www.oncotarget.com Oncotarget, 2025, Vol. 16, pp: 834-847

Research Paper

Machine learning-based survival prediction in colorectal cancer combining clinical and biological features

Lucas M. Vieira[1,2], Natasha A.N. Jorge[3], João B. Sousa[4], João C. Setubal[5], Peter F. Stadler[3] and Maria E.M.T. Walter[1]

[1]Department of Computer Science, University of Brasília, Campus Universitario Darcy Ribeiro, Prédio CIC/EST, Brasília, DF 71910-900, Brazil

[2]Current Affiliation - Department of Pharmacology, School of Medicine, University of California San Diego, California, CA 92093, USA

[3]Bioinformatics, Institute for Informatics, Leipzig University, Leipzig, Saxony 610101, Germany

[4]Division of Coloproctology, Department of Surgery, University of Brasilia, Campus Universitario Darcy Ribeiro, Faculdade de Medicina, Brasília, DF 70910-900, Brazil

[5]Institute of Chemistry, Department of Biochemistry, University of São Paulo, Av. Prof. Lineu Prestes, São Paulo, SP 05508-000, Brazil

Correspondence to: *Lucas M. Vieira,* ***email***: *lvieira@health.ucsd.edu*

Keywords: colorectal cancer; machine learning; feature selection; non-coding RNAs; genes

Received: *April 19, 2025* ***Accepted***: *November 24, 2025* ***Published***: *December 15, 2025*

ABSTRACT

Colorectal cancer (CRC) is one of the most common and lethal types of cancer worldwide. Understanding both the biological and clinical aspects of the patient is essential to uncover the mechanism underlying the prognosis of the disease. However, most current approaches focus primarily on clinical or biological elements, which can limit their ability to capture the full complexity of the prognosis of CRC. This study aims to enhance understanding of the mechanisms of CRC by combining clinical and biological data from CRC patients with machine learning techniques (ML) to explore the importance of features and predict patient survival. First, we performed differential expression analysis and inspected patient survival curves to identify relevant biological features. Then, we applied ML techniques to understand the individual impact of each clinical and biological feature on patient survival. *E2F8*, *WDR77*, and *hsa-miR-495-3p* stood out as biological features, while pathological stage, age, new tumor event, lymph node count, and chemotherapy have shown themselves as interesting clinical features. Furthermore, our ML model achieved an accuracy of 89.58% to predict patient survival. The clinical and biological features proposed here in conjunction with ML can improve the interpretation of CRC mechanisms and predict patient survival.

INTRODUCTION

Colorectal cancer (CRC) is one of the most common and lethal cancers in the world, accounting for around 10% of all cancer diagnoses in the world [1–4]. CRC occurs in the digestive tract, specifically in the colon, rectum, and rectosigmoid junction. The behavior and treatment of CRC can differ according to its anatomical site [5]. Although prognosis, prevention, and treatment have advanced due to the growing number of people diagnosed with CRC each year, a better understanding of the mechanisms of CRC development and progression continues to be crucial [3, 6–8].

Given the importance of mRNAs, miRNAs, and lncRNAs in cancer, recent studies also show the importance of their underlying interaction system in cancer progression, the so-called competing endogenous RNAs (ceRNAs) mechanism [9–12]. A ceRNA network can play an essential role in cancer development [13–15]. Therefore, exploring miRNAs, lncRNAs, mRNAs, and the ceRNA

networks formed by them, together with clinical factors, could lead to a better understanding of the underlying mechanisms of CRC [16].

In this context, this article aims to present a method to predict patient survival in CRC by highlighting biological and clinical markers to characterize CRC behavior (and help in patient prognosis), taking into account the different anatomical sites: colon, rectum, and rectosigmoid junction. In more detail, it is common knowledge that interactions among proteins, miRNAs, and lncRNAs affect cancer since they can regulate suppressive and oncogenic functions in various types of cancer [14]. As is already known, understanding these mechanisms can help prevent tumor emergence and cancer development, as well as facilitate its identification. Although several studies present relationships between protein and ncRNAs with cancer [17–22], a few focus on predicting the prognosis of cancer patients using computational techniques with data from protein markers, miRNA, lncRNA, and clinical characteristics of patients [23–25], despite the fact that several databases present disease-related information [26–29] and specific information about cancer [30–34]. In recent years, research has been developed on different aspects of CRC, such as gene biomarkers for diagnosis [4, 35], prediction models for prognostics [36–38] and patient survival [39–42]. Recently, some interesting reviews about CRC and machine learning (ML) have been published [43, 44].

Other studies relate the importance of patient clinical aspects for cancer, e.g., the impact of race, age, and demographics on CRC emergence behavior [42, 45–47], but few [44, 48] explore the importance of these clinical aspects in combination with biological aspects to predict CRC prognosis and patient survival through machine learning.

RESULTS

In this section, we first describe the data after the pre-processing execution, then the selected features, and finally, the results of the model construction.

Data pre-processing

Using the method detailed in Section *Data pre-processing*, biological and clinical features were associated with a patient. For clinical features, it was noticed that, in many cases, information necessary to collect clinical metadata was missing. Of these features, the ones with the most missing values were race, ethnicity, weight, and height.

To address this problem, first, the clinical features were divided into two groups. Group (i) included age at initial pathological diagnosis, gender, number of lymph nodes, number of positive lymph nodes, chemotherapy, pathologic stage, vital status, and new tumor event. Group (ii) included race, ethnicity, weight, and height.

Then, considering these two groups of clinical features, data was grouped into three cases in the *missing features handler* step:

- Case 1. Filtered data with missing biological or Group (i) clinical features;
- Case 2. Filtered data with missing biological or Group (i) or Group (ii) clinical features; and
- Case 3. All data but replacing missing clinical features by using the most frequent value. In this case, missing values were filled using the most frequent value in the dataset. For example, if a patient's race was missing and the most common race in the data was "white", the patient was assigned to "white". This approach was based on specialist recommendations, as features like race have fixed categories (e.g., "white" or "non-white"), and other imputation techniques, such as mean or median, could introduce non-existent values.

After filtering and transforming the data for each case, as described, the number of patients was: (i) for Case 1, 357 with colon cancer, 74 with rectum cancer, and 63 with rectosigmoid junction cancer; (ii) for Case 2, 177 with colon cancer, 27 with rectum cancer and 33 with rectosigmoid junction cancer; and (iii) for Case 3, 391 with colon cancer, 85 with rectum cancer, and 69 with rectosigmoid junction cancer. With all the features set up, prediction models were constructed for Cases 1, 2, and 3.

It is also important to note that, for the feature *new tumor event*, TCGA only indicates whether such an event occurred, without distinguishing between a metachronous tumor and a recurrence of the original tumor. Given that the incidence of metachronous colorectal cancer in sporadic populations is approximately 3–5% overall [49–51], and that only 20% of the selected cases were annotated as having a new tumor event, it is reasonable to assume that, on average, most of these events in our cohort represent recurrences rather than true metachronous tumors.

Feature selection

During the feature selection phase, a grid search was applied to Least Absolute Shrinkage and Selection Operator (LASSO), in conjunction with a five cross-fold validation, to optimize its parameters and provide a more reliable process. Figure 1 - Case 1 shows the 10 features selected as important for Case 1. Among these features, six were clinical: pathological stage, number of lymph nodes, age, number of positive lymph nodes, new tumor event, and chemotherapy; and four were biological features: *E2F8*, *hsa-miR-495-3p*, *WDR77*, and *KCNQ1OT1*. When using Shapley Additive Explanations (SHAP) to improve understanding, it is possible to notice

the impact of features such as pathological stage and age, which are directly related to a patient's survival, as the greater its values are, the greater the chance of this patient not surviving.

