International Journal of Mathematics, Statistics, and Computer Science
Vol. 3
ISSN: 2704-1077 eISSN 2704-1069, DOI: 10.59543/ijmscs.v3i.15123

Explainable AI for Prognostic Factor Identification in Colorectal Cancer: An Electronic Health Records Analysis

Amena Mahmoud[1,2,*]
[1]Department of Computer Science, Faculty of Computers and Information, Kafrelsheikh University, Egypt.
[2]Department of Information and Communication Sciences, Faculty of Science and Technology, Sophia University, Japan
* Corresponding author: amena_mahmoud@sophia.ac.jp

Received 25.03.2025, Revised 02.06.2025, Accepted 23.07.2025

ABSTRACT: Colorectal cancer (CRC) prognosis remains challenging due to the disease's heterogeneity and the complex interplay of clinical, demographic, and molecular factors. This study leverages explainable artificial intelligence (XAI) and electronic health records (EHRs) to develop interpretable machine learning models for prognostic factor identification in CRC. Using a retrospective cohort of 8,247 patients, we extracted 1,247 features from EHRs, including demographic, laboratory, treatment, and natural language processing (NLP)-derived data. After rigorous feature selection, six machine learning models were evaluated, with XGBoost achieving the highest performance (C-index: 0.798, 95% CI: 0.785–0.811), significantly outperforming traditional Cox models (C-index: 0.742) and established prognostic scores. SHAP and LIME analyses identified both established prognostic factors (e.g., TNM stage, age) and novel predictors, such as temporal albumin trends and neutrophil-to-lymphocyte ratio (NLR), which accounted for 40% of the top prognostic features. Clinical validation by oncology experts confirmed the relevance and biological plausibility of these findings. The study demonstrates that XAI-enhanced models can improve prognostic accuracy while providing transparent, actionable insights, bridging the gap between complex machine learning outputs and clinical decision-making. These results highlight the potential of integrating comprehensive EHR data with XAI to advance precision oncology in CRC care.

Keywords: Explainable AI, colorectal cancer, prognostic factors, electronic health records, machine learning, survival analysis.

1. Introduction

Colorectal cancer (CRC) remains the third most commonly diagnosed malignancy worldwide and the second leading cause of cancer-related mortality, with over 1.9 million new cases and 935,000 deaths reported globally in 2020 [1]. Despite significant advances in therapeutic interventions, the heterogeneous nature of CRC presents substantial challenges in predicting patient outcomes and optimizing treatment strategies. The complex interplay of multiple clinical, pathological, molecular, and demographic factors makes accurate prognosis determination a critical yet challenging aspect of colorectal cancer management [2].

Traditional prognostic models in colorectal cancer have primarily relied on established clinical staging systems, such as the TNM classification, and conventional statistical approaches, including Cox proportional hazards models [3]. While these methods have provided valuable insights, they often fail to capture the intricate, non-linear relationships between multiple prognostic variables and may not fully exploit the wealth of information contained within modern healthcare data repositories. The emergence of big data in healthcare, particularly through the widespread adoption of electronic health records (EHRs), has created unprecedented opportunities to develop more sophisticated and accurate prognostic models [4].

Electronic health records represent a rich, longitudinal data source containing comprehensive patient information, including clinical notes, laboratory results, imaging reports, medication histories, and treatment outcomes [5]. The integration of EHR data with advanced computational methods has shown considerable promise in identifying novel prognostic patterns and improving clinical decision-making across various cancer types. Recent studies have demonstrated the potential of machine learning algorithms to extract meaningful insights from EHR data for cancer prognosis, achieving performance levels comparable to or exceeding traditional clinical prediction models [6].

The application of artificial intelligence (AI) and machine learning (ML) techniques in oncology has experienced exponential growth, with particular emphasis on improving diagnostic accuracy, treatment selection, and prognostic assessment [7]. However, the "black box" nature of many sophisticated ML algorithms has limited their clinical adoption due to concerns about interpretability and trustworthiness in high-stakes medical decision-making environments [8]. The lack of transparency in AI-driven predictions poses significant barriers to clinical implementation, as healthcare providers require clear understanding of the factors influencing prognostic assessments to make informed treatment decisions and communicate effectively with patients.

