COMPARATIVE ANALYSIS OF COLORECTAL CANCER BASED ON MACHINE LEARNING ALGORITHMS

K.Muthuchamy, *Research Scholar, Assistant Professor, Department of Computer Science and Information Technology, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai-117.* muthuchamyka@gmail.com

Dr. SK. Piramu Preethika, *Assistant Professor, Department of Computer Science and Information Technology, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai-117.* Piramu.scs@vistas.ac.in

Abstract

Colorectal cancer (CRC) is the second leading cause of cancer-related deaths. Computational intelligence (CI) has emerged as a promising tool to improve diagnosis, staging, and treatment, but evidence remains scattered across the literature. This work aimed to predict Colorectal cancer patients using machine learning (ML) methods. A retrospective analysis included in PubMed and EMBASE identified systematic reviews, following PRISMA guidelines. Extracted data covered CI techniques, evaluation methods, target outcomes, and dataset characteristics. 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 GINI index criteria. which yields the highest accuracy (~96.2%) with better precision, recall, and F1-scores. Time from diagnosis, age, tumor size, metastatic status, lymph node involvement, and treatment type emerged as crucial predictors of Colorectal cancer based on mean GINI index. The NB models achieved the highest predictive values for CRC patient. 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 patient. The NB model exhibited optimal efficacy in 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- GINI Index- machine learning algorithms*

Introduction

Colorectal colon, with the third highest diagnosis rate, is the second dangerous cancer in the world. Colorectal cancer constitutes a substantial public health challenge, particularly among men. Originating from malignant cells within the rectum, a segment of the large intestine, colorectal cancer progresses through distinct stages, often asymptomatically during its early phases. Early detection via systematic screening is pivotal for improving clinical outcomes and reducing mortality. However, traditional diagnostic modalities, including endoscopy and pathological biopsy, are hindered by their significant time, cost, and accessibility constraints[1,2]. Diagnosing is one of the core principles of medicine based on the integration of multi-source data analysis and clinician experience. Because of the variety of tumor symptoms, rapid tumor growth, individual differences, and drug sensitivity, it is difficult for doctors to diagnose tumors accurately. Artificial Intelligence (AI) supports clinicians in qualitative diagnosis and detecting the stage of colon cancer, which currently relies on endoscopy and pathological biopsy [3].

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 [4]. 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 [5].

To address this critical need, establishing a CRC monitoring system that regularly screens individuals based on their risk factors 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 [6-10]. Despite the considerable scientific attention toward predicting CRC patient using ML approaches [11-13], none of these studies have reported the 96% 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 [14]. Consequently, the aim of this study 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 96% 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

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

The investigation was carried out across prominent tertiary databases. A span of nearly 1 year 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.

The primary explanatory variables encompassed age at diagnosis, tumor size (centimeters), 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.

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 datasets.

In this retrospective study for CRC patients, a comprehensive strategy addressed potential biases. Inclusion and exclusion criteria were meticulously defined, encompassing diverse CRC patients. Data collection involved robust examination of medical records to enhance data accuracy. A wide range of covariates allowed exploration of confounding variables. Adoption of different ML models facilitated comparative analysis. Transparent data analysis and external review further enhanced credibility.

This study consisted of 785 individuals diagnosed with CRC. This investigation embraced a comprehensive approach, encompassing the entire eligible patient population throughout the study duration, thus precluding the need for sampling.

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 Python software, wherein statistical significance was indicated by P-values of 0.05, coupled with a 96% confidence interval (CI).

The study population was divided randomly into two distinct samples: 80% were employed as training data for outcome prediction, and the remaining 20% 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 [4-5]. Tuning of each algorithm’s parameters was executed to optimize outcome risk prediction accuracy.

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) [6]. Model performance was measured based on mean GINI index criteria. which yields the highest accuracy (~96.2%) with better precision, recall, and F1-scores.

A hybrid approach encompassing both statistical techniques and clinical considerations was undertaken for variable selection. The mean Gini index 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 [15-17]. Concurrently, clinical variables deemed unrelated were omitted from initial random forest analysis, aligning with clinical perspectives.

