DOI: 10.58490/ctump.2025i10TA.4407

PROGNOSTIC VALUE OF DEMOGRAPHICS, HISTOPATHOLOGY, AND MISMATCH REPAIR STATUS IN COLORECTAL CANCER: A TIME-TO-EVENT MACHINE LEARNING APPROACH

Hoang Duc Trinh*[1,2*], Huynh Minh Thien*[3]***, Tran-Thi Huong Ly*[3]***,
Quoc Thang Pham*[1]***, Vo Van Kha*[3]

1. University of Medicine and Pharmacy at Ho Chi Minh City
2. Can Tho University of Medicine and Pharmacy
3. Can Tho Oncology Hospital
**Corresponding author: hdtrinh@ctump.edu.vn*

Received: 14/11/2025
Reviewed: 20/12/2025
Accepted: 25/12/2025

ABSTRACT

Background: *Colorectal cancer prognosis depends on multiple factors including demographics, histopathology, and mismatch repair (MMR) status. Traditional Cox regression models have limitations in handling high-dimensional clinical data.* ***Objectives:*** *This study applied time-to-event machine learning algorithms to investigate the prognostic values of demographics, histopathology, and MMR status in predicting overall survival of colorectal cancer patients.* ***Materials and Methods:*** *A total of 165 colorectal cancer patients from Can Tho Oncology Hospital (2019-2021) were recruited. Input features included age, sex, pTNM stage, histological type, tumor grade, lymphovascular invasion (LVI), tumor-infiltrating lymphocytes (TILs), perineural invasion (PNI), and MMR protein status. The dataset was divided into training (70%) and testing sets. Five time-to-event algorithms were trained with 1000 bootstraps and hyperparameter tuning, then validated on the testing set. SurvSHAP package was used for feature importance ranking.* ***Results:*** *Gradient Boost Survival outperformed other models with acceptable discrimination (C-index: 0.812, 95% CI: 0.784-0.840) and calibration (integrated Brier score: 0.057, 95% CI: 0.056-0.058). Age, lymphovascular invasion, and MMR status were identified as the three most important predictive features. Feature importance increased during the first 24 months and then stabilized.* ***Conclusion:*** *Time-to-event ensemble machine learning models effectively predict survival prognosis in colorectal cancer. MMR status, combined with demographic and histopathological features, represents an important predictor of overall survival.*

Keywords: *Colorectal cancer, overall survival, mismatch repair, time-to-event analysis, machine learning*

I. INTRODUCTION

Colorectal cancer ranks among the most common malignancies worldwide, contributing significantly to cancer-related mortality [1], [2], [3]. Treatment response and prognosis depend on multiple factors, including histopathological characteristics and mismatch repair (MMR) status [4], [5]. Deficient MMR (d-MMR), characterized by mutational inactivation of DNA repair proteins, results in microsatellite instability (MSI) and exhibits distinct features such as poor differentiation, dense tumor-infiltrating lymphocytes, and variable chemotherapy responses compared to proficient MMR (p-MMR) cancers [6], [7], [8].

While early evidence suggested d-MMR tumors have better prognosis than p-MMR [9], recent studies show inconsistent results [7], [10]. This controversy suggested that combined features should be utilized for colorectal cancer prognosis rather than relying solely on MMR status. Several predictive models using Cox proportional hazards regression have been constructed [11], but this approach has limitations in handling high-dimensional clinical data due to statistical assumptions and feature interactions [12]. Time-to-event machine learning models have demonstrated superior performance in predicting cancer prognosis compared with traditional Cox models [12], [13].

Studies applying time-to-event machine learning to colorectal cancer survival prognosis are rare, and none have investigated the predictive ability of MMR-related features using these advanced methods. This study aimed to evaluate the prognostic value of demographic characteristics, histopathological features, and MMR status for predicting overall survival in patients with colorectal cancer using time-to-event machine learning algorithms.

II. MATERIALS AND METHODS

2.1. Participants

The study included 165 patients with confirmed colorectal cancer diagnosis who underwent treatment at Can Tho Oncology Hospital between January 1, 2019 and December 31, 2021, with follow-up ending December 31, 2023.

- **Inclusion criteria:** (1) follow-up time ≥1 month; (2) histopathologically confirmed colorectal diagnosis; (3) available MMR evaluations.

- **Exclusion criteria:** (1) coexisting primary tumors; (2) missing essential demographic or cancer stage data. MMR evaluation was conducted through immunohistochemistry.

