RESEARCH ARTICLE

Editorial Process: Submission:10/13/2023 Acceptance:01/19/2024

Classification and Diagnostic Prediction of Colorectal Cancer Mortality Based on Machine Learning Algorithms: A Multicenter National Study

Gohar Mohammadi[1]*, Mehdi Azizmohammad Looha[2], Mohammad Amin Pourhoseingholi[3], Mostafa Rezaei Tavirani[4], Samaneh Sohrabi[5], Amirali Zareie Shab Khaneh[6], Hassan Piri[7], Maryam Alaei[8], Naser Parvani[5], Iman Vakilzadeh[5], Sara javadi[9], Zeynab Moradian Haft Cheshmeh[10], Zahra Razzaghi[11], Reza Mahmoud Robati[12], Mona Zamanian Azodi[4], Saba Zarean Shahraki[13], Raheleh Talebi[14], Jamshid Charati Yazdani[15], Mohammad Esmaeil Motlagh[16], Soheila Khodakarim[17], Melika Hadavi[6]

Abstract

Introduction: Colorectal cancer (CRC) ranks as the second leading cause of cancer-related deaths. This study aimed to predict survival outcomes of CRC patients using machine learning (ML) methods. **Material and Methods:** A retrospective analysis included 1853 CRC patients admitted to three prominent tertiary hospitals in Iran from October 2006 to July 2019. Six ML methods, namely logistic regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Neural Network (NN), Decision Tree (DT), and Light Gradient Boosting Machine (LGBM), were developed with 10-fold cross-validation. Feature selection employed the Random Forest method based on mean decrease GINI criteria. Model performance was assessed using Area Under the Curve (AUC). **Results:** Time from diagnosis, age, tumor size, metastatic status, lymph node involvement, and treatment type emerged as crucial predictors of survival based on mean decrease GINI. The NB (AUC = 0.70, 95% Confidence Interval [CI] 0.65–0.75) and LGBM (AUC = 0.70, 95% CI 0.65–0.75) models achieved the highest predictive AUC values for CRC patient survival. **Conclusions:** This study highlights the significance of variables including time from diagnosis, age, tumor size, metastatic status, lymph node involvement, and treatment type in predicting CRC survival. The NB model exhibited optimal efficacy in mortality prediction, maintaining a balanced sensitivity and specificity. Policy recommendations encompass early diagnosis and treatment initiation for CRC patients, improved data collection through digital health records and standardized protocols, support for predictive analytics integration in clinical decisions, and the inclusion of identified prognostic variables in treatment guidelines to enhance patient outcomes.

Keywords: Colorectal cancer- data mining- feature selection- mortality prediction- machine learning algorithms

Asian Pac J Cancer Prev, **25 (1)**, 333-342

*[1]Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [2]Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [3]Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [4]Proteomics Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [5]Vice Chancellor in Administration and Resources Development Affairs, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [6]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. [7]Shahid Beheshti University of Medical Sciences, Tehran, Iran. [8]Cardiovascular Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [9]Vice Chancellor for Research & Technology, Shiraz University of Medical Sciences, Shiraz, Iran. [10]Department of Epidemiology, Faculty of Health, Iran University of Medical Science,Tehran, Iran. [11]Laser Application in Medical Sciences Research Center. Shahid Beheshti University of Medical Sciences, Tehran, Iran. [12]Department of Dermatology, Director of Skin Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [13]Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [14]Department of Mathematics at Architecture and Computer Engineering, University of Applied Sciences (unit 10), Tehran, Iran. [15]Health Sciences Research Center, Mazandaran University of Medical Sciences, Sari, Iran. [16]Department of Pediatrics, Faculty of Medicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. [17]Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran. *For Correspondence: g.mohammadi@sbmu.ac.ir*

Introduction

The global incidence of cancer has witnessed a steady rise over the years, attributable to a constellation of factors encompassing unhealthy dietary habits, obesity, genetic predisposition, and advancing age [1]. According to the GLOBOCAN 2020 report, CRC accounted for approximately 1.93 million new cases worldwide, constituting 10% of global cancer incidence. Furthermore, the disease led to 0.94 million deaths, corresponding to 9.4% of cancer-related mortalities in 2020 [2]. In Iran, CRC ranked as the fourth most diagnosed cancer among men and the second among women in 2020, following lung and breast cancer, respectively [3]. The imperative of early detection for effective intervention and improved CRC outcomes is evident.

To address this critical need, establishing a CRC monitoring system that regularly screens individuals based on their risk factors becomes paramount to enhance early-stage prognosis accuracy. In pursuit of this objective, researchers have increasingly turned to predictive methods, with data mining and machine learning (ML) approaches taking center stage [4-8]. The applicability of ML extends beyond CRC and includes the prediction of survival outcomes in various cancer types, leading to several comparative studies among a subset of these methodologies [9, 10, 1, 11, 12].

Given the pronounced mortality rate of CRC in Iran and its escalating global trend, employing ML techniques for early-stage disease prediction holds promise in improving patient survival rates. Despite the considerable scientific attention toward predicting CRC survival using ML approaches [13-15, 5], none of these studies have reported the 95% confidence interval (CI) for their methods. The inclusion of CIs in ML algorithms holds significance for debugging, facilitating accurate performance assessment and comparison, conveying precision and uncertainty, and estimating true errors and generalization capabilities [16]. Consequently, the aim of this research was to construct a data analysis framework through a comparative assessment of the supervised ML algorithms in terms of accuracy, precision, and sensitivity, accompanied by a 95% CI. This was achieved using a nationwide multicenter database to enhance the precision of early CRC detection. Additionally, we employed selection techniques to identify the optimal feature subset.

Materials and Methods

Study design

This retrospective study was designed to predict survival outcome among patients diagnosed with CRC through the application of machine learning models.