To facilitate the creation of a user-friendly predictive model yielding numerical probabilities of fibrosis incidence [18], a nomogram was employed, serving as a graphical representation of the statistical predictive framework.

Results

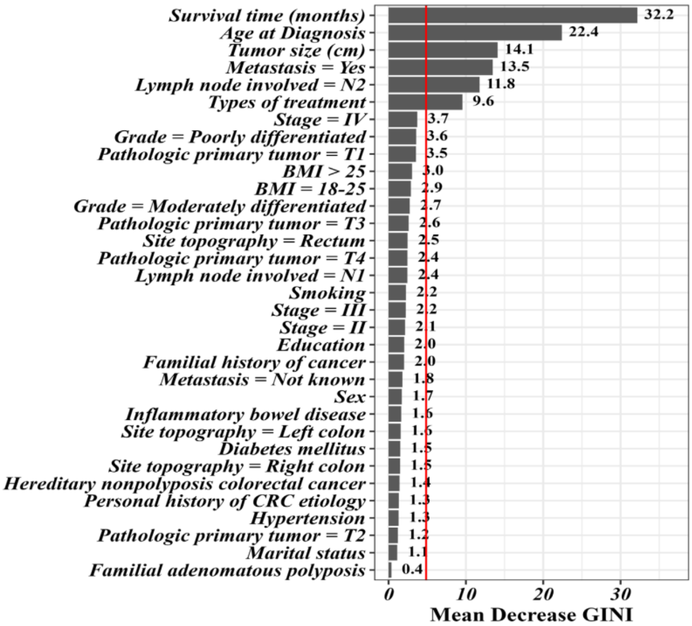
The descriptive statistics were reported in Table 1. A total of 785 patients were included in the study with mean age of 54.20 ± 14.52 years. Tumor size, survival time, BMI categories, marital status, familial history of cancer, site topography, grade, pathologic primary tumor, lymph node involvement, metastasis, and stage all exhibited statistically significant associations with vital status.

Table 1. Descriptive Statistics and Comparative Analysis of Patient Variables

Variable	Levels	Total	Vital status	P-value	Variable
Age at diagnosis	-----	54.20 ± 14.52	54.82 ± 13.22	55.24 ± 14.43	0.062
Sex	Male	562 (59.37)	452 (59.01)	297 (60.37)	0.63
	Female	223 (40.63)	134 (40.99)	185 (39.63)	
Marital status	Married	652 (94.87)	315 (95.57)	357 (92.89)	0.023
	Other	233 (5.13)	51 (4.43)	55 (7.11)	
BMI – four categories	<18.5	240 (7.47)	90 (5.81)	68 (11.79)	<0.001
	18.6-24.9	220 (54.67)	148 (54.36)	174 (55.69)	
	25-29.9	165 (29.63)	321 (30.60)	67 (27.03)	
	>30	400 (8.22)	227 (9.23)	47 (5.49)	
BMI – three categories	<18	245 (7.74)	83 (6.03)	55 (12.20)	<0.001
	18-25	260 (54.40)	745 (54.14)	121 (55.28)	
	>25	280 (37.85)	548 (39.83)	76 (32.52)	
Smoking	No	440 (71.54)	981 (71.29)	354 (71.95)	0.816
	Yes	345 (28.46)	395 (28.71)	138 (28.05)	
Education	Illiterate	121 (28.24)	380 (27.62)	148 (30.08)	0.286
	Primary school	234 (32.78)	459 (33.36)	153 (31.10)	
	High school	201 (21.41)	305 (22.17)	96 (19.51)	
	University	229 (17.57)	232 (16.86)	95 (19.31)	
Hypertension	No	666 (88.95)	216 (88.37)	447 (90.85)	0.153
	Yes	229 (11.05)	60 (11.63)	45 (9.15)	
Diabetes	No	105 (89.91)	122 (89.53)	47 (90.85)	0.434
	Yes	680 (10.09)	44 (10.47)	45 (9.15)	
Familial history of cancer	No	220 (64.50)	85 (62.86)	38 (68.70)	0.021
	Yes	665 (35.50)	51 (37.14)	154 (31.30)	
Site topography	Right colon	200 (30.75)	42 (30.60)	154 (31.30)	0.02
	Left colon	101 (58.41)	82 (59.74)	68 (54.47)	
	Rectum	283 (9.77)	22 (8.87)	61 (12.40)	
	Transverse	101 (1.07)	11 (0.80)	9 (1.83)	
Grade	Well differentiated	142 (55.63)	19 (59.52)	221 (44.92)	<0.001
	Moderately differentiated	67 (36.15)	47 (34.16)	25 (41.67)	
	Poorly differentiated	124 (8.22)	87 (6.32)	66 (13.41)	
Pathologic primary tumor	T0	174 (57.34)	37 (60.83)	32 (47.15)	<0.001
	T1	243 (18.31)	24 (15.55)	29 (26.22)	
	T2	243 (7.63)	13 (8.21)	30 (6.10)	
	T3	92 (15.59)	22 (14.68)	90 (18.29)	
	T4	21 (1.12)	10 (0.73)	11 (2.24)	
Stage	I	281 (15.00)	27 (16.50)	52 (10.57)	<0.001
	II	181 (36.36)	56 (38.23)	54 (31.30)	