2.2. Methods

- **Study-design:** A retrospective cohort study.

- **Sample size:** 165 patients.

- **Sampling method:** Convenience random sampling.

- **Study contents:**

- **Inputs and Outcomes:** Input demographic characteristics included age and sex. Histopathological features from H&E staining and immunohistochemistry included pTNM stage, histological types, tumor grade, lymphovascular invasion (LVI), tumor-infiltrating lymphocytes (TILs), and perineural invasion (PNI). Of note, to determine TILs, we counted lymphocytes in representative tumor regions in HE slides using an optical microscope at a magnification of x400. We calculated the average cell number from five consecutive microscopic fields to retrieve the final TILs values. High TILs and low TILs were considered if TILs values were ≥ 3 and < 3, respectively. MMR status data included presence of d-MMR and absence of specific proteins (MLH1, PMS2, MSH2, MSH6). Chemotherapy received during follow-up was also included. The outcome was overall survival, defined as duration from diagnosis commencement to death in months.

- **Data Preprocessing:** Data were checked for missing values, and categorical variables were one-hot encoded. The dataset was divided into training (70%) and testing (30%) sets. The training set was used for model development, with performance evaluated on the test set.

- **Model Development:** Five time-to-event machine learning algorithms were employed: Cox proportional hazard model with Breslow's method, Cox elastic-net regression, Survival Tree, Random Survival Forest, and Gradient Boost Survival.

Hyperparameter tuning was performed using GridSearchCV with 1000 bootstraps on the training set. Models with optimized parameters were validated on the testing set with 1000 bootstraps. The best-performing model underwent feature importance ranking using SurvSHAP function, which evaluates both overall feature importance and temporal changes in feature importance.

- Performance metrics: Time-dependent concordance index (C-index) evaluated discriminative ability, with higher values indicating better discrimination and 0.5 representing no-skill prediction. Integrated Brier score assessed calibration, with lower scores indicating better models. Time-dependent area under the ROC curve (time-dependent ROC-AUC) was also evaluated. Performances were compared based on mean and 95% confidence intervals from 1000 bootstraps on the testing dataset.

- Statistical Analysis: Complete-case approach was used to handle missing values. Continuous variables were expressed as mean ± standard deviation, categorical variables as percentages. Training and testing sets were compared using Mann-Whitney U test or Fisher's exact test. Survival rates were compared by log-rank test. P-value < 0.05 was considered significant. Statistical analysis and machine learning were performed using R (version 4.2.3) and Python (version 3.10.2).

- Ethic approval: This study was approved by the Institute Research Board of the University of Medicine and Pharmacy in Ho Chi Minh City (approval number 20/HDDD-DHYD, January 10, 2022). Informed consent was obtained from all participants.

III. RESULTS

3.1. Participant Characteristics

Age and sex distributions were similar between training and testing sets, predominantly middle-aged and elderly patients. Histopathological characteristics and MMR status showed no significant differences between datasets. The cohort mainly consisted of colorectal adenocarcinoma with tumor grade II. Disease staging included 88 patients at stages I-II and 77 at stages III-IV. Large proportions exhibited LVI, PNI, and TILs. Follow-up time and death events were comparable between datasets. Survival rate differences were not significant between datasets (log-rank test p=0.24), indicating suitability for machine learning training-testing procedures (Table 1).

Table 1. Participants' demographic and histopathological characteristics

Characteristic	Testing set N = 50	Training set N = 115	p-value
Age (mean, SD)	59 (11)	61 (14)	0.316
Sex (number of males, %)	30 (60%)	63 (55%)	0.610
Histological types (n, %)			1.000
Non-mucinous adenocarcinoma	48 (96%)	110 (95.7%)	
Mucinous adenocarcinoma	2 (4.0%)	5 (4.3%)	
Tumor grade (n, %)			0.554
I	50 (100%)	112 (97.4%)	
III	0 (0%)	3 (2.6%)	
pTNM stage (n, %)			0.498
I-II	29 (58%)	59 (51%)	
III-IV	21 (42%)	56 (49%)	
MMR (number of d-MMR, %)	13 (26%)	31 (27%)	1.000