Figure 1 - Case 2 shows the 7 features selected as important for Case 2. Among these features, five were clinical: pathological stage, new tumor event, number of lymph nodes, weight, age, and chemotherapy; and two

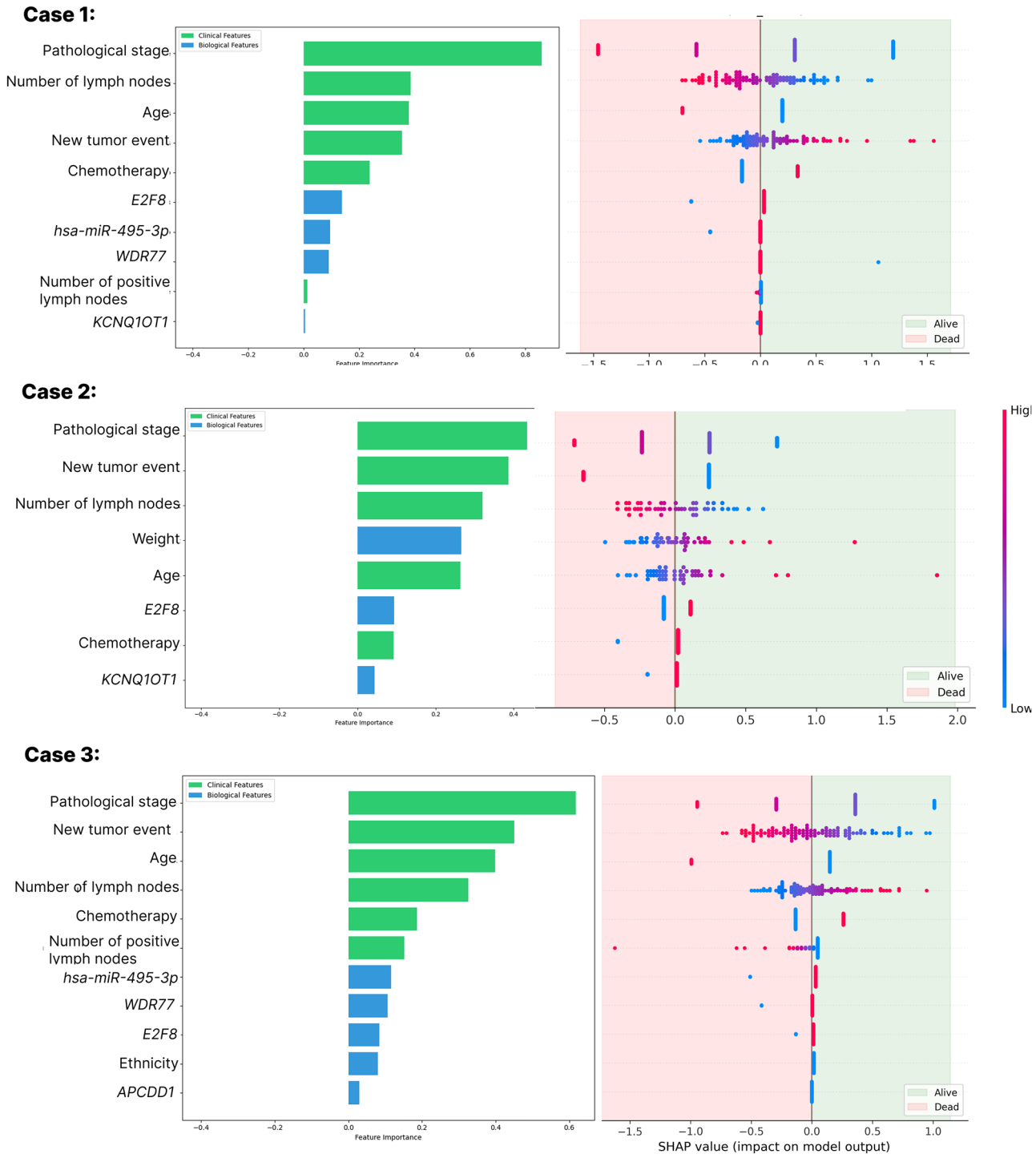

Figure 1: LASSO feature ranking and SHAP explanatory for Cases 1, 2, and 3 feature selection models. A positive SHA*P* value indicates a positive impact on prediction, leading the model to predict 1 (Patient survival). A negative value indicates an adverse effect, leading the model to predict 0 (Patient non-survival). The color of the SHAP data points shows the values as a heatmap where blue is the lowest value (e.g., 0) and red is the highest value (e.g., 1). For Cases 1 and 2, pathological stage and E2F8 expression are the most relevant clinical and biological features respectively. On the other hand, for group 3, pathological stage and hsa-miR-495-3p expression are the most relevant features.

were biological features: *E2F8* and *KCNQ1OT1*. In the analysis with SHAP, it is possible to notice the impact of features such as pathological stage, which is directly related to a patient's survival, as the greater its values are, the greater the chance of this patient not surviving. Unlike Case 1, here, the age feature is inversely related to patient survival.

Figure 1 - Case 3 shows the 11 features selected as important for Case 3. Among these features, seven were clinical: pathological stage, new tumor event, age, number of lymph nodes, chemotherapy, number of positive lymph nodes, and ethnicity; and four were biological features: *hsa-miR-495-3p*, *WDR77*, *E2F8*, and APCDD1. With SHAP, it is possible to notice the impact of features such as pathological stage and age, which is directly related to patient survival, as the greater its values are, the greater the chance of this patient not surviving. Unlike Case 1 and similarly to Case 2, here, the *age* feature is inversely related to patient survival.

Figure 2 shows the standard features among the three groups. Compared to individual results, it is noticeable that the feature selection varies across the groups. Particularly for biological features, only the molecule *E2F8* was consistently ranked as important in all the models.

Finally, similar behavior is observed in Groups 1 and 3, including similarities in clinical and biological features, such as *WDR77* and *hsa-miR-495-3p*. This is different from Case 2, which may indicate that, when filtering with Case 2, the amount of remaining data may affect the prediction and selection of features.

Model construction

After using LASSO in Phase 2, the features were used as input to construct the prediction model. During this phase, a grid search was applied in conjunction with a five cross-fold validation to optimize its parameters and prediction accuracy. Table 1 shows the performance evaluation of all the models constructed to predict patient survival using the selected features.

The SVM model led to the best results in predicting patient survival for Case 1, achieving an accuracy of 86.87% on the test data, with 83.49% of