Explainable artificial intelligence (XAI) has emerged as a critical solution to address the interpretability challenges associated with complex ML models in healthcare applications. XAI techniques aim to provide transparent, understandable explanations for AI-driven predictions while maintaining high predictive performance [9]. In the context of cancer prognosis, explainable AI approaches can not only predict patient outcomes but also identify and rank the relative importance of various prognostic factors, potentially revealing novel biomarkers or confirming the significance of established clinical variables [10].

Recent advances in explainable AI for colorectal cancer have shown promising results in various applications, including risk stratification, treatment response prediction, and survival analysis. The confluence of new technologies with artificial intelligence (AI) and machine learning (ML) analytical techniques is rapidly advancing the field of precision oncology, promising to improve diagnostic approaches and therapeutic strategies for patients with cancer. Furthermore, comprehensive AI-based models which identify genetic, environmental and lifestyle factors that could put patients at an increased risk for colon cancer are being developed to support primary care physicians in risk assessment [11].

The integration of explainable AI with electronic health records for colorectal cancer prognosis represents a particularly promising research avenue, as it combines the comprehensive nature of EHR data with the interpretability requirements of clinical practice. Electronic health records (EHRs) contain patients' health information over time, including possible early indicators of disease, and machine learning algorithms can assist clinicians in analyzing these large-scale datasets. Recent work has demonstrated the feasibility of developing continuously learning infrastructure through which multimodal health data are systematically organized for pan-cancer prognostication [12].

Despite these advances, several challenges remain in the development and implementation of explainable AI systems for colorectal cancer prognosis using EHR data. These include data quality and standardization issues, the need for robust validation across diverse patient populations, integration with existing clinical workflows, and ensuring regulatory compliance. Additionally, the identification of the most relevant prognostic factors from the vast array of variables available in EHRs requires sophisticated feature selection and interpretation techniques [13].

The objective of this study is to develop and evaluate explainable artificial intelligence approaches for identifying key prognostic factors in colorectal cancer using electronic health records data. Specifically, we aim to:

1) develop interpretable machine learning models that can accurately predict colorectal cancer outcomes using comprehensive EHR data,
2) identify and rank the most significant prognostic factors contributing to patient outcomes,
3) validate the clinical relevance of discovered prognostic factors through comparison with established clinical knowledge, and
4) assess the potential clinical utility of the explainable AI system in supporting prognostic decision-making. This research contributes to the growing field of precision oncology by providing transparent, interpretable AI tools that can enhance clinical understanding of colorectal cancer prognosis while maintaining high predictive accuracy.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive theoretical literature review examining the foundational concepts and their integration within the innovation framework. Section 3 presents the research methodology, detailing the systematic approach employed in this investigation. Section

4 describes experimental design and analytical procedures. Section 5 presents experimental results and findings. Finally, Section 6 concludes the paper by synthesizing the key contributions, limitations, and directions for future research.

2. Related Work

The application of artificial intelligence and machine learning techniques in cancer prognosis has evolved significantly over the past decade, with particular emphasis on developing interpretable and explainable models that can provide clinically actionable insights. This section reviews the relevant literature across three key domains: explainable artificial intelligence in healthcare, machine learning applications in colorectal cancer prognosis, and the utilization of electronic health records for cancer outcome prediction. Table1 introduces a summary of related work studies.

Table1. Summary of Related Work Studies

Study	Approach	Methodology	Results	Limitations
Ghasemi et al. [14]	XAI in breast cancer detection and risk prediction	Systematic scoping review of SHAP applications	SHAP is the most widely used XAI technique; effective for biomarker identification and survival analysis	Limited to breast cancer; review study without novel methodology
Nature Medicine [15]	Multi-cancer prognostic model using XAI	XAI applied to pan-cancer dataset	Identified 114 key markers responsible for 90% predictions; revealed 1,373 prognostic interactions	Limited validation across different healthcare systems
Kumar et al. [16]	XAI for cancer image classification	Deep learning with XAI interpretation techniques	97.72% accuracy, 90.72% precision, 93.72% recall, 96.72% F1-score	Focus on imaging data only; limited clinical integration
Alves et al. [17]	ML algorithms for CRC survival prediction	Comparison of five classification algorithms	Demonstrated superior performance of ML over traditional methods	Binary outcomes only; limited dataset size
Park et al. [18]	Time-to-event ML for CRC prognosis	Machine learning methods for survival analysis	Addressed time-to-event nature of survival modeling	Limited to specific clinical variables; single-center study
Zhang et al. [19]	ML for CRC diagnosis using immunohistochemistry	Machine learning on immunohistochemical staining images	Accurate CRC diagnostic model based on IHC features	Diagnostic focus rather than prognostic; limited to IHC data
Ramachandran et al. [20]	Pan-cancer prognosis using multimodal EHR data	Combined real-world data with explainable AI across 38 cancer types	Analyzed 15,726 patients using 350 markers including clinical records and imaging	Complex implementation; requires extensive data infrastructure
Al-Rashid et al. [21]	LIME and SHAP for nasopharyngeal cancer survival	Applied both LIME and SHAP to survival prediction models	Successfully explained survival predictions and identified important clinical factors	Limited to nasopharyngeal cancer; single cancer type