Study setting

The investigation was carried out across three prominent tertiary hospitals in Iran: Imam Khomeini Hospital of Mazandaran, Taleghani Hospital in Tehran, and Shahid Faghihi Hospital in Shiraz. A span of nearly 13 years, commencing from October 2006 and concluding in July 2019, served as the temporal framework for data collection. The study encompassed patients who received a CRC diagnosis within this stipulated timeframe. Comprehensive patient data were meticulously sourced from medical records, followed by a subsequent follow-up protocol employing telephonic communication to ascertain the mortality status of each participant.

Participants

Enrollment in the study encompassed patients diagnosed with diverse histological subtypes of CRC across all stages as defined by the TNM classification system. However, exclusions were made for individuals with inaccessible medical records, non-Iranian residency status, Iranian individuals residing abroad, CRC patients admitted prior to or subsequent to the designated study duration, and cases wherein death declarations were inaccurately recorded.

Variable

The principal outcome variable under scrutiny was the survival outcome (deceased versus survived). The primary explanatory variables encompassed age at diagnosis, tumor size (centimeters), hospital of admission (Taleghani, Sari, and Shiraz), gender (male or female), and marital status (married or other). Supplementary covariates included Body Mass Index (BMI) categories (<18.5, 18.6-24.9, 25-29.9, and >30), Nutritional Index (NI) categories (<18, 18-25, and >25), smoking status ("No" or "Yes"), educational attainment (Illiterate, Primary school, High school, University), hypertension status ("No" or "Yes"), diabetes mellitus status ("No" or "Yes"), and family history of cancer ("No" or "Yes"). Additionally, CRC site (topography) categorization (Right colon, Left colon, Rectum, Transverse), tumor grade classification (Well differentiated, Moderately differentiated, Poorly differentiated), Pathologic Primary Tumor (T0, T1, T2, T3, and T4), lymph node involvement (N0, N1, N2), metastasis status ("No," "Yes," or "Not known"), CRC stage (I, II, III, IV), treatment type (Surgery, Chemotherapy & radiography & immunotherapy), Familial Adenomatous Polyposis status ("No" or "Yes"), Hereditary Nonpolyposis Colorectal Cancer status ("No" or "Yes"), Inflammatory Bowel Disease status ("No" or "Yes"), and Personal History of CRC ("No" or "Yes") were analyzed.

Data sources/measurement

The data utilized for this study were meticulously extracted from the medical records of the enrolled patients. Information pertaining to survival outcomes, treatment modalities, and diverse clinical attributes were sourced from hospital records. To ensure uniformity in evaluation methods, all participants received diagnoses and treatments at the same tertiary medical institutions, adhering to standardized medical protocols. Furthermore, patient particulars were assessed through telephone interviews conducted by the Colorectal Research Centers, further enhancing data accuracy and completeness.

Bias

In this retrospective study on survival estimates

for CRC patients, a comprehensive strategy addressed potential biases. Inclusion and exclusion criteria were meticulously defined, encompassing diverse CRC patients from three respected tertiary hospitals in Iran. Data collection involved robust examination of medical records and thorough telephone interviews to enhance data accuracy. Standardized medical protocols minimized treatment variability. A wide range of covariates allowed exploration of confounding variables. Conducting the study across three distinct hospitals reduced institutional bias. Adoption of different ML models facilitated comparative analysis. Transparent data analysis and external review further enhanced credibility. Despite these measures, potential biases persist due to the retrospective approach and possible incomplete data, yet methodical data collection and rigorous analysis bolster study resilience.

Study size

This study consisted of 1853 individuals diagnosed with CRC and admitted to three prominent tertiary hospitals in Iran within the timeframe spanning October 2006 to July 2019. This investigation embraced a comprehensive approach, encompassing the entire eligible patient population throughout the study duration, thus precluding the need for sampling.

Statistical methods

A presentation of descriptive statistics ensued by categorizing patients based on vital followed by a comparison of their respective characteristics. Qualitative data were depicted in terms of frequencies (expressed as percentages), while quantitative data were represented by their mean ± standard deviation or median (interquartile range [IQR]). The analytical processes were executed utilizing R software (version 4.2.1), wherein statistical significance was indicated by P-values of 0.05, coupled with a 95% confidence interval (CI).

Machine learning algorithm

The study population was divided randomly into two distinct samples: 70% were employed as training data for outcome prediction, and the remaining 30% constituted validation data for algorithm testing. Each patient was uniquely assigned to either the training or validation sample. Leveraging demographic, clinical, and laboratory variables, four machine learning models, logistic regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Neural Network (NN), Decision Tree (DT), and Light GBM (LGBM) were developed [1-3]. Tuning of each algorithm's parameters was executed to optimize outcome risk prediction accuracy.

Performance evaluation

In the validation dataset, a 10-fold cross-validation method and Receiver Operating Characteristic (ROC) analysis were employed to evaluate the six models. The evaluation encompassed sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), and accuracy (ACC) [4]. Model performance was gauged utilizing the Area Under the Curve (AUC) to quantify discrimination, and the optimal predictive model was selected for clinical decision-making based on performance outcomes.

Variable importance and variable selection

A hybrid approach encompassing both statistical techniques and clinical considerations was undertaken for variable selection. The mean decrease Gini (MDG) method was harnessed within the context of random forest analysis to identify pivotal variables, aggregating the cumulative reduction in Gini impurity during tree node splits [17-19]. Concurrently, clinical variables deemed unrelated were omitted from initial random forest analysis, aligning with clinical perspectives.

Nomogram analysis

To facilitate the creation of a user-friendly predictive model yielding numerical probabilities of fibrosis incidence [20], a nomogram was employed, serving as a graphical representation of the statistical predictive framework.