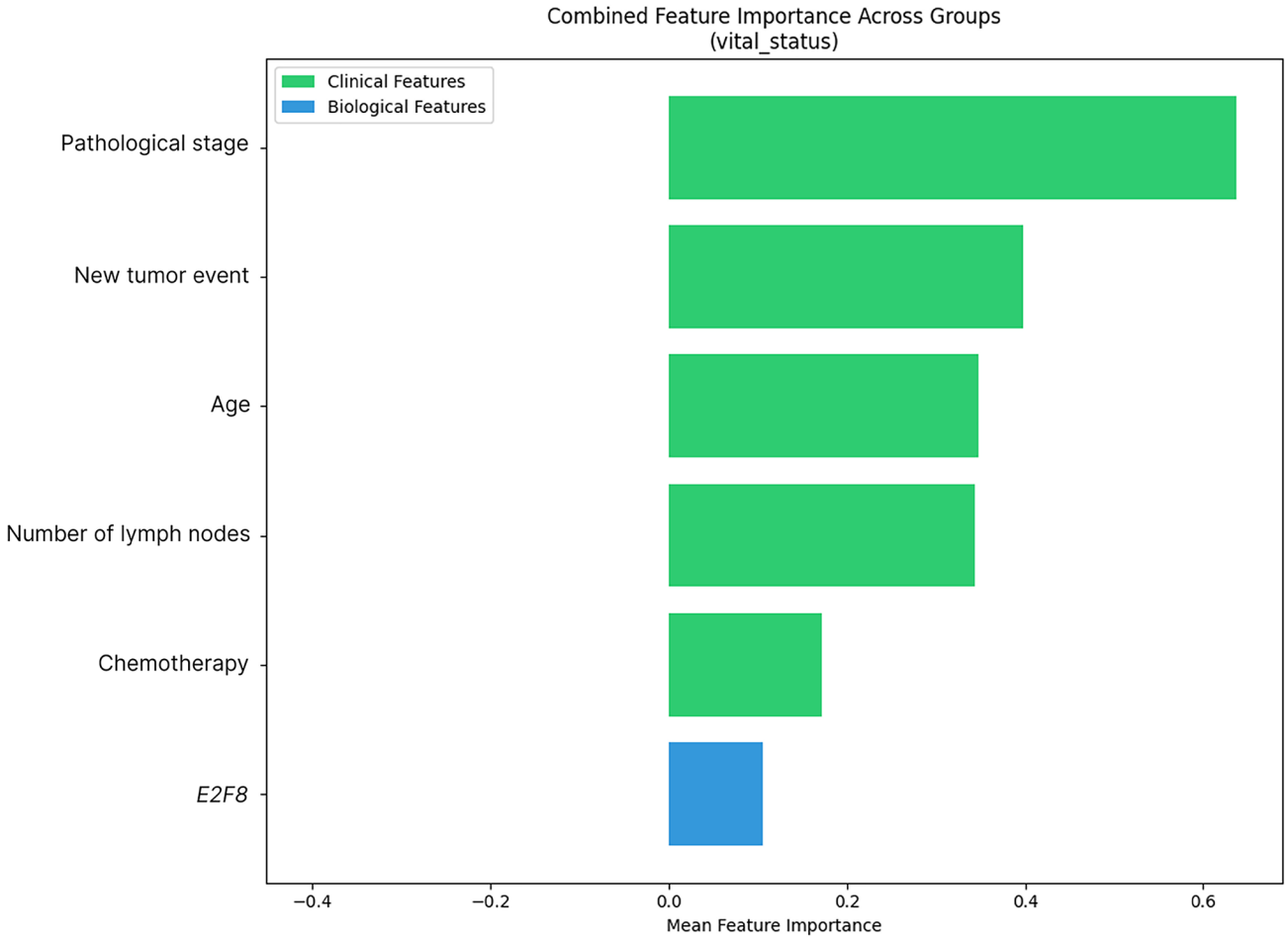

Figure 2: LASSO ranking for features relevant in groups 1, 2, and 3. The features are displayed with a mean importance value greater than 0 in all three groups. The meaning of importance scores of these selected features were computed and displayed to highlight their significance among the groups.

Table 1: Model performance for each group

Case	Model	Accuracy (%)	AUC (%)	Best parameters
1	LR	79.79%	81.77%	C: 0.1, weight: {0: 0.7, 1: 1.3}, solver: liblinear
	SVM	**86.87%**	**83.49%**	C: 0.1, weight: balanced, kernel: linear
	RandomForest	83.84%	79.37%	Weight: balanced, max depth: None, n: 200
	AdaBoost	80.81%	81.05%	Learning rate: 0.1, n: 50
	Stacking	81.82%	78.03%	ada n: 100, final estimator C: 1.0, rf n: 100
	Voting	80.81%	79.21%	ada n: 100, rf n: 100, svm C: 1
2	LR	70.83%	**82.05%**	C: 0.01, weight: {0: 0.7, 1: 1.3}, solver: liblinear
	SVM	77.08%	81.20%	C: 0.1, weight: balanced, kernel: linear
	RandomForest	85.42%	73.79%	Weight: balanced, max depth: None, n: 100
	AdaBoost	**89.58%**	76.50%	Learning rate: 0.1, n: 200
	Stacking	81.25%	77.49%	ada n: 100, final estimator C: 1.0, rf n: 100
	Voting	79.17%	74.07%	ada n: 100, rf n: 100, svm C: 1
3	LR	79.81%	79.02%	C: 0.01, weight: balanced, solver: lbfgs
	SVM	77.98%	**80.41%**	C: 0.1, weight: balanced, kernel: linear
	RandomForest	81.65%	76.25%	Weight: balanced, max depth: 20, n: 200
	AdaBoost	81.65%	76.44%	Learning rate: 0.1, n: 50
	Stacking	79.82%	78.84%	ada n: 100, final estimator C: 0.1, rf n: 100
	Voting	**82.57%**	78.27%	ada n: 100, rf n: 100, svm C: 1

The highest accuracy and AUC are highlighted in bold for each group.

AUC. The AB model led to the best accuracy for Case 2, achieving an accuracy of 89.58% and an AUC of 76.50%. The Voting model led to the best accuracy for Case 3, achieving an accuracy of 82.57% and an AUC of 78.27%. As shown, both the classical linear and the ensemble models led to a good performance in predicting patient survival using clinical and biological data. Still, the ensemble methods, in particular the Voting and Stacking approaches, display the overall best performance between groups.

Finally, the bootstrap analysis reveals varying levels of uncertainty among the adopted models and metrics. For Case 1, SVM shows the best performance with relatively narrow confidence intervals (±4.7% for accuracy). In Group 2, the AB model achieves the highest overall accuracy (89.58%), with wider confidence intervals (±7.0% for accuracy), indicating less certainty in its performance. In Group 3, the Voting classifier demonstrates the most consistent performance with narrow margins (±3.9% for accuracy). Statistical significance testing through bootstrap confidence intervals confirms that advanced ML models (particularly SVM, AB, and Voting classifiers) provide meaningful accuracy improvements over baseline logistic regression in all groups. However, AUC differences are generally not statistically significant due to substantial confidence interval overlaps across models.

DISCUSSION

Feature selection and ML methods are widely used to understand large volumes of data better and to generate information. With the significant growth of biological data on CRC and the amount of information that can be extracted from these data to study CRC prognosis, the use of feature extraction techniques is of interest in improving ML methods [52, 53].

In this work, we used feature selection to identify biological and clinical features that are relevant to CRC patient survival. Also, ML models were created to predict patient survival, which can help doctors better understand key points in CRC prognosis. The proposed method combines biological and clinical features to predict patient survival, using as input data from patients from the United States, available in the TCGA database.

During the feature selection phase, using LASSO and SHAP, the results of this work suggest that the combination of biological and clinical features can help to understand and predict a patient's prognosis. The biological feature with the most significant average impact in all cases was the *E2F8* gene, which was also identified by other studies as a potential CRC biomarker and is significantly associated with cell proliferation and indicates the CRC stage [54–56]. Although not present in Case 2, *WDR77* and *hsa-miR-495-3p* were also shown as

relevant biological features for most of the groups, which was also identified by other studies [57, 58] as important in cancer or CRC development. Finally, age, pathological stage, chemotherapy, and lymph node count, which are clinical features previously identified as relevant in CRC development [59–63] were also highlighted as important in our study, even when combined with biological features.

Given these findings from the feature selection phase, our study supports the relevance of well-known clinical factors while also highlighting novel molecular factors. Furthermore, integrating clinical and molecular information, by combining established and newly identified features, improved the accuracy of prognostic prediction in colorectal cancer and contributed to refining prognosis across different stages of disease progression.

In the model construction phase, as expected with datasets storing only a few hundred cases and imbalanced labels, where non-survival cases are more prevalent, our bootstrapped confidence intervals indicated margins of approximately 5% to 10%. This, in comparison with the LR model used as a baseline, suggests that the ML models delivered meaningful improvements. Specifically, accuracy increased by 4.6% to 11.1%, while AUC showed only a low improvement of up to 4%. Although the results suggest a degree of similarity between LR and the more complex models for AUC, this can be attributed to the use of LASSO for feature selection, which can substantially enhance the performance of LR.

The ensemble methods and SVM consistently outperformed baseline LR for survival prediction. Using the ensemble AB model, we had the best accuracy of 89.58%, with an AUC of 76.50%. For patient survival prediction, AUC may be a more suitable metric due to the imbalance in the datasets since identifying high-risk patients is more critical than achieving high overall prediction accuracy. However, in this case, the ensemble methods demonstrated more excellent stability between both Accuracy and AUC. This suggests that combining simple models or even complex ones, such as SVM, RF, and AB, with a Voting or Stacking strategy can reduce overfitting since the metrics indicate more consistent performance. Focusing on stability across these metrics provides valuable insights into the robustness of the models, particularly in scenarios where no single metric fully captures performance.

Previous studies [24, 44, 52, 53] devised models to predict CRC-related outcomes through a variety of ML techniques. Alboaneen et al. [44] provided a systematic review of the application of ML for CRC detection and diagnosis, highlighting that ensemble methods such as RF can achieve strong performance. However, the results were different, depending on the dataset type and the pre-processing. Buk et al. [42] used clinical features and data from Brazilian CRC patients to predict survival, achieving an accuracy of 77% and AUC of 86% using RF. They also applied SHAP explanations to identify the most important features, among which were clinical stage and age. Achilonu et al. [24] created a pipeline using clinical features to predict CRC survival in South African patients. Specifically, their best model used an artificial neural network (ANN) and achieved an accuracy of 82.0%. Kang et al. [53] used LASSO to select biological and clinical features, such as age and sex, to predict lymph node metastasis in CRC. It obtained better results using LASSO than models without it, achieving the best AUC of 79.5%. In particular, Su et al. [52] also used TGCA-COAD data and applied LASSO to select biological features, along with SVM, RF, and DT, to predict colon cancer diagnosis and staging. For cancer diagnosis, they achieved high accuracy, reaching 98%. However, cancer staging proved to be more challenging, with a maximum accuracy of 91% and an AUC of 82%. Although these performance metrics might indicate overfitting, the techniques used have proven to be useful.