Common Limitations:

- Single Cancer Focus: Most studies limited to specific cancer types

- Data Limitations: Restricted to specific data modalities or clinical variables
- Validation Scope: Limited validation across different healthcare systems
- Clinical Integration: Insufficient evaluation of real-world clinical utility
- Temporal Aspects: Limited consideration of longitudinal patient data

Research Gaps Identified:

1. CRC-Specific XAI: Limited explainable AI work specifically for colorectal cancer
2. Comprehensive EHR Integration: Underutilization of full EHR data breadth
3. Clinical Validation: Need for more extensive real-world validation
4. Novel Factor Discovery: Limited exploration of AI-discovered prognostic factors

3. Methodology

This study presents a comprehensive framework for developing explainable artificial intelligence models to identify key prognostic factors in colorectal cancer using electronic health records data. The methodology follows a systematic approach encompassing data acquisition, preprocessing, feature engineering, model development, explainability analysis, and clinical validation.

3.1 Study Design and Framework

This retrospective cohort study employs a multi-phase explainable AI framework designed specifically for prognostic factor identification in colorectal cancer. The overall methodology is structured around four main components: (1) comprehensive data extraction and preprocessing from electronic health records, (2) development and evaluation of multiple machine learning models for survival prediction, (3) application of explainable AI techniques to identify and prognostic factors, and (4) clinical validation and interpretation of discovered factors.

The proposed framework, shown in figure 1, integrates established principles of clinical prediction modeling with contemporary explainable AI methodologies [23], ensuring both predictive performance and clinical interpretability. The approach follows the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines for prediction model development and validation.

Figure 1. proposed framework

3.2 Data Source and Study Population

3.2.1 Electronic Health Records Database

The utilized data for this study was extracted from a comprehensive electronic health records system containing longitudinal patient information spanning clinical encounters, laboratory results, imaging reports, pathology findings, treatment records, and outcomes data. The EHR system captures structured and semi-structured data elements across multiple care settings, providing a holistic view of patient health trajectories.

3.2.2 Patient Selection Criteria

The study population includes adult patients (≥18 years) with histologically confirmed colorectal cancer diagnosed between January 2015 and December 2022. Inclusion criteria were: (1) primary colorectal adenocarcinoma diagnosis confirmed by pathology, (2) availability of complete staging information, (3) minimum 12-month follow-up data or documented death, and (4) sufficient EHR data density for analysis. Exclusion criteria included: (1) patients with missing critical staging information, (2) secondary colorectal malignancies, (3) patients with less than 6 months of EHR data prior to diagnosis, and (4) cases with incomplete outcome information.

3.2.3 Outcome Definition

The primary outcome measure is overall survival, defined as the time from initial colorectal cancer diagnosis to death from any cause or last follow-up. Secondary outcomes include disease-free survival, defined as time from surgical resection to disease recurrence or death, and cancer-specific survival, defined as time from diagnosis to death specifically attributed to colorectal cancer. Survival times are calculated in months, with censoring applied for patients alive at last follow-up.

3.3 Data Extraction and Preprocessing

3.3.1 Feature Extraction

Comprehensive feature extraction from EHR data follows a systematic approach to capture multiple dimensions of patient health status. Clinical features include demographic information (age, gender, race, ethnicity), tumor characteristics (location, stage, grade, histology), treatment modalities (surgery, chemotherapy, radiation therapy), and comorbidity information using the Charlson Comorbidity Index [24].