Results

Descriptive Statistics

The descriptive statistics were reported in Table 1. A total 1,873 patients were included in the study with mean age of 55.20 ± 14.55 years. Hospital location demonstrated significant variations in vital status distribution ($p < 0.001$). Tumor size, survival time, hospital location, BMI categories, marital status, familial history of cancer, site topography, grade, pathologic primary tumor, lymph node involvement, metastasis, and stage all exhibited statistically significant associations with vital status.

Variable Importance Analysis: Unveiling Key Predictors for Outcome Prediction in CRC Patients

Figure 1 displays mean decrease GINI values, revealing each variable's contribution to predictive performance. "Time from diagnosis (months)" had the highest GINI value of 32.18, indicating its strong predictive impact. "Age at Diagnosis" followed with a substantial GINI value of 22.41, emphasizing its significance. "Tumor size (cm)", Metastasis = Yes", "Lymph node involved = N2", "Types of treatment", "Stage = IV" and "Grade = Poorly differentiated" had high GINI values, underlining their meaningful contributions. Variables with GINI values above the average (4.85) were selected to emphasize their higher influence on the predictive model.

Performance Comparison of Machine Learning Models for Mortality Prediction in CRC Patients

As the next step, we aimed to predict the occurrence of death using various machine learning models using logistic regression with all variables, logistic regression, SVM, NB, NN, DT, and LGBM with selected variables. The results of the performance comparison are summarized in Table 2. When considering SE, SP, PPV, NPV, and ACC, some patterns emerge. The NB model achieves the highest AUC value of 0.70, indicating good discriminatory ability.

Table 1. Descriptive Statistics and Comparative Analysis of Patient Variables and Vital Status

Variable	Levels	Total (n=1873)	Vital status		P-value
			Survived or discharged (n=1396)	Deceased (n=492)	
Age at diagnosis	-----	55.20 ± 14.55	54.82 ± 14.22	56.24 ± 15.43	0.064
Tumor size (cm)	-----	4.81 ± 3.11	4.71 ± 2.89	5.08 ± 3.65	0.023
Survival time	-----	23.98 (10.03, 40.39)	25.33 (9.81, 41.74)	19.37 (10.61, 35.08)	<0.001
Hospital	Taleghani	1127 (60.17)	889 (64.61)	236 (47.97)	<0.001
	Sari	219 (11.69)	111 (8.07)	108 (21.95)	
	Shiraz	527 (28.14)	376 (27.33)	148 (30.08)	
Sex	Male	1112 (59.37)	812 (59.01)	297 (60.37)	0.63
	Female	761 (40.63)	564 (40.99)	195 (39.63)	
Marital status	Married	1777 (94.87)	1315 (95.57)	457 (92.89)	0.024
	Other	96 (5.13)	61 (4.43)	35 (7.11)	
BMI – four categories	<18.5	140 (7.47)	80 (5.81)	58 (11.79)	<0.001
	18.6-24.9	1024 (54.67)	748 (54.36)	274 (55.69)	
	25-29.9	555 (29.63)	421 (30.60)	133 (27.03)	
	>30	154 (8.22)	127 (9.23)	27 (5.49)	
BMI – three categories	<18	145 (7.74)	83 (6.03)	60 (12.20)	<0.001
	18-25	1019 (54.40)	745 (54.14)	272 (55.28)	
	>25	709 (37.85)	548 (39.83)	160 (32.52)	
Smoking	No	1340 (71.54)	981 (71.29)	354 (71.95)	0.816
	Yes	533 (28.46)	395 (28.71)	138 (28.05)	
Education	Illiterate	529 (28.24)	380 (27.62)	148 (30.08)	0.286
	Primary school	614 (32.78)	459 (33.36)	153 (31.10)	
	High school	401 (21.41)	305 (22.17)	96 (19.51)	
	University	329 (17.57)	232 (16.86)	95 (19.31)	
Hypertension	No	1666 (88.95)	1216 (88.37)	447 (90.85)	0.153
	Yes	207 (11.05)	160 (11.63)	45 (9.15)	
Diabetes	No	1684 (89.91)	1232 (89.53)	447 (90.85)	0.434
	Yes	189 (10.09)	144 (10.47)	45 (9.15)	
Familial history of cancer	No	1208 (64.50)	865 (62.86)	338 (68.70)	0.021
	Yes	665 (35.50)	511 (37.14)	154 (31.30)	
Site topography	Right colon	576 (30.75)	421 (30.60)	154 (31.30)	0.02
	Left colon	1094 (58.41)	822 (59.74)	268 (54.47)	
	Rectum	183 (9.77)	122 (8.87)	61 (12.40)	
	Transverse	20 (1.07)	11 (0.80)	9 (1.83)	
Grade	Well differentiated	1042 (55.63)	819 (59.52)	221 (44.92)	<0.001
	Moderately differentiated	677 (36.15)	470 (34.16)	205 (41.67)	
	Poorly differentiated	154 (8.22)	87 (6.32)	66 (13.41)	
Pathologic primary tumor	T0	1074 (57.34)	837 (60.83)	232 (47.15)	<0.001
	T1	343 (18.31)	214 (15.55)	129 (26.22)	
	T2	143 (7.63)	113 (8.21)	30 (6.10)	
	T3	292 (15.59)	202 (14.68)	90 (18.29)	
	T4	21 (1.12)	10 (0.73)	11 (2.24)	
Lymph node involved	N0	924 (49.33)	731 (53.13)	191 (38.82)	<0.001
	N1	798 (42.61)	578 (42.01)	217 (44.11)	
	N2	151 (8.06)	67 (4.87)	84 (17.07)	
Metastasis	No	1091 (58.25)	857 (62.28)	230 (46.75)	<0.001
	Yes	284 (15.16)	148 (10.76)	136 (27.64)	
	Not known	498 (26.59)	371 (26.96)	126 (25.61)	
Stage	I	281 (15.00)	227 (16.50)	52 (10.57)	<0.001
	II	681 (36.36)	526 (38.23)	154 (31.30)	
	III	631 (33.69)	460 (33.43)	169 (34.35)	
	IV	280 (14.95)	163 (11.85)	117 (23.78)	