It is worth noting that the methods described in these studies, including our research, used different data and features as input and presented relatively good accuracy in predicting specific patient prognosis factors. ML methods use various clinical features to predict CRC prognosis targets [23, 24, 42, 44]. Achilonu et al. [24] and Gupta et al. [25] show that clinical features, such as age, gender, race, recurrence, chemotherapy, smoking, and alcohol consumption, can also lead to reasonable accuracy in predicting CRC prognosis factors. Our study identified some of the clinical features proposed in the cited articles as relevant, such as age, gender, race, recurrence, and chemotherapy. Although smoking and alcohol consumption have been shown to be appropriate in the cited works [23–25], these clinical features were not included in this study because their values were missing from the available TCGA data for most patients.

ML algorithms showed overall performance, particularly stacking, which displayed good overall results. Furthermore, the results of our work suggest that the combination of biological and clinical features can help predict patient prognosis. The biological feature with the most significant average impact in all cases was the *E2F8* gene, which was also identified by other studies as a potential CRC biomarker and significantly associated with cell proliferation and stages of CRC [54–56]. Although not present in Case 2, *WDR77* and *hsa-miR-495-3p* were also shown as relevant biological features for most of the groups, which was also identified by other studies [57, 58] as important in cancer or CRC development. Finally, age, pathological stage, chemotherapy, and lymph node count, which are clinical features previously identified as relevant in CRC development [59–63] were also highlighted as important in our study, even when combined with biological features.

Our study has some limitations. Initially, although several novel lncRNAs, mRNAs, and miRNAs with

clinical significance for CRC were found, the study was developed using TCGA data and no further experimental validation was carried out. It is also essential to observe that TCGA consists of data collected exclusively from patients in the United States. In addition, some clinical features known to be relevant for CRC prognosis, such as smoking status and body mass index (BMI), were excluded from the training features due to missing information in most cases. For BMI, weight and height were available in some instances and showed trends supporting its prognostic importance, but the incomplete data limited their inclusion. The relatively small dataset (~545 patients), the use of only open-source data, and the imputation of missing categorical variables (e.g., race) also represent important constraints.

As acknowledged, these limitations underscore the value of dividing the data into groups for further exploration, which helped highlight the importance of collecting additional data, even though potential bias may have arisen from the imputation of missing categorical variables. CRC was treated as a single disease in our prediction instead of dividing it into its anatomical sites (colon, rectum, and rectosigmoid junction) to mitigate this limiting factor. This study also showed that even basic patient information, such as age and weight, when accurately recorded by doctors and stored on databases, can strongly contribute to improving CRC prognosis studies.

Finally, the development of this analysis with data collected from patients of other countries could give doctors a regional-specific view and a better understanding of CRC-specific characteristics for each anatomical site as potentially related to the region where patients live. Research on biological and clinical features in CRC is still under development. Further experimental studies and a more significant amount of CRC data are required to improve understanding of CRC prognosis.

In conclusion, this study proposed a model to understand biological and clinical features and predict CRC patient survival. For the best overall, the AB model resulted in an accuracy of 89.58% and an AUC of 76.50%. Furthermore, the results of this work suggest that the combination of biological and clinical features can help to predict patient prognosis. These results highlight evidence for prospective research of CRC associating *E2F8* as a potential prognostic biomarker and the importance of clinical features, e.g., age. The findings of this study can contribute to further studies on CRC using bioinformatic and ML techniques.

MATERIALS AND METHODS

The model is designed to predict CRC patient survival. In practice, survival data record whether a patient survived after treatment until the last known medical appointment. The pipeline (Figure 3) is composed of three main phases: (1) data pre-processing; (2) feature selection, in which features are ranked and then filtered; and (3) model construction, in which the prediction model is constructed and evaluated. The pipeline was implemented in Python, using the scikit-learn [64] package for the ML algorithms implementation.

The input for the method was constructed with biological and clinical information downloaded from two databases, using the GDC interface (https://portal.gdc.cancer.gov/), TCGA rectal adenocarcinoma (TCGA-COAD)[1]; and TCGA rectal adenocarcinoma (TCGA-READ)[2]. Data were exclusively selected from adenocarcinoma, the most common type of CRC, and filtered to minimize variance by removing possible outlier cases. Therefore, the expression raw count data was collected from miRNA-seq and RNA-seq files at TCGA, with 541 primary tumors (TP) and 48 non-tumor tissues (NT), from 545 patients, whose 391 had colon cancer, 85 had rectum cancer and 69 had rectosigmoid junction cancer. The ages of the patients ranged from 31 to 90 years old, with an average age of 66. Of these, 185 patients (34%) received chemotherapy, 105 (19%) had a relapse, and 108 patients (20%) died. Details of the data can be seen on the project's GitHub[3].

[1] https://portal.gdc.cancer.gov/projects/TCGA-COAD
[2] https://portal.gdc.cancer.gov/projects/TCGA-READ
[3] https://github.com/lmacielvieira/crc-bio-cli-ml

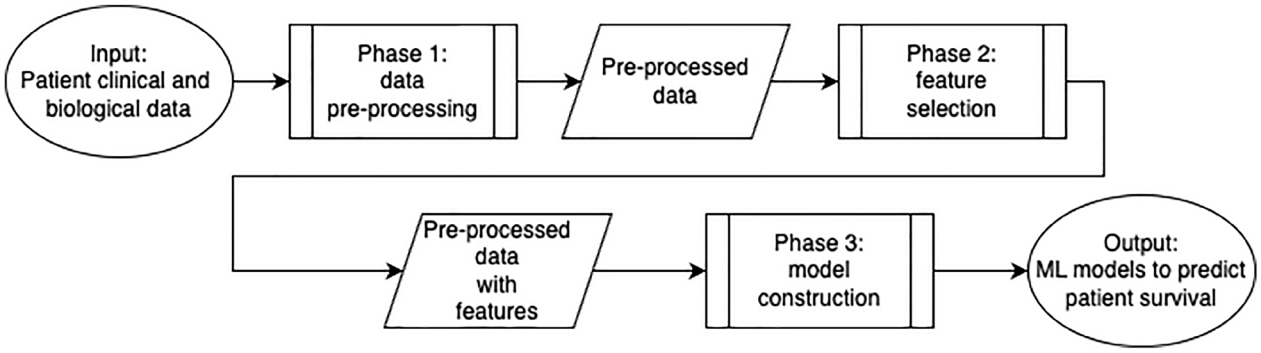

Figure 3: A method to predict CRC patient survival. The pipeline is divided into three main phases: (1) data pre-processing, in which the patient's clinical and biological data is processed, (2) feature selection, in which clinical and biological features are ranked and then filtered; and (3) model construction, in which the prediction model is constructed and evaluated.

Table 2: Candidate molecules to be used as biological features in the ML model to predict CRC patient survival

Molecule	Type	Potential roles in CRC
ANKRD6	Gene	Immune invasion [70]
APCDD1	Gene	CRC recurrence [71]
DMD	Gene	Lymph node metastasis [72]
E2F8	Gene	Cell proliferation [56]
H19	lncRNA	Cell migration and invasion [73]
HECW2	Gene	CRC progression [74]
HOXD13	Gene	CRC progression [75]
KCNQ1OT1	lncRNA	Chemo resistance [76]
LINC00894	lncRNA	Cell proliferation [77]
NRG1	Gene	Tumorigenesis [78]
RANBP1	Gene	CRC progression [79]
SNHG16	lncRNA	Cell growth [80]
TMEM198	Gene	CRC prognosis [65]
UST	Gene	CRC prognosis [65]
WDR77	Gene	Cell proliferation [57]
hsa-miR-1271-5p	miRNA	Cell proliferation [81]
hsa-miR-130a-3p	miRNA	Cell proliferation [82]
hsa-miR-130b-3p	miRNA	Cell growth [83]
hsa-miR-495-3p	miRNA	Cell proliferation [58]

Data pre-processing

The data pre-processing phase consists of three steps: (1) *feature extraction*, which extracts the clinical and biological features from the input data; (2) *normalization*, which normalizes the clinical and biological data to numerical values; and (3) *missing features handler*, which construct the cases to be analyzed, taking into account the missing features of the data.