III	163 (33.69)	40 (33.43)	19 (34.35)
IV	228 (14.95)	13 (11.85)	17 (23.78)

Figure 1 displays mean GINI index 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:

As the next step, we aimed to predict the patient 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 0.86, indicating good discriminatory ability. This model also shows balanced sensitivity and specificity values, with a SE of 0.62 and SP of 0.75. The NB model demonstrates a PPV of 0.47 and a NPV of 0.85, suggesting its effectiveness in correctly classifying both positive and negative outcomes.

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

Variables	AUC (96% CI)	SE (96% CI)	SP (96% CI)	PPV (96% CI)	NPV (96% CI)	ACC (96% CI)
Logistic regression - all variables (LO-A)	0.47 (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)	0.67 (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	0.69 (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)

Machine (SVM)						
Naïve Bayes (NB)	0.86 (0.65, 0.75)	0.62 (0.51, 0.68)	0.75 (0.69, 0.78)	0.47 (0.40, 0.54)	0.85 (0.78, 0.86)	0.70 (0.66, 0.73)
Neural Network (NN)	0.58 (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)	0.62 (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)

Discussion

This work aimed to identify key variables influencing CRC patients and establish an optimized prediction model. Analysis of mean GINI index values revealed critical variables in prediction. Variables such as Age at Diagnosis, Tumor size, Metastasis (Yes vs. No), Lymph node involvement (N2 vs. others), and Types of treatment (Chemotherapy & Radiation Therapy & Immunotherapy vs. surgery) emerged as vital components of the prediction model, underscoring their significance in prediction outcomes. Among machine learning models, the NB model demonstrated the highest efficacy based on measures, 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.

References

1. *Zhao T, Zeng Z, Li T, Tao W, Yu X, Feng T, et al. USC-ENet: a high-efficiency model for the diagnosis of liver tumors combining B-modes ultrasound andclinical data. Health Inf Sci Syst. 2023;11(1):15. pmid:36950106.*
2. *Thakur T, Batra I, Luthra M, Vimal S, Dhiman G, Malik A, et al. Gene Expression-Assisted Cancer Prediction Techniques. J Healthc Eng.2021;2021:4242646. pmid:34545300*
3. *Gupta N, Kupfer SS, Davis AM. Colorectal Cancer Screening. JAMA. 2019;321(20):2022–3. pmid:31021387*
4. *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.*
5. *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.*
6. *Meera C, Nalini D. Breast cancer prediction system using data mining methods. Int J Pure Appl Math. 2018;119(12):10901- 11.*
7. *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.*
8. *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.*
9. *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.*
10. *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.*
11. *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.*

12. *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.*
13. *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*
14. *Zhang J. Estimating confidence intervals on accuracy in classification in machine learning. 2019*
15. *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.*
16. *18. Strobl C. Statistical issues in machine learning towards reliable split selection and variable importance measures. Cuvillier Verlag; 2008.*
17. *19. Calle ML, Urrea V. Stability of random forest importance measures. Brief bioinformatics. 2011;12(1):86-9.*
18. *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.*