Table 1. Continued

Variable	Levels	Total (n=1873)	Vital status		P-value
			Survived or discharged (n=1396)	Deceased (n=492)	
First treatment	Surgery	1336 (71.33)	1051 (76.38)	281 (57.11)	<0.001
	Chemotherapy & Radiation Therapy & Immunotherapy	537 (28.67)	325 (23.62)	211 (42.89)	
FAP	No	1858 (99.20)	1363 (99.06)	490 (99.59)	0.379
	Yes	15 (0.80)	13 (0.94)	2 (0.41)	
HNPCC	No	1705 (91.03)	1254 (91.13)	446 (90.65)	0.783
	Yes	168 (8.97)	122 (8.87)	46 (9.35)	
IBD	No	1850 (98.77)	1361 (98.91)	484 (98.37)	0.347
	Yes	23 (1.23)	15 (1.09)	8 (1.63)	
Personal history of CRC	No	1764 (94.18)	1296 (94.19)	464 (94.31)	1
	Yes	109 (5.82)	80 (5.81)	28 (5.69)	
Status	Survived or discharged	1376 (73.47)	1376 (100.00)	0 (0.00)	------
	Deceased	492 (26.27)	0 (0.00)	492 (100.00)	
	NA	5 (0.27)	0	0	

Note: Descriptive statistics were reported in terms of frequency (%) for categorical data. Symmetric and asymmetric numeric data were expressed as mean ± standard deviation (SD) and median (interquartile range [IQR]), respectively. The relationship between categorical variables and vital status was investigated using the Fisher exact test. Additionally, the differences in symmetric and asymmetric variables between the deceased and survived groups were assessed using the independent t-test and Mann-Whitney U test, respectively.

This model also shows balanced sensitivity and specificity values, with a SE of 0.60 and SP of 0.73. The NB model demonstrates a PPV of 0.45 and a NPV of 0.83, suggesting its effectiveness in correctly classifying both positive and negative outcomes. The LO model with selected variables and the SVM model both exhibit similar AUC values of 0.68. It is worth highlighting the LGBM model, which demonstrates an AUC value of 0.70 as well. This signifies its proficiency in outcome discrimination, akin to the NB model. However, the LGBM model also showcases