The feature extraction step uses the input data described previously to extract, process, and associate biological and clinical features for each patient. For biological features, as proposed by Vieira et al. [65], a customized script in R was developed. This custom script performs differential expression analysis to identify potential differentially expressed (DE) molecules. Based on these DE molecules, it constructs competing endogenous RNA (ceRNA) networks and subsequently conducts survival analysis using the DE molecules involved in these interactions. The output consists of candidate biological markers associated with patient survival, which are then used as biological features in downstream analyses.

The differential expression analysis was performed using GDCRNATools v1.6 [66], incorporating the limma package and voom normalization [67], with the thresholds $FDR \leq 0.05$ and $|logFC| \geq 2$. The ceRNA networks were constructed by predicting mRNA–miRNA–lncRNA interactions using the spongeScan algorithm [68] in conjunction with the StarBase v2.0 database [69], via a built-in function in GDCRNATools. The survival analysis was conducted using Cox proportional hazards (CoxPH) modeling and Kaplan–Meier (KM) estimation to calculate hazard ratios and generate survival curves, with ($p < 0.05$) considered statistically significant.

Finally, we select target biomarkers, in which: (i) the molecules are differentially expressed (DE); (ii) the biomarkers are presented in the CRC ceRNA networks; and (iii) the biomarkers affect patient survival. These criteria guaranteed the selection of molecules with a potential role in the CRC patient prognosis [65] and led to the compilation of a list of 19 molecules, as shown in Table 2.

To parameterize these molecules as biological features, we developed a customized Python script using the expression data calculated in the previous script as input. This script utilized two key elements from the input: the molecule's average voom-normalized expression, and its normalized counts for each patient. Using this information, the script determined whether each molecule was overexpressed in a patient by checking if its normalized expression count was higher than the normalized average expression $E_{patient} > AveExpr$.

To identify clinical characteristics, we proceeded as follows. First, the raw clinical metadata available

Table 3: List of numerical values used in the feature vector

Feature	Associated values
Age	Numerical value = age of the patient
Chemotherapy	1 = received chemo; 0 = did not receive chemo
Ethnicity	1 = Latino; 0 = non Latino
Gender	0 = female; 1 = male
Height	Numerical value = height of the patient
Race	1 = non-white; 0 = white
Pathological stage	Stage IV = 3; stage III = 2; stage II = 1; stage I = 0
Vital status	1 = survival; 0 = non-survival
Number of positive lymph nodes	Numerical value = number of lymph nodes
Number of lymph nodes	Numerical value = number of positive lymph nodes
Weight	Numerical value = weight of the patient
New tumor event	1 = new tumor; 0 = no new tumor
ANKRD6	1 = overexpressed; 0 = not overexpressed
APCDD1	1 = overexpressed; 0 = not overexpressed
DMD	1 = overexpressed; 0 = not overexpressed
E2F8	1 = overexpressed; 0 = not overexpressed
H19	1 = overexpressed; 0 = not overexpressed
HECW2	1 = overexpressed; 0 = not overexpressed
HOXD13	1 = overexpressed; 0 = not overexpressed
KCNQ1OT1	1 = overexpressed; 0 = not overexpressed
LINC00894	1 = overexpressed; 0 = not overexpressed
NRG1	1 = overexpressed; 0 = not overexpressed
RANBP1	1 = overexpressed; 0 = not overexpressed
SNHG16	1 = overexpressed; 0 = not overexpressed
TMEM198	1 = overexpressed; 0 = not overexpressed
UST	1 = overexpressed; 0 = not overexpressed
WDR77	1 = overexpressed; 0 = not overexpressed
hsa-miR-1271-5p	1 = overexpressed; 0 = not overexpressed
hsa-miR-130a-3p	1 = overexpressed; 0 = not overexpressed
hsa-miR-130b-3p	1 = overexpressed; 0 = not overexpressed
hsa-miR-495-3p	1 = overexpressed; 0 = not overexpressed

at TCGA was analyzed. These clinical features were divided into 9 groups - clinical, demographic, diagnosis, exposure, family history, follow-up, molecular test, pathological details, and treatment. A doctor specialist in CRC assisted in the process of manually choosing the most relevant characteristics from the available data. The following features were selected: age at initial pathological diagnosis, ethnicity, gender, race, vital status, number of positive lymph nodes, number of lymph nodes, pathological stage, weight, height, chemotherapy, new tumor event, and vital status.

To normalize and prepare the data to be used in the prediction models, in the *normalization* step, the clinical and biological features were transformed into numerical values, as shown in Table 3. These numerical values were later used in the charts to show the importance of the features.

Finally, in the *missing features handler* step, the data points with any missing feature were removed or had their values replaced by the feature median value. Experiments were developed for both cases and are detailed in Section *Results*.

Feature selection

The feature selection phase consisted of two steps using Least Absolute Shrinkage and Selection Operator

(LASSO): (1) *feature ranking*, in which the grid search is used to select the best training parameters and to rank each biological and clinical feature by importance in the final prediction, and (2) *feature selection*, in which the most relevant features are selected.

After using LASSO for feature selection, Shapley Additive Explanations (SHAP) was used to assess the impact of each selected feature on the prediction. This was done to provide an intuitive understanding of how each predictor contributes to classifying the patient's vital status.

Model construction

The model construction phase consisted of five steps: (1) *data split*, which divides the pre-processed data into training and testing data in an 80% and 20% ratio; (2) *feature selection*, in which the features obtained at Phase 2 (feature selection) are selected; (3) *parameter optimization*, where grid search and cross-validation were used to optimize the ML hyperparameters; (4) *ML classifiers construction*, in which ML models using Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Adaptative Boosting (AB), Stacking (SE), and Voting (VE) are built; and (5) *performance evaluation*, in which the ML classifiers are evaluated and compared.

In the *ML classifiers construction* step, the six classifiers (LR, SVM, RF, AB, SE, and VE) were used since each behaves differently based on the pattern of the input data. Therefore, the goal was to explore these classifiers to find the best option to predict the expected outcome. In particular, the SE and VE classifiers used the combination of SVM, RF, and AB, adopting an ensemble voting and stacking strategy to verify whether using multiple classifiers together could improve the outcome. Then, in the *performance evaluation* step, the performance of each ML model was evaluated, taking the testing data as input and comparing the models through several metrics, such as AUC, accuracy, precision, and recall.

Finally, to validate the performance of the proposed model, the LR model was constructed as a baseline for comparative analysis due to its interpretability and simplicity. Furthermore, to address potential prediction errors arising from the limited number of data, stratified bootstrapping was adopted to preserve class balance, using 1,000 iterations and a confidence level of 95% to estimate prediction confidence intervals.

Abbreviations

AB: Adaptative Boosting; AUC: Area Under the Curve; ceRNAs: competing endogenous RNAs; CRC: Colorectal cancer; DE: differentially expressed; LASSO: Least Absolute Shrinkage and Selection Operator; LR: Logistic Regression; ML: Machine learning; RF: Random Forest.; SE: Stacking ensemble; SHAP: Shapley Additive Explanations; SVM: Support Vector Machine; VE: Voting ensemble.

Data availability

The data, custom scripts, and ML models built are all accessible at https://github.com/lmacielvieira/crc-bio-cli-ml.

AUTHOR CONTRIBUTIONS

L.M.V. contributed to the conception and design of the study, implemented the algorithms, and performed the analysis. J.B.S. collaborated with the discussion from a medical perspective. N.A.N.J. collaborated on bioinformatics methods and biological assumptions. J.C.S., M.E.M.T.W., and P.F.S. reviewed and collaborated on the methodology and discussion from a bioinformatics perspective. All the authors contributed to the article's writing. All authors approved the submitted version.