Characteristic	Testing set N = 50	Training set N = 115	p-value
Lymphovascular Invasion (n, %)	38 (76%)	89 (77%)	0.843
Perineural Invasion (n, %)	30 (60%)	74 (64%)	0.603
Tumor-infiltrating lymphocytes (n, %)	25 (50%)	58 (50%)	1.000
Absence of MLH1 (n, %)	4 (8.0%)	13 (11%)	0.591
Absence of PMS2 (n, %)	11 (22%)	25 (22%)	1.000
Absence of MSH2 (n, %)	2 (4.0%)	6 (5.2%)	1.000
Absence of MSH6 (n, %)	3 (6.0%)	11 (9.6%)	0.555
Chemotherapy (n, %)	28 (56%)	71 (62%)	0.495
Follow-up time (mean, SD)	37 (11)	36 (13)	0.657
Deaths (n, %)	3 (6.0%)	14 (12%)	0.278

Note: differences in continuous variables between the two datasets were investigated by Mann–Whitney U test, whereas differences in continuous variables between the two datasets were investigated by Fisher's exact test. Note: MMR = mismatch repair; d-MMR = deficiency mismatch repair; MLH1 = MutL homolog 1; PMS2 = Postmeiotic Segregation Increased 2; MSH2 = MutS homolog 2; MSH6 = MutS homolog 6; MSH2: MutS homolog 2

3.2. Model Performances

Five time-to-event algorithms were trained with hyperparameter tuning using bootstrap methods. The Gradient Boost Survival model demonstrated superior performance in both discriminative ability (C-index: 0.812, 95% CI: 0.784-0.840) and calibration ability (Integrated Brier score: 0.057, 95% CI: 0.056-0.058), significantly outperforming other models based on bootstrapped 95% confidence intervals (Table 2). Regarding time-dependent ROC-AUC, all models showed stable prediction performance from approximately the tenth month, with Gradient Boost Survival exhibiting the highest values. The Gradient Boost Survival and Survival Tree models could predict overall survival in the early follow-up period, unlike other models. The Cox Elastic Net model showed prediction ability not substantially higher than random prediction.

Table 2. Performances of five time-to-event survival models

Models	C-index	Integrated Brier score	Time-dependent AUC
Cox proportion hazard	0.637 (0.598 - 0.676)	0.102 (0.071 - 0.133)	0.640 (0.592 - 0.688)
Cox Elastic	0.589 (0.562 - 0.616)	0.057 (0.055 - 0.059)	0.596 (0.559 - 0.633)
Survival Tree	0.744 (0.709 - 0.779)	0.592 (0.573 - 0.611)	0.765 (0.723 - 0.807)
Random Survival Forest	0.721 (0.696 - 0.746)	0.580 (0.561 - 0.599)	0.730 (0.696 - 0.764)
Gradient Boost Survival	0.812 (0.784-0.840)	0.057 (0.056 - 0.058)	0.835 (0.801 - 0.869)

3.3. Feature Importance Ranking

Given the Gradient Boost Survival model's superior performance, feature importance ranking was conducted using this model. The analysis revealed that age, lymphovascular invasion (LVI), and MMR status were the three most important and significant features for predicting overall survival (Figure 1A). The impacts of these features gradually increased during the first 24 months of follow-up and then stabilized (Figure 1B). This temporal pattern suggests that the predictive value of these features becomes more pronounced over time, reaching maximum importance at approximately two years post-diagnosis.