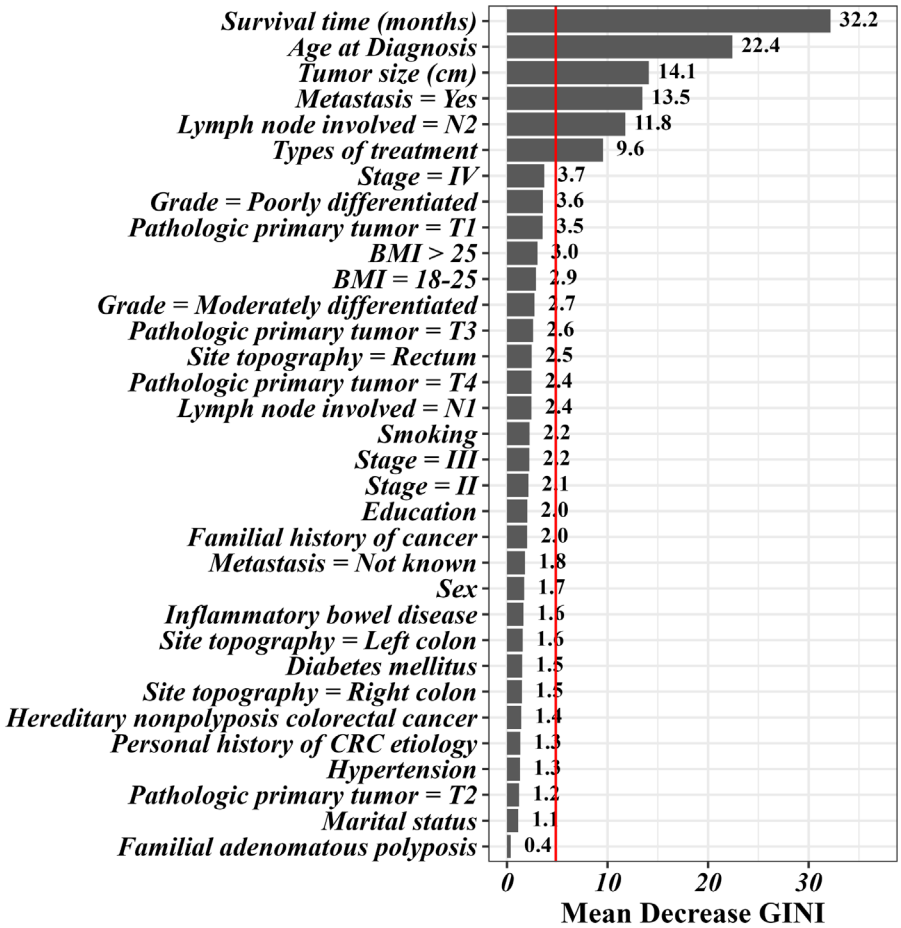

Figure 1. Assessment of Variable Importance Using Mean Decrease GINI

Table 2. Performance Comparison of Machine Learning Models for Death Prediction

Variables	AUC (95% CI)	SE (95% CI)	SP (95% CI)	PPV (95% CI)	NPV (95% CI)	ACC (95% CI)
Logistic regression - all variables (LO-A)	0.49 (0.46, 0.51)	1.00 (0.98, 1.00)	0.00 (NA, 0.01)	0.28 (0.00, 1.00)	NA (0.00, 1.00)	0.28 (0.25, 0.32)
Logistic regression - selected variables (LO-S)	0.68 (0.62, 0.73)	0.53 (0.45, 0.61)	0.75 (0.70, 0.79)	0.43 (0.38, 0.52)	0.81 (0.76, 0.85)	0.69 (0.65, 0.73)
Support Vector Machine (SVM)	0.68 (0.63, 0.74)	0.58 (0.50, 0.66)	0.70 (0.66, 0.75)	0.42 (0.37, 0.50)	0.82 (0.77, 0.85)	0.68 (0.64, 0.71)
Naïve Bayes (NB)	0.70 (0.65, 0.75)	0.60 (0.51, 0.68)	0.73 (0.69, 0.78)	0.45 (0.40, 0.54)	0.83 (0.78, 0.86)	0.70 (0.66, 0.73)
Neural Network (NN)	0.48 (0.43, 0.54)	0.19 (0.13, 0.27)	0.86 (0.82, 0.89)	0.33 (0.27, 0.43)	0.75 (0.65, 0.80)	0.68 (0.64, 0.72)
Decision Tree (DT)	0.60 (0.56, 0.64)	0.27 (0.20, 0.35)	0.91 (0.88, 0.94)	0.53 (0.44, 0.62)	0.77 (0.70, 0.83)	0.75 (0.71, 0.78)
Light GBM (LGBM)	0.70 (0.65, 0.75)	0.75 (0.67, 0.82)	0.56 (0.51, 0.61)	0.39 (0.34, 0.48)	0.86 (0.80, 0.88)	0.61 (0.57, 0.65)

This table presents the area under the curve (AUC), sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), and accuracy (ACC) of various machine learning models applied to outcome prediction. The models include logistic regression with all variables (LO), logistic regression with selected variables (LO), Support Vector Machine (SVM), Naïve Bayes (NB), Neural Network (NN), Decision Tree (DT), and Light GBM (LGBM). Each model's performance is quantified with corresponding 95% confidence intervals (CI) for the presented metrics.

relatively high SE and NPV, implying its competence in identifying true positive cases and true negative cases, respectively.

In conjunction with the provided metrics, we conducted a supplementary analysis of the models' efficacy utilizing ROC curves. Proximity of the curve to the upper-left corner of the graph corresponds to heightened aptitude of the model in delineating between affirmative and negative occurrences. Notably, the NB and LGBM models exhibited conspicuous distinction, showcasing the most elevated AUC values within the cohort (Figure 2).

Nomogram: Integrating Predictive Factors for Mortality Analysis in CRC Patients

In the concluding phase of our investigation, a significant outcome arose with the introduction of a nomogram as a pivotal tool for predictive analysis, as shown in Figure 3. This nomogram effectively integrates a constellation of essential factors, employing logistic regression based on GINI criteria to predict mortality status. The utility of the nomogram lies in its user-friendly nature, encompassing a straightforward 3-step process. Initially, the value associated with a specific patient is located on the scale for each variable, and the corresponding score is determined using the scoring scale. Subsequently, the cumulative score is computed by aggregating the obtained scores from the previous step. This cumulative score is then positioned on the total score scale. Finally, the probability of an event correlated with the subject's total score is ascertained from the probability scale.

Discussion

This study aimed to identify key variables influencing mortality prediction among CRC patients and establish an optimized prediction model. Analysis of mean decrease GINI values revealed critical variables in mortality prediction. Variables such as Time from diagnosis, Age at Diagnosis, Tumor size, Metastasis (Yes vs. No), Lymph node involvement (N2 vs. others), and Types

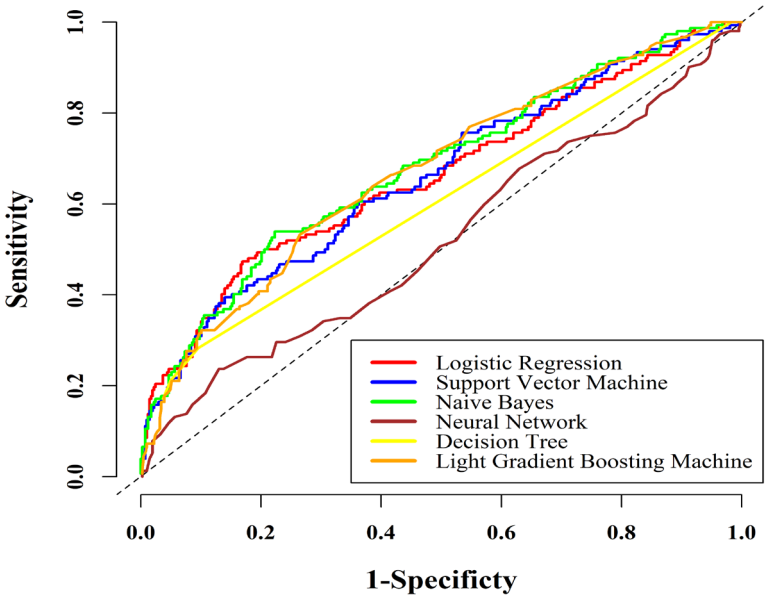

Figure 2. ROC Curves for Mortality Prediction Using Different Machine Learning Models