ACKNOWLEDGMENTS

The results shown here are based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

CONFLICTS OF INTEREST

Authors have no conflicts of interest to declare.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brasil, project number: 306947/2021-8. This study was supported in part by the German Federal Ministry of Education and Research BMBF through DAAD project 57616814 (SECAI, School of Embedded Composite AI) and jointly with the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the program Center of Excellence for AI-research *Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig*, project identification number: SCADS24B.

REFERENCES

1. Gupta S, May FP, Kupfer SS, Murphy CC. Birth Cohort Colorectal Cancer (CRC): Implications for Research and Practice. Clin Gastroenterol Hepatol. 2024; 22:455–69.e7. https://doi.org/10.1016/j.cgh.2023.11.040. [PubMed]
2. Kopetz S, Murphy DA, Pu J, Ciardiello F, Desai J, Van Cutsem E, Wasan HS, Yoshino T, Saffari H, Zhang X, Hamilton P, Xie T, Yaeger R, Tabernero J. Molecular

profiling of BRAF-V600E-mutant metastatic colorectal cancer in the phase 3 BEACON CRC trial. Nat Med. 2024; 30:3261–71. https://doi.org/10.1038/s41591-024-03235-9. [PubMed]

3. Freitas SC, Sanderson D, Caspani S, Magalhães R, Cortés-Llanos B, Granja A, Reis S, Belo JH, Azevedo J, Gómez-Gaviro MV, Sousa CT. New Frontiers in Colorectal Cancer Treatment Combining Nanotechnology with Photo-and Radiotherapy. Cancers (Basel). 2023; 15:383. https://doi.org/10.3390/cancers15020383. [PubMed]

4. Koppad S, Basava A, Nash K, Gkoutos GV, Acharjee A. Machine learning-based identification of colon cancer candidate diagnostics genes. Biology. 2022; 11:365. https://doi.org/10.3390/biology11030365.

5. Xie W, Zhong YS, Li XJ, Kang YK, Peng QY, Ying HZ. Postbiotics in colorectal cancer: Intervention mechanisms and perspectives. Front Microbiol. 2024; 15:1360225. https://doi.org/10.3389/fmicb.2024.1360225.

6. Li Q, Geng S, Luo H, Wang W, Mo YQ, Luo Q, Wang L, Song GB, Sheng JP, Xu B. Signaling pathways involved in colorectal cancer: pathogenesis and targeted therapy. Signal Transduct Target Ther. 2024; 9:266. https://doi.org/10.1038/s41392-024-01953-7. [PubMed]

7. Leiphrakpam PD, Are C. PI3K/Akt/mTOR Signaling Pathway as a Target for Colorectal Cancer Treatment. Int J Mol Sci. 2024; 25:3178. https://doi.org/10.3390/ijms25063178. [PubMed]

8. Gao R, Fang C, Xu J, Tan H, Li P, Ma L. LncRNA CACS15 contributes to oxaliplatin resistance in colorectal cancer by positively regulating ABCC1 through sponging miR-145. Arch Biochem Biophys. 2019; 663:183–91. https://doi.org/10.1016/j.abb.2019.01.005. [PubMed]

9. Ding J, Zhao Z, Song J, Luo B, Huang L. MiR-223 promotes the doxorubicin resistance of colorectal cancer cells via regulating epithelial-mesenchymal transition by targeting FBXW7. Acta Biochim Biophys Sin (Shanghai). 2018; 50:597–604. https://doi.org/10.1093/abbs/gmy040. [PubMed]

10. Thorenoor N, Faltejskova-Vychytilova P, Hombach S, Mlcochova J, Kretz M, Svoboda M, Slaby O. Long non-coding RNA ZFAS1 interacts with CDK1 and is involved in p53-dependent cell cycle control and apoptosis in colorectal cancer. Oncotarget. 2016; 7:622–37. https://doi.org/10.18632/oncotarget.5807. [PubMed]

11. Wang Q, Zhang H, Shen X, Ju S. Serum microRNA-135a-5p as an auxiliary diagnostic biomarker for colorectal cancer. Ann Clin Biochem. 2017; 54:76–85. https://doi.org/10.1177/0004563216638108. [PubMed]

12. Inoue A, Yamamoto H, Uemura M, Nishimura J, Hata T, Takemasa I, Ikenaga M, Ikeda M, Murata K, Mizushima T, Doki Y, Mori M. MicroRNA-29b is a Novel Prognostic Marker in Colorectal Cancer. Ann Surg Oncol. 2015 (Suppl 3); S1410–18. https://doi.org/10.1245/s10434-014-4255-8. [PubMed]

13. Lee YJ, Kim WR, Park EG, Lee DH, Kim JM, Shin HJ, Jeong HS, Roh HY, Kim HS. Exploring the Key Signaling Pathways and ncRNAs in Colorectal Cancer. Int J Mol Sci. 2024; 25:4548. https://doi.org/10.3390/ijms25084548. [PubMed]

14. Ye J, Li J, Zhao P. Roles of ncRNAs as ceRNAs in Gastric Cancer. Genes (Basel). 2021; 12:1036. https://doi.org/10.3390/genes12071036. [PubMed]

15. Conte F, Fiscon G, Sibilio P, Licursi V, Paci P. An Overview of the Computational Models Dealing with the Regulatory ceRNA Mechanism and ceRNA Deregulation in Cancer. Methods Mol Biol. 2021; 2324:149–64. https://doi.org/10.1007/978-1-0716-1503-4_10. [PubMed]

16. Ren JX, Chen L, Guo W, Feng KY, Cai YD, Huang T. Patterns of Gene Expression Profiles Associated with Colorectal Cancer in Colorectal Mucosa by Using Machine Learning Methods. Comb Chem High Throughput Screen. 2024; 27:2921–34. https://doi.org/10.2174/0113862073266300231026103844. [PubMed]

17. Ashouri K, Wong A, Mittal P, Torres-Gonzalez L, Lo JH, Soni S, Algaze S, Khoukaz T, Zhang W, Yang Y, Millstein J, Lenz HJ, Battaglin F. Exploring Predictive and Prognostic Biomarkers in Colorectal Cancer: A Comprehensive Review. Cancers (Basel). 2024; 16:2796. https://doi.org/10.3390/cancers16162796. [PubMed]

18. Peng WX, Koirala P, Mo YY. LncRNA-mediated regulation of cell signaling in cancer. Oncogene. 2017; 36:5661–67. https://doi.org/10.1038/onc.2017.184. [PubMed]

19. Zhang R, Xia LQ, Lu WW, Zhang J, Zhu JS. LncRNAs and cancer. Oncol Lett. 2016; 12:1233–39. https://doi.org/10.3892/ol.2016.4770. [PubMed]

20. Ma C, Nong K, Zhu H, Wang W, Huang X, Yuan Z, Ai K. H19 promotes pancreatic cancer metastasis by derepressing let-7's suppression on its target HMGA2-mediated EMT. Tumour Biol. 2014; 35:9163–69. https://doi.org/10.1007/s13277-014-2185-5. [PubMed]

21. Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, Mosquera JM, Pauwels J, Park K, Kossai M, MacDonald TY, Fontugne J, Erho N, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun. 2014; 5:5383. https://doi.org/10.1038/ncomms6383. [PubMed]

22. Xiang JF, Yin QF, Chen T, Zhang Y, Zhang XO, Wu Z, Zhang S, Wang HB, Ge J, Lu X, Yang L, Chen LL. Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res. 2014; 24:513–31. https://doi.org/10.1038/cr.2014.35. [PubMed]

23. Gründner J, Prokosch HU, Stürzl M, Croner R, Christoph J, Toddenroth D. Predicting clinical outcomes in colorectal cancer using machine learning. Building continents of knowledge in oceans of data: The future of co-created eHealth. USA: IOS Press Ebooks; 2018; 101–5. https://doi.org/10.3233/978-1-61499-852-5-101.

