justified by the value of classification using AUC. Also, the most essential variables on the survival of colorectal cancer patient were metastasis history, cancer location and gender, respectively. In our investigation, the main variable was metastatic history, whereas the survival of patients with CRC was dependent variable in the Indonesian study. Moreover, in their study, both stage of tumor and age were of last importance, which was inconsistent with our survey.

Conclusion

The results of this study indicated that NN can be the best method among ML-based approaches. In addition, the tumor stage, size of tumor, age, treatment type, number of involved lymph nodes were important variables in RF method, respectively.

Abbreviations

AUC: Area under the curve

BMI: Body mass index

CI: Confidence interval

CRC: Colorectal cancer

CV: Cross-validation

DT: Decision tree

LR: Logistic regression

ML: Machine learning

NB: Naïve Bayes

NN: Neural network

RF: Random forest

ROC: Receiver operating characteristic curve

SVM: Support vector machine

Declarations

Ethical Approval and Consent to participate

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Iran University of Medical Sciences (IR.IUMS.REC.1400.459).

Consent for publication

All authors consent to publication.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no conflict of interest.

Funding

This research received no external funding.

Author Contributions

Methodology and formal analysis, N.B. and A.T.; software, N.B. and A.T.; validation, A.T., N.B.; investigation, N.B. and A.T.; resources, A.T., M.G. and A.A.; responsible for data collection M.A.P.; data curation, N.B. and A.T.; writing-original draft preparation, A.T., N.B.; writing-review and editing, A.T., N.B.; visualization, N.B., A.T. and M.A.; supervision, A.T., N.B. and A.A.; All authors have read and agreed to the published version of the manuscript.

Acknowledgement

We thanked the Iran University of Medical Sciences.

Authors' information

¹Urology and Nephrology Research Center, Shahid Beheshti University of Medical Sciences. ²Assistant Professor of Molecular Medicine, Colorectal Research Center, Iran University of Medical Sciences, P.O. Box 1445613131. ³Department of computer and Data Science, Faculty of mathematical Sciences, Shahid Beheshti University. ⁴Biostatistician, Colorectal Research Center, Iran University of Medical Sciences, P.O. Box 1445613131. ⁵Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences. ⁶Skane University, Ystad Hospital, Department of surgery, orthopaedics, intensive and perioperative care, Ystad, Sweden. ⁷Gastroenterology and Hepatology, Colorectal Research Center, Iran University of Medical Sciences. ⁸Department of Internal Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences.

References

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics*, 2016. CA: a cancer journal for clinicians, 2016. **66**(1): p. 7-30.
2. Wieszczy, P., et al., *Colorectal cancer incidence and mortality after removal of adenomas during screening colonoscopies*. Gastroenterology, 2020. **158**(4): p. 875–883. e5.
3. Ferlay, J., et al., *Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods*. International journal of cancer, 2019. **144**(8): p. 1941–1953.
4. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction*. Computational and structural biotechnology journal, 2015. **13**: p. 8–17.
5. Manilich, E.A., et al., *A novel data-driven prognostic model for staging of colorectal cancer*. Journal of the American College of Surgeons, 2011. **213**(5): p. 579–588. e2.

6. Wen, R., et al., *Machine learning-based random forest predicts anastomotic leakage after anterior resection for rectal cancer*. Journal of gastrointestinal oncology, 2021. **12**(3): p. 921.
7. Greener, J.G., et al., *A guide to machine learning for biologists*. Nature Reviews Molecular Cell Biology, 2021: p. 1–16.
8. Zhou, C.-M., et al., *Machine learning predicts lymph node metastasis of poorly differentiated-type intramucosal gastric cancer*. Scientific Reports, 2021. **11**(1): p. 1–7.
9. Nicolò, C., et al., *Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer*. JCO clinical cancer informatics, 2020. **4**: p. 259–274.
10. Tapak, L., et al., *Prediction of survival and metastasis in breast cancer patients using machine learning classifiers*. Clinical Epidemiology and Global Health, 2019. **7**(3): p. 293–299.
11. Borumandnia, N., et al., *Nomogram to predict the overall survival of colorectal cancer patients: a multicenter national study*. International journal of environmental research and public health, 2021. **18**(15): p. 7734.
12. Talebi, A., et al., *Biological and clinical relevance of long non-coding RNA PCAT-1 in cancer, a systematic review and meta-analysis*. Asian Pacific journal of cancer prevention: APJCP, 2019. **20**(3): p. 667.
13. Talebi, A., et al., *Survival analysis in gastric cancer: a multi-center study among Iranian patients*. BMC surgery, 2020. **20**(1): p. 1–8.
14. Cueto-López, N., et al., *A comparative study on feature selection for a risk prediction model for colorectal cancer*. Computer methods and programs in biomedicine, 2019. **177**: p. 219–229.
15. Nartowt, B.J., et al., *Robust machine learning for colorectal cancer risk prediction and stratification*. Frontiers in big Data, 2020. **3**: p. 6.
16. Achilonu, O.J., et al., *Predicting Colorectal Cancer Recurrence and Patient Survival Using Supervised Machine Learning Approach: A South African Population-Based Study*. Frontiers in Public Health, 2021. **9**.
17. Boyne, D.J., et al., *Development of a Model for Predicting Early Discontinuation of Adjuvant Chemotherapy in Stage III Colon Cancer*. JCO Clinical Cancer Informatics, 2020. **4**: p. 972–984.
18. Gupta, P., et al., *Prediction of colon cancer stages and survival period with machine learning approach*. Cancers, 2019. **11**(12): p. 2007.
19. Anuraga, G. and J. Fernanda. *Random forest prognostic factor in colorectal cancer*. in *Journal of Physics: Conference Series*. 2019. IOP Publishing.

Figures

Figure 1

ROC curves for different ML algorithms (left) Train data, (right) Test data

Figure 2

Calibration plots of class probabilities against those predicted by the ML algorithms