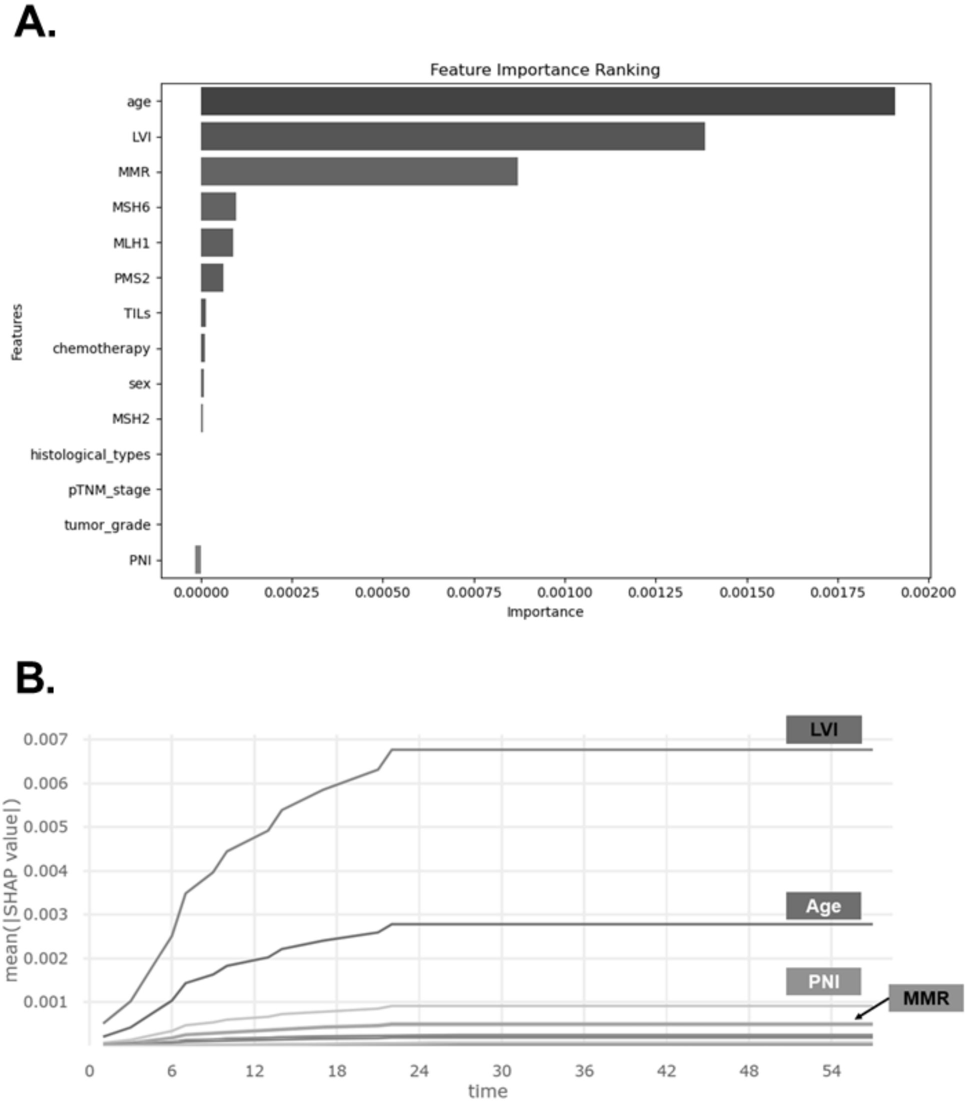

Figure 1. Feature importance ranking using the Gradient Boost Survival model based on aggregated SurvSHAP(t) values (A) and the changes of importance for each feature over time based on SurvSHAP(t) values (B). Due to the small SHAP values of most features, only 5 features that showed large changes in their importance over time were labeled.

IV. DISCUSSION

This is the first study using time-to-event machine learning algorithms to evaluate combined demographic, histopathological, and MMR-related features in colorectal cancer overall survival prediction. Using hospital-collected data, we found Gradient Boost Survival outperformed other models including Cox PH, Cox Elastic Net, Survival Tree, and Random Survival Forest. The model exhibited stable time-dependent ROC-AUC from the tenth month onwards. Critically, age, LVI, and MMR status emerged as the three most important predictive features, with their importance gradually increasing during the first 24 months before stabilizing.

Our findings align with previous studies showing machine learning advantages over classical approaches [12]. The Gradient Boost Survival algorithm optimizes loss functions through regression trees, effectively handling nonlinearities and interaction effects that challenge classical Cox PH models [12], [13]. This study supports the use of ensemble machine learning as an effective approach for cancer survival prediction.

The SurvSHAP analysis revealed age, LVI, and MMR status as primary predictors, consistent with existing literature. Age has been associated with overall and progression-free survival in colorectal cancer patients [14]. LVI predicts poor prognosis in stage III colorectal cancer [15]. Deficient MMR worsens survival in young metastatic colorectal cancer patients and significantly associates with overall survival [10]. Additionally, specific MMR gene dysfunctions (MLH1, PMS2, MSH2, MSH6) contributed differentially to our prediction model, suggesting each carrier mutation exhibits distinct cancer risk profiles requiring further investigation.

Study limitations include a relatively small sample size and especially a small number of events, although bootstrap validation demonstrated good discrimination and calibration. The model relied solely on available demographics and histopathological features, lacking data on tumor location, sidedness, and biomarkers (KRAS, BRAF) due to missing values. External validation was not performed. Future research should employ multi-center datasets with comprehensive clinical and preclinical data for external validation.