24. Achilonu OJ, Fabian J, Bebington B, Singh E, Eijkemans MJC, Musenge E. Predicting Colorectal Cancer Recurrence and Patient Survival Using Supervised Machine Learning Approach: A South African Population-Based Study. Front Public Health. 2021; 9:694306. https://doi.org/10.3389/fpubh.2021.694306. [PubMed]

25. Gupta P, Chiang SF, Sahoo PK, Mohapatra SK, You JF, Onthoni DD, Hung HY, Chiang JM, Huang Y, Tsai WS. Prediction of Colon Cancer Stages and Survival Period with Machine Learning Approach. Cancers (Basel). 2019; 11:2007. https://doi.org/10.3390/cancers11122007. [PubMed]

26. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013; 41:D983–86. https://doi.org/10.1093/nar/gks1099. [PubMed]

27. Gong J, Liu W, Zhang J, Miao X, Guo AY. lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. Nucleic Acids Res. 2015; 43:D181–86. https://doi.org/10.1093/nar/gku1000. [PubMed]

28. Ning S, Zhao Z, Ye J, Wang P, Zhi H, Li R, Wang T, Li X. LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. BMC Bioinformatics. 2014; 15:152. https://doi.org/10.1186/1471-2105-15-152. [PubMed]

29. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008; 36:D154–58. https://doi.org/10.1093/nar/gkm952. [PubMed]

30. The Cancer Genome Atlas Program (TCGA). National Cancer Institute. https://www.cancergenome.nih.gov.

31. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004; 6:1–6. https://doi.org/10.1016/s1476-5586(04)80047-2. [PubMed]

32. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004; 91:355–58. https://doi.org/10.1038/sj.bjc.6601894. [PubMed]

33. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, Gao Y, Guo M, Yue M, Wang L, Li X. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. 2016; 44:D980–85. https://doi.org/10.1093/nar/gkv1094. [PubMed]

34. Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics. 2013; 29:638–44. https://doi.org/10.1093/bioinformatics/btt014. [PubMed]

35. Vaziri-Moghadam A, Foroughmand-Araabi MH. Integrating machine learning and bioinformatics approaches for identifying novel diagnostic gene biomarkers in colorectal cancer. Sci Rep. 2024; 14:24786. https://doi.org/10.1038/s41598-024-75438-6.

36. Xiao B, Yang M, Meng Y, Wang W, Chen Y, Yu C, Bai L, Xiao L, Chen Y. Construction of a prognostic prediction model for colorectal cancer based on 5-year clinical follow-up data. Sci Rep. 2025; 15:2701. https://doi.org/10.1038/s41598-025-86872-5. [PubMed]

37. Tan Y, Hu B, Li Q, Cao W. Prognostic value and clinicopathological significance of pre-and post-treatment systemic immune-inflammation index in colorectal cancer patients: a meta-analysis. World J Surg Oncol. 2025; 23:11. https://doi.org/10.1186/s12957-025-03662-z. [PubMed]

38. Shimura T, Yin C, Ma R, Zhang A, Nagai Y, Shiratori A, Ozaki H, Yamashita S, Higashi K, Sato Y, Imaoka H, Kitajima T, Kawamura M, et al. The prognostic importance of the negative regulators of ferroptosis, GPX4 and HSPB1, in patients with colorectal cancer. Oncol Lett. 2025; 29:144. https://doi.org/10.3892/ol.2025.14890. [PubMed]

39. Wu T, Fang L, Ruan Y, Shi M, Su D, Ma Y, Ma M, Wang B, Liao Y, Han S, Lu X, Zhang C, Liu C, Zhang Y. Tumor aggression-defense index-a novel indicator to predicts recurrence and survival in stage II-III colorectal cancer. J Transl Med. 2025; 23:107. https://doi.org/10.1186/s12967-025-06141-x. [PubMed]

40. Wu SJ, Wu CY, Ye K. Risk factors, monitoring, and treatment strategies for early recurrence after rectal cancer surgery. World J Gastrointest Surg. 2025; 17:100232. https://doi.org/10.4240/wjgs.v17.i1.100232. [PubMed]

41. Tsai TJ, Syu KJ, Huang XY, Liu YS, Chen CW, Wu YH, Lin CM, Chang YY. Identifying timing and risk factors for early recurrence of resectable rectal cancer: A single center retrospective study. World J Gastrointest Surg. 2024; 16:2842–52. https://doi.org/10.4240/wjgs.v16.i9.2842. [PubMed]

42. Cardoso LB, Parro VC, Peres SV, Curado MP, Fernandes GA, Filho VW, Toporcov TN. Machine learning for predicting survival of colorectal cancer patients. Sci Rep. 2023; 13:8874. https://doi.org/10.1038/s41598-023-35649-9.

43. Onuiri EE, Akande O, Kalesanwo OB, Adigun T, Rosanwo K, Umeaka KC. A systematic review of machine learning prediction models for colorectal cancer patient survival using clinical data and gene expression profiles. Revue d'Intelligence Artificielle. 2023; 37:1273–80. https://doi.org/10.18280/ria.370520.

44. Alboaneen D, Alqarni R, Alqahtani S, Alrashidi M, Alhuda R, Alyahyan R, Alshammari T. Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. BDCC. 2023; 7:74. https://doi.org/10.3390/bdcc7020074.

45. Anuraga G, Tang WC, Phan NN, Ta HDK, Liu YH, Wu YF, Lee KH, Wang CY. Comprehensive Analysis of Prognostic and Genetic Signatures for General Transcription Factor

III (GTF3) in Clinical Colorectal Cancer Patients Using Bioinformatics Approaches. Curr Issues Mol Biol. 2021; 43:2–20. https://doi.org/10.3390/cimb43010002. [PubMed]

46. Gabriel E, Attwood K, Al-Sukhni E, Erwin D, Boland P, Nurkin S. Age-related rates of colorectal cancer and the factors associated with overall survival. J Gastrointest Oncol. 2018; 9:96–110. https://doi.org/10.21037/jgo.2017.11.13. [PubMed]

47. Yang Y, Wang G, He J, Ren S, Wu F, Zhang J, Wang F. Gender differences in colorectal cancer survival: A meta-analysis. Int J Cancer. 2017; 141:1942–49. https://doi.org/10.1002/ijc.30827. [PubMed]

48. de Back TR, Wu T, Schafrat PJ, Ten Hoorn S, Tan M, He L, van Hooff SR, Koster J, Nijman LE, Vink GR, Beumer IJ, Elbers CC, Lenos KJ, et al. A consensus molecular subtypes classification strategy for clinical colorectal cancer tissues. Life Sci Alliance. 2024; 7:e202402730. https://doi.org/10.26508/lsa.202402730. [PubMed]

49. Tziris N, Dokmetzioglou J, Giannoulis K, Kesisoglou I, Sapalidis K, Kotidis E, Gambros O. Synchronous and metachronous adenocarcinomas of the large intestine. Hippokratia. 2008; 12:150–52. [PubMed]

50. Zhang Y, Win AK, Makalic E, Buchanan DD, Pai RK, Phipps AI, Rosty C, Boussioutas A, Karahalios A, Jenkins MA. Associations between pathological features and risk of metachronous colorectal cancer. Int J Cancer. 2024; 155:1023–32. https://doi.org/10.1002/ijc.34979. [PubMed]

51. Zhang Y, Karahalios A, Aung YK, Win AK, Boussioutas A, Jenkins MA. Risk factors for metachronous colorectal cancer and advanced neoplasia following primary colorectal cancer: a systematic review and meta-analysis. BMC Gastroenterol. 2023; 23:421. https://doi.org/10.1186/s12876-023-03053-2. [PubMed]

52. Su Y, Tian X, Gao R, Guo W, Chen C, Chen C, Jia D, Li H, Lv X. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Comput Biol Med. 2022; 145:105409. https://doi.org/10.1016/j.compbiomed.2022.105409. [PubMed]

53. Kang J, Choi YJ, Kim IK, Lee HS, Kim H, Baik SH, Kim NK, Lee KY. LASSO-Based Machine Learning Algorithm for Prediction of Lymph Node Metastasis in T1 Colorectal Cancer. Cancer Res Treat. 2021; 53:773–83. https://doi.org/10.4143/crt.2020.974. [PubMed]

54. Yao H, Lu F, Shao Y. The E2F family as potential biomarkers and therapeutic targets in colon cancer. PeerJ. 2020; 8:e8562. https://doi.org/10.7717/peerj.8562. [PubMed]

55. Xu Z, Qu H, Ren Y, Gong Z, Ri HJ, Chen X. An Update on the Potential Roles of E2F Family Members in Colorectal Cancer. Cancer Manag Res. 2021; 13:5509–21. https://doi.org/10.2147/CMAR.S320193. [PubMed]