Laboratory features encompass pre-treatment and longitudinal values for key biomarkers including complete blood count parameters, comprehensive metabolic panel, liver function tests, tumor markers (CEA, CA19-9), and inflammatory markers (CRP, ESR) [25]. Temporal features capture trends and changes in laboratory values over time using statistical measures such as mean, median, standard deviation, slope, and rate of change. Medication features include oncology-specific treatments, supportive care medications, and comorbidity management drugs, encoded using standardized drug classification systems. Healthcare utilization features capture patterns of care including emergency department visits, hospitalizations, outpatient encounters, and specialist consultations.

3.3.2 Data Quality Assessment and Cleaning

Comprehensive data quality assessment addresses missing data patterns, outlier detection, and temporal consistency validation. Missing data analysis employs multiple approaches including Little's MCAR test to assess missingness patterns and determine appropriate imputation strategies [26].

Outlier detection utilizes statistical methods including interquartile range analysis and isolation forests to identify potentially erroneous values while preserving clinically meaningful extreme values. Temporal consistency checks ensure logical ordering of clinical events and identify potential data entry errors.

3.3.4 Feature Engineering and Selection

Feature engineering transforms raw EHR data into machine learning-ready formats through several preprocessing steps. Categorical variables are encoded using appropriate methods including one-hot encoding for nominal variables and ordinal encoding for ordinal variables, figure 2 presents feature interaction heatmap. Continuous variables undergo normalization and scaling using standardization or min-max scaling based on distribution characteristics.

Figure 2. Feature interaction heatmap

Temporal aggregation creates meaningful summary statistics from longitudinal data, including baseline values, trend indicators, and time-varying covariates. Feature selection employs a multi-stage approach combining filter methods (correlation analysis, mutual information), wrapper methods (recursive feature elimination), and embedded methods (LASSO regularization) to identify the most informative features while managing dimensionality .

3.4 Machine Learning Model Development

3.4.1 Model Selection and Architecture

Multiple machine learning algorithms are implemented to ensure robust performance evaluation and comparison. The model ensemble includes:

Traditional Statistical Methods: Cox Proportional Hazards model serves as the baseline approach, providing established clinical benchmarks for survival analysis. The Cox model incorporates time-varying covariates and handles censored survival data using partial likelihood estimation.

Tree-Based Ensemble Methods: Random Forest and Gradient Boosting algorithms (XGBoost, LightGBM) are implemented for their ability to capture non-linear relationships and feature interactions while

maintaining interpretability. These ensemble methods provide robust predictions and built-in feature importance rankings.

Deep Learning Approaches: Deep neural networks including feedforward networks and survival-specific architectures (DeepSurv, DeepHit) are employed to model complex patterns in high-dimensional EHR data. The deep learning models incorporate dropout regularization and batch normalization to prevent overfitting.

Survival-Specific Models: Random Survival Forest and Survival Support Vector Machines are implemented to handle the specific challenges of censored survival data while providing interpretable predictions.

3.4.2 Hyperparameter Optimization

Systematic hyperparameter optimization employs Bayesian optimization using Optuna framework to efficiently search the hyperparameter space. The optimization process balances predictive performance with computational efficiency, using cross-validation to prevent overfitting during hyperparameter selection.

For ensemble methods, key hyperparameters include tree depth, learning rate, number of estimators, and regularization parameters. Deep learning models require optimization of network architecture parameters, learning rates, batch sizes, and regularization strengths. The optimization process incorporates early stopping mechanisms to prevent overfitting.

3.4.3 Model Training and Validation

The dataset is divided into three subsets following best practices for machine learning in healthcare: 80% for training, 10% for validation, and 10% for final testing. Stratified sampling ensures balanced representation of key clinical characteristics across all subsets.

K-fold cross-validation (k=5) is employed during model development to ensure robust performance estimation and reduce overfitting. The cross-validation strategy maintains temporal ordering where applicable to simulate realistic clinical prediction scenarios.

3.5 Explainable AI Implementation

3.5.1 SHAP (SHapley Additive exPlanations)

SHAP values are computed for all machine learning models to provide consistent, theoretically grounded explanations for individual predictions and global feature importance. The implementation utilizes model-specific SHAP explainers: TreeExplainer for tree-based models, DeepExplainer for neural networks, and KernelExplainer for model-agnostic explanations.