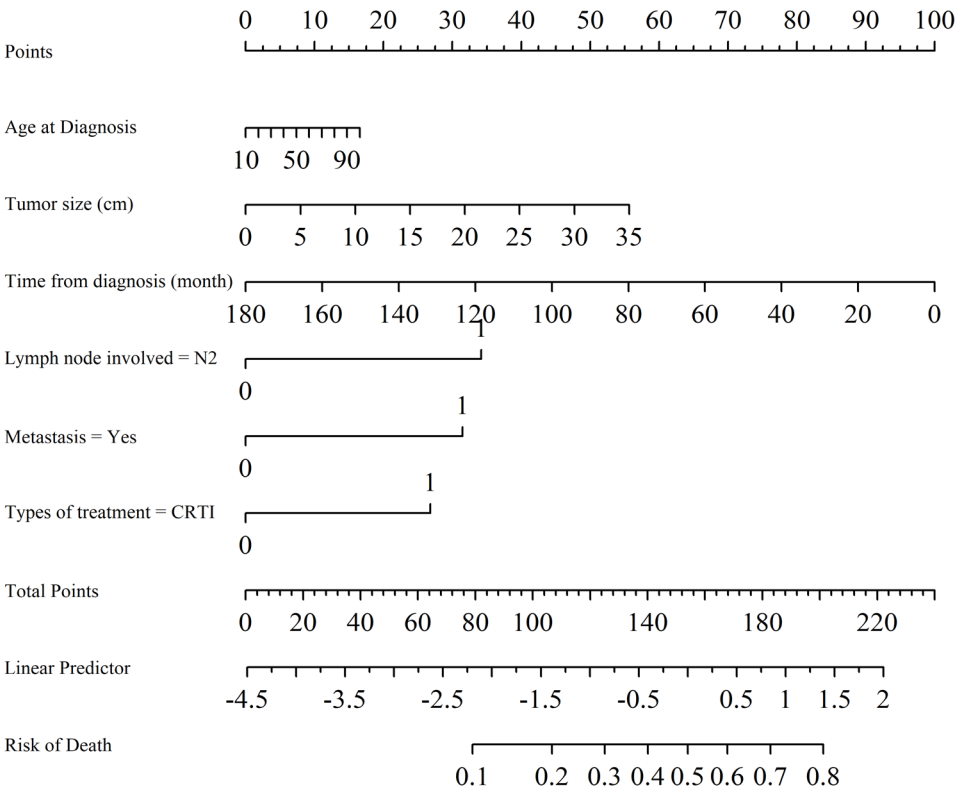

Figure 3. Nomogram for Predicting Mortality Using Logistic Regression

of treatment (Chemotherapy & Radiation Therapy & Immunotherapy vs. surgery) emerged as vital components of the prediction model, underscoring their significance in mortality outcomes. Among machine learning models, the NB model demonstrated the highest efficacy based on AUC, maintaining a balance between sensitivity and specificity. Logistic Regression, SVM, and LGBM also performed competitively. However, the NN model showed relatively lower AUC and sensitivity, suggesting the need for architectural refinement or feature engineering to enhance its predictive capacity.

In this study, we utilized machine learning models to predict survival outcomes in a CRC patient cohort, employing a comprehensive and rigorous methodology. However, it's essential to acknowledge and address inherent limitations. The retrospective nature of the study, relying on historical medical records, may introduce information bias and incomplete data capture. Despite our efforts to collect comprehensive patient information, missing or incomplete data necessitated the exclusion of some cases, impacting result robustness. Exclusions based on inaccessible records, non-Iranian residency, and overseas residence may limit result generalizability. Unmeasured or residual confounding variables, like socioeconomic status and treatment adherence, should be considered. The study's confinement to three Iranian tertiary hospitals may restrict broader applicability due to geographical and healthcare variations. The hybrid variable selection approach introduces subjectivity and potential bias, impacting variable importance determination. The selective omission of specific clinical variables aligned with clinical perspectives may affect model predictability.

This study's strength lies in its comprehensive and meticulous methodology. It spanned nearly 13 years across three prominent tertiary hospitals in Iran, encompassing a diverse patient population. Strict inclusion and exclusion criteria ensured participant relevance. Accurate data collection from medical records and telephone interviews was conducted, and standardized medical protocols reduced treatment variability. A wide array of covariates, including demographics, clinical factors, and treatments, were analyzed to explore potential confounding variables. The study's multicenter nature and comparative analysis of machine learning models helped mitigate potential biases. Transparency in data analysis methodologies, external review, and validation further bolstered credibility. Despite potential bias sources, the meticulous approach, large sample size, statistical rigor, and consideration of clinical perspectives contribute to robust and meaningful insights.

The present study has demonstrated a commendable performance in terms of the AUC predicting the survival status of CRC patients. Nonetheless, the literature exhibits heterogeneity in reported AUC values across studies predicting survival outcomes among CRC patients. For instance, Zhao et al. employed Bayesian additive regression trees (BART), a statistical learning method, to scrutinize patient-specific tumor attributes for enhanced prognostic prediction in CRC. Their BART model, utilizing seven robust and relevant variables, yielded AUCs ranging

from 0.67 to 0.83 (median: 0.74) through five-fold cross-validation. The relatively elevated AUC could potentially be attributed to a richer dataset, as their investigation analyzed 75 clinicopathologic, immune, microbial, and genomic factors within 815 stage II-III patients across two comprehensive U.S.-wide prospective cohorts, resulting in the identification of seven consistent survival predictors [21]. Additionally, Gupta et al. pursued a study employing data mining techniques and extant CRC risk prediction models, wherein deep autoencoders achieved exceptional performance outcomes, attaining 95% AUC. This heightened AUC value, in contrast to our study, could be attributed to the substantial sample sizes within their investigation, utilizing the Surveillance, Epidemiology, and End Results (SEER) dataset [22]. However, a study by Achilonu et al., characterized by a sample size smaller than ours, exhibited a higher AUC of 0.82 (95% CI 0.776–0.856). Notably, their research established a correlation between recurrence and diminished survival, a phenomenon absent in our study [23]. The observed augmentation in AUC values within their study may be attributed to superior data quality. In contrast, our study involved the integration of three discrete historical datasets to enable a multicenter framework. Nevertheless, while a single-center design, analogous to that of Achilonu et al., possesses the capability to mitigate specific biases and enhance predictive precision, it concurrently engenders the potential repercussion of curtailing the generalizability of findings.