V. CONCLUSION

Time-to-event machine learning algorithms successfully predict overall survival in colorectal adenocarcinoma patients. Ensemble models outperform classical Cox models and survival tree models. Demographics, histopathology, and MMR status represent important predictors for patient overall survival. This approach demonstrates practical applicability using readily available clinical data and offers enhanced interpretability through time-dependent feature importance analysis, potentially improving therapeutic planning and patient management in colorectal cancer.

REFERENCES

1. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A, *et al.* Colorectal cancer statistics. 2017. *CA Cancer J Clin.* 2017. 67(3), 177-93. DOI: 10.3322/caac.21395.
2. Baidoun F, Elshiwy K, Elkeraie Y, Merjaneh Z, Khoudari G, Sarmini MT, *et al.* Colorectal Cancer Epidemiology: Recent Trends and Impact on Outcomes. *Curr Drug Targets.* 2021. 22(9), 998-1009. DOI: 10.2174/1389450121999201117115717.
3. Pardamean CI, Sudigyo D, Budiarto A, Mahesworo B, Hidayat AA, Baurley JW, *et al.* Changing Colorectal Cancer Trends in Asians: Epidemiology and Risk Factors. *Oncol Rev.* 2023. 1710576. DOI: 10.3389/or.2023.10576.
4. Hou JT, Zhao LN, Zhang DJ, Lv DY, He WL, Chen B, *et al.* Prognostic Value of Mismatch Repair Genes for Patients With Colorectal Cancer: Meta-Analysis. *Technol Cancer Res Treat.* 2018. 171533033818808507. DOI: 10.1177/1533033818808507.
5. Taieb J, Svrcek M, Cohen R, Basile D, Tougeron D, Phelip JM. Deficient mismatch repair/microsatellite unstable colorectal cancer: Diagnosis, prognosis and treatment. *Eur J Cancer.* 2022. 175136-57. DOI: 10.1016/j.ejca.2022.07.020.
6. Kim JK, Chen CT, Keshinro A, Khan A, Firat C, Vanderbilt C, *et al.* Intratumoral T-cell repertoires in DNA mismatch repair-proficient and -deficient colon tumors containing high or

low numbers of tumor-infiltrating lymphocytes. *Oncoimmunology*. 2022. 11(1), 2054757, DOI: 10.1080/2162402X.2022.2054757.

7. Sherman SK, Schuitevoerder D, Chan CHF, Turaga KK. Metastatic Colorectal Cancers with Mismatch Repair Deficiency Result in Worse Survival Regardless of Peritoneal Metastases. *Ann Surg Oncol*. 2020. 27(13), 5074-83. DOI: 10.1245/s10434-020-08733-x.
8. Jin Z, Sinicrope FA. Mismatch Repair-Deficient Colorectal Cancer: Building on Checkpoint Blockade. *J Clin Oncol*. 2022. 40(24), 2735-50. DOI: 10.1200/JCO.21.02691.
9. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. 2005. 23(3), 609-18. DOI: 10.1200/JCO.2005.01.086.
10. van der Heide DM, Turaga KK, Chan CHF, Sherman SK. Mismatch Repair Status Correlates with Survival in Young Adults with Metastatic Colorectal Cancer. *J Surg Res*. 2021. 266104-12. DOI: 10.1016/j.jss.2021.03.040.
11. Gong Q, Zhang HH, Sun SB, Ge WM, Li Y, Zhu YC, *et al*. Mismatch repair-deficient status associates with favorable prognosis of Eastern Chinese population with sporadic colorectal cancer. *Oncol Lett*. 2018. 15(5), 7007-13. DOI: 10.3892/ol.2018.8192.
12. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep*. 2021. 11(1), 6968. DOI: 10.1038/s41598-021-86327-7.
13. Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*. 2020. 21(212), 1-6.
14. Fang L, Yang Z, Zhang M, Meng M, Feng J, Chen C. Clinical characteristics and survival analysis of colorectal cancer in China: a retrospective cohort study with 13,328 patients from southern China. *Gastroenterol Rep (Oxf)*. 2021. 9(6), 571-82. DOI: 10.1093/gastro/goab048.
15. Zhong JW, Yang SX, Chen RP, Zhou YH, Ye MS, Miao L, *et al*. Prognostic Value of Lymphovascular Invasion in Patients with Stage III Colorectal Cancer: A Retrospective Study. *Med Sci Monit*. 2019. 256043-50. DOI: 10.12659/MSM.918133.