56. Yan PY, Zhang XA. Knockdown of E2F8 Suppresses Cell Proliferation in Colon Cancer Cells by Modulating the NF-κB Pathway. Ann Clin Lab Sci. 2019; 49:474–80. [PubMed]

57. Wang Y, Wu Q, Liu J, Wang X, Xie J, Fu X, Li Y. WDR77 in Pan-Cancer: Revealing expression patterns, genetic insights, and functional roles across diverse tumor types, with a spotlight on colorectal cancer. Transl Oncol. 2024; 49:102089. https://doi.org/10.1016/j.tranon.2024.102089. [PubMed]

58. Zhang JL, Zheng HF, Li K, Zhu YP. miR-495-3p depresses cell proliferation and migration by downregulating HMGB1 in colorectal cancer. World J Surg Oncol. 2022; 20:101. https://doi.org/10.1186/s12957-022-02500-w. [PubMed]

59. Sinicrope FA. Increasing Incidence of Early-Onset Colorectal Cancer. N Engl J Med. 2022; 386:1547–58. https://doi.org/10.1056/NEJMra2200869. [PubMed]

60. Ulanja MB, Ntafam C, Beutler BD, Antwi-Amoabeng D, Rahman GA, Ulanja RN, Mabrouk T, Governor SB, Djankpa FT, Alese OB. Race, age, and sex differences on the influence of obesity on colorectal cancer sidedness and mortality: A national cross-sectional study. J Surg Oncol. 2023; 127:109–18. https://doi.org/10.1002/jso.27096. [PubMed]

61. Kim HJ, Choi GS. Clinical Implications of Lymph Node Metastasis in Colorectal Cancer: Current Status and Future Perspectives. Ann Coloproctol. 2019; 35:109–17. https://doi.org/10.3393/ac.2019.06.12. [PubMed]

62. Grass F, Behm KT, Duchalais E, Crippa J, Spears GM, Harmsen WS, Hübner M, Mathis KL, Kelley SR, Pemberton JH, Dozois EJ, Larson DW. Impact of delay to surgery on survival in stage I-III colon cancer. Eur J Surg Oncol. 2020; 46:455–61. https://doi.org/10.1016/j.ejso.2019.11.513. [PubMed]

63. Chan GHJ, Chee CE. Making sense of adjuvant chemotherapy in colorectal cancer. J Gastrointest Oncol. 2019; 10:1183–92. https://doi.org/10.21037/jgo.2019.06.03. [PubMed]

64. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011; 12:2825–30. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

65. Vieira LM, Jorge NAN, de Sousa JB, Setubal JC, Stadler PF, Walter MEM. Competing Endogenous RNA in Colorectal Cancer: An Analysis for Colon, Rectum, and Rectosigmoid Junction. Front Oncol. 2021; 11:681579. https://doi.org/10.3389/fonc.2021.681579. [PubMed]

66. Li R, Qu H, Wang S, Wei J. GDCRNATools: An R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinform. 2018; 34:2515–17. https://doi.org/10.1093/bioinformatics/bty124.

67. Ritchie M, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47. https://doi.org/10.1093/nar/gkv007.

68. Furió-Tarí P, Tarazona S, Gabaldón T, Enright AJ, Conesa A. spongeScan: A web for detecting microRNA binding

elements in lncRNA sequences. Nucleic Acids Res. 2016; 44:W176–80. https://doi.org/10.1093/nar/gkw443. [PubMed]

69. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014; 42:D92–97. https://doi.org/10.1093/nar/gkt1248. [PubMed]

70. Bai R, Wu D, Shi Z, Hu W, Li J, Chen Y, Ge W, Yuan Y, Zheng S. Pan-cancer analyses demonstrate that ANKRD6 is associated with a poor prognosis and correlates with M2 macrophage infiltration in colon cancer. Chin J Cancer Res. 2021; 33:93–102. https://doi.org/10.21147/j.issn.1000-9604.2021.01.10. [PubMed]

71. Xu B, Lian J, Pang X, Gu Y, Zhu J, Zhang Y, Lu H. Identification of colon cancer subtypes based on multi-omics data-construction of methylation markers for immunotherapy. Front Oncol. 2024; 14:1335670. https://doi.org/10.3389/fonc.2024.1335670. [PubMed]

72. Liu C, Wu W, Chang W, Wu R, Sun X, Wu H, Liu Z. miR-31-5p-*DMD* axis as a novel biomarker for predicting the development and prognosis of sporadic early-onset colorectal cancer. Oncol Lett. 2022; 23:157. https://doi.org/10.3892/ol.2022.13277. [PubMed]

73. Ghafouri-Fard S, Esmaeili M, Taheri M. H19 lncRNA: Roles in tumorigenesis. Biomed Pharmacother. 2020; 123:109774. https://doi.org/10.1016/j.biopha.2019.109774. [PubMed]

74. Li F, Wang L, Wang Y, Shen H, Kou Q, Shen C, Xu X, Zhang Y, Zhang J. HECW2 promotes the progression and chemoresistance of colorectal cancer via AKT/mTOR signaling activation by mediating the ubiquitin-proteasome degradation of lamin B1. J Cancer. 2023; 14:2820–32. https://doi.org/10.7150/jca.87545. [PubMed]

75. Yin J, Guo Y. HOXD13 promotes the malignant progression of colon cancer by upregulating PTPRN2. Cancer Med. 2021; 10:5524–33. https://doi.org/10.1002/cam4.4078. [PubMed]

76. Zheng ZH, You HY, Feng YJ, Zhang ZT. LncRNA KCNQ1OT1 is a key factor in the reversal effect of curcumin on cisplatin resistance in the colorectal cancer cells. Mol Cell Biochem. 2021; 476:2575–85. https://doi.org/10.1007/s11010-020-03856-x. [PubMed]

77. Chen B, Liu D, Chen R, Guo L, Ran J. Elevated LINC00894 relieves the oncogenic properties of thyroid cancer cell by sponging let-7e-5p to promote TIA-1 expression. Discov Oncol. 2022; 13:56. https://doi.org/10.1007/s12672-022-00520-2. [PubMed]

78. Jonna S, Feldman RA, Swensen J, Gatalica Z, Korn WM, Borghaei H, Ma PC, Nieva JJ, Spira AI, Vanderwalde AM, Wozniak AJ, Kim ES, Liu SV. Detection of NRG1 Gene Fusions in Solid Tumors. Clin Cancer Res. 2019; 25:4966–72. https://doi.org/10.1158/1078-0432.CCR-19-0160. [PubMed]

79. Zheng D, Cao M, Zuo S, Xia X, Zhi C, Lin Y, Deng S, Yuan X. RANBP1 promotes colorectal cancer progression by regulating pre-miRNA nuclear export via a positive feedback loop with YAP. Oncogene. 2022; 41:930–42. https://doi.org/10.1038/s41388-021-02036-5. [PubMed]

80. Ke D, Wang Q, Ke S, Zou L, Wang Q. Long-Non Coding RNA SNHG16 Supports Colon Cancer Cell Growth by Modulating miR-302a-3p/AKT Axis. Pathol Oncol Res. 2020; 26:1605–13. https://doi.org/10.1007/s12253-019-00743-9. [PubMed]

81. Zhang XW, Li SL, Zhang D, Sun XL, Zhai HJ. RP11-619L19.2 promotes colon cancer development by regulating the miR-1271-5p/CD164 axis. Oncol Rep. 2020; 44:2419–28. https://doi.org/10.3892/or.2020.7794. [PubMed]

82. Song GL, Xiao M, Wan XY, Deng J, Ling JD, Tian YG, Li M, Yin J, Zheng RY, Tang Y, Liu GY. MiR-130a-3p suppresses colorectal cancer growth by targeting Wnt Family Member 1 (WNT1). Bioengineered. 2021; 12:8407–18. https://doi.org/10.1080/21655979.2021.1977556. [PubMed]

83. Song D, Zhang Q, Zhang H, Zhan L, Sun X. MiR-130b-3p promotes colorectal cancer progression by targeting CHD9. Cell Cycle. 2022; 21:585–601. https://doi.org/10.1080/15384101.2022.2029240. [PubMed]