Several studies have consistently identified the prominent machine learning algorithms employed for predicting CRC survival, including NN, LR, RF, SVM, DT, and LGBM. Additionally, a prevalent practice across these investigations involved the application of feature selection techniques to identify the most relevant subset of variables contributing to survival outcome prediction [23-26, 5]. In congruence with these established approaches, our study similarly incorporated these data mining methods and feature selection strategies. The significance of feature selection is paramount in addressing the intricacies of high-dimensional data analysis. Particularly pertinent to high-dimensional datasets, the strategic elimination of irrelevant and redundant features through feature selection confers advantages in terms of predictive performance enhancement, computational efficiency, and interpretability of outcomes. In the context of method comparison, benchmark studies have garnered notable attention, primarily focusing on classification datasets. These studies meld feature selection methods with classification techniques to gauge the predictive efficacy of the chosen features. Undeniably, feature selection constitutes a pivotal facet in the construction of machine learning models, offering potential for performance optimization, computational resource conservation, and refinement of result interpretation [27].

Our investigation revealed that several key variables significantly influenced the prediction of survival outcomes in CRC patients. Notably, the variables of time elapsed from diagnosis, age at diagnosis, tumor size, metastatic status, lymph node involvement, and type of treatment emerged as pivotal determinants. This finding aligns with previous research conducted by Lee et al., who demonstrated that an extended interval between confirmed CRC diagnosis and the initiation of treatment correlated with a markedly elevated risk of mortality, a trend observed consistently across all cancer stages. This may be attributed to the shorter timeframe from diagnosis to outcome events associated with prolonged diagnostic-to-treatment intervals [28]. The influence of age, a well-established prognostic factor, on CRC patient survival was substantiated in our study, as evident in studies by Achilonu et al. and Gao et al. [29, 23]. Additionally, tumor size exhibited considerable clinical significance, displaying prognostic and predictive value for colon cancer, warranting its selective incorporation into staging systems to enhance risk prediction for mortality [30, 31]. Moreover, the intricate role of metastasis in prognosticating CRC survival has been underscored in various studies [29, 30]. The presence of lymph node metastasis (LNM) holds sway over prognosis and clinical decision-making in colorectal cancer cases. Krogue et al. contributed insights through their work, wherein machine-derived features, generated via their method, exhibited significant associations with LNM, even after accounting for established clinicopathologic variables, as confirmed by external validation [32].

In conclusion, this study aimed to identify key predictors and establish an optimized model for predicting mortality in CRC patients. Variables such as time from diagnosis, age at diagnosis, tumor size, metastatic status, lymph node involvement, and type of treatment were identified as crucial for mortality prediction, with the NB model exhibiting the most balanced performance. The LR, SVM, and LGBM models also showed competitive predictive capacity, while the NN model had relatively lower performance. Despite inherent limitations, including potential data bias and generalizability constraints, the multicenter design, extensive variable analysis, and rigorous methodologies contribute to the study's robustness and meaningful insights.

While our study demonstrated commendable performance in predicting CRC survival, variation in reported AUC values across different studies highlights the complexity of predicting survival outcomes in CRC patients. The incorporation of richer datasets and different ML techniques has led to differing AUC values in the literature. Acknowledging these discrepancies is crucial for understanding the context of our study's findings. Furthermore, the integration of feature selection methods into ML models, such as the one used in our study, is vital for optimizing predictive performance, computational efficiency, and result interpretation. This approach aligns with established practices in the field and contributes to the methodological rigor of our study.

Based on the findings of this study, it is recommended that early diagnosis and timely treatment initiation for CRC patients be promoted by policy makers in the healthcare sector, with the aim of reducing the interval between diagnosis and treatment to mitigate mortality risks associated with delayed care. The allocation of resources for comprehensive data capture through digital health

records and standardized protocols should be considered to address potential incomplete data and information bias. Multicenter collaborations should be encouraged to capture diverse patient demographics and treatment strategies, thereby enhancing the generalizability of results. The integration of predictive analytics, particularly the data mining models, into clinical decision-making processes can be supported to aid personalized patient care. Research collaborations and data sharing initiatives should be strengthened to address the heterogeneity in reported AUC values across studies and promote standardized methodologies for predictive model comparison and validation. Furthermore, the incorporation of identified prognostic variables into treatment guidelines should be considered by policy makers to optimize patient management strategies and enhance outcomes for CRC patients.

Author Contribution Statement

Methodology and formal analysis, M.A.L., M.A.P., A.Z.S.K., S.J., M.H. and S.K. ; software, M.A.L., G.M., S.S. and Z.M.H.C.; validation, M.A.L., S.Z.S., M.Z.A. and M.E.M; investigation, M.R.T. and H.P.; resources, S.K., J.C.Z. and M.A.P.; responsible for data collection M.A.P.; data curation, M.A.L., R.T., R.M.R., Z.R. and N.P.; writing-original draft preparation, M.A.L.; writing-review and editing, I.V., M.Z.A. and R.T.; visualization, M.A.L.; supervision, G.M. and M.A.P; All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We express our gratitude to Imam Khomeini Hospital of Mazandaran, Taleghani Hospital in Tehran, and Shahid Faghihi Hospital in Shiraz for their collaboration in facilitating the collection of the dataset.

Scientific Writing Guideline

This research adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline.

Ethical Approval and Consent to Participate

The study was conducted according to the guidelines of the Declaration of Helsinki, and was approved by Research Ethics Committee of SBMU, Tehran, Iran (IR.SBMU.RETECH.REC.1402.329).

Conflict of Interest

The authors declare no conflict of interest.

References

1. Alboaneen D, Alqarni R, Alqahtani S, Alrashidi M, Alhuda R, Alyahyan E, et al. Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. Big Data Cogn Compu. 2023;7(2):74.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA cancer J Clin. 2021;71(3):209-49.
3. The International Association of Cancer Registries (IACR). Cancer today. 2023.
4. Meera C, Nalini D. Breast cancer prediction system using data mining methods. Int J Pure Appl Math. 2018;119(12):10901-11.
5. Shanbehzadeh M, Nopour R, Kazemi-Arpanahi H. Comparison of four data mining algorithms for predicting colorectal cancer risk. J Adv Med Biomed Res. 2021;29(133):100-8.
6. Patil S, Moafa IH, Alfaifi MM, Abdu AM, Jafer MA, Raju L, et al. Reviewing the role of artificial intelligence in cancer. Asian Pac J Cancer Biol. 2020;5(4):189-99.
7. Maher RS, Bhawiskar SK, editors. Review on automated skin cancer detection using image processing methods. International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022); 2023: Atlantis Press.
8. Gangopadhyay A. Artificial intelligence and cancer control in low-middle income countries-relevance in the covid-19 era. Asian Pac J Cancer Care. 2023;8(3):663-5.
9. Ebrahim M, Sedky AAH, Mesbah S. Accuracy assessment of machine learning algorithms used to predict breast cancer. Data. 2023;8(2):35.
10. Wu R, Luo J, Wan H, Zhang H, Yuan Y, Hu H, et al. Evaluation of machine learning algorithms for the prognosis of breast cancer from the surveillance, epidemiology, and end results database. Plos One. 2023;18(1):e0280340.
11. Azari H, Nazari E, Mohit R, Asadnia A, Maftooh M, Nassiri M, et al. Machine learning algorithms reveal potential mirnas biomarkers in gastric cancer. Sci Rep. 2023;13(1):6147.
12. Kamal VK, Kumari D. Use of artificial intelligence/machine learning in cancer research during the covid-19 pandemic. Asian Pac J Cancer Care. 2020;5(S1):251-3.
13. Skrede O-J, 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(10221):350-60.
14. Osman MH, Mohamed RH, Sarhan HM, Park EJ, Baik SH, Lee KY, et al. Machine learning model for predicting postoperative survival of patients with colorectal cancer. Cancer Res Treat. 2022;54(2):517-24.
15. Nartowt BJ, Hart GR, Muhammad W, Liang Y, Stark GF, Deng J. Robust machine learning for colorectal cancer risk prediction and stratification. Front Big Data. 2020;3:6.
16. Zhang J. Estimating confidence intervals on accuracy in classification in machine learning. 2019.
17. Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology. 2007;88(11):2783-92.
18. Strobl C. Statistical issues in machine learning towards reliable split selection and variable importance measures. Cuvillier Verlag; 2008.
19. Calle ML, Urrea V. Stability of random forest importance measures. Brief bioinformatics. 2011;12(1):86-9.
20. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008;26(8):1364-70.
21. Zhao M, Lau MC, Haruki K, Väyrynen JP, Gurjao C, Väyrynen SA, et al. Bayesian risk prediction model for colorectal cancer mortality through integration of clinicopathologic and genomic data. NPJ Precis Oncol. 2023;7(1):57. https://doi.org/10.1038/s41698-023-00406-8.
22. Gupta S, Kalaivani S, Rajasundaram A, Ameta GK, Oleiwi AK, Dugbakie BN. Prediction performance of deep learning for colon cancer survival prediction on seer data. Biomed Res Int. 2022;2022:1467070. https://doi.

org/10.1155/2022/1467070.
23. Achilonu OJ, Fabian J, Bebington B, Singh E, Nimako G, Eijkemans MJC, et al. Predicting colorectal cancer recurrence and patient survival using supervised machine learning approach: A south african population-based study. Front Public Health. 2021;9. https://doi.org/10.3389/fpubh.2021.694306.
24. Bai L, Bu F, Li X, Yang X, Guo S, Min L, et al. Mc-kv: A prognosis-oriented classifier based on semi-supervised learning for molecular subtyping of colorectal cancer. Adv Theory Simul. 2023;6(6):2300156. https://doi.org/https://doi.org/10.1002/adts.202300156.
25. Chen PC, Yeh YM, Lin BW, Chan RH, Su PF, Liu YC, et al. A prediction model for tumor recurrence in stage ii-iii colorectal cancer patients: From a machine learning model to genomic profiling. Biomedicines. 2022;10(2):340. https://doi.org/10.3390/biomedicines10020340.
26. Pourhoseingholi MA, Kheirian S, Zali MR. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. Acta Inform Med. 2017;25(4):254-8. https://doi.org/10.5455/aim.2017.25.254-258.
27. Bommert A, Welchowski T, Schmid M, Rahnenführer J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. Briefings in Bioinformatics. 2021;23(1): bbab354. https://doi.org/10.1093/bib/bbab354.
28. Lee YH, Kung PT, Wang YH, Kuo WY, Kao SL, Tsai WC. Effect of length of time from diagnosis to treatment on colorectal cancer survival: A population-based study. PLoS One. 2019;14(1):e0210465. https://doi.org/10.1371/journal.pone.0210465.
29. Gao P, Zhou X, Wang Zn, Song Yx, Tong Ll, Xu Yy, et al. Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. The tnm staging system. PLOS One. 2012;7(7):e42015. https://doi.org/10.1371/journal.pone.0042015.
30. Liu Z, Xu Y, Xu G, Baklaushev VP, Chekhonin VP, Peltzer K, et al. Nomogram for predicting overall survival in colorectal cancer with distant metastasis. BMC Gastroenterol. 2021;21(1):103. https://doi.org/10.1186/s12876-021-01692-x.
31. Xu W, He Y, Wang Y, Li X, Young J, Ioannidis JPA, et al. Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: An umbrella review of systematic reviews and meta-analyses of observational studies. BMC Med. 2020;18(1):172. https://doi.org/10.1186/s12916-020-01618-6.
32. Krogue JD, Azizi S, Tan F, Flament-Auvigne I, Brown T, Plass M, et al. Predicting lymph node metastasis from primary tumor histology and clinicopathologic factors in colorectal cancer using deep learning. Commun Med. 2023;3(1):59. https://doi.org/10.1038/s43856-023-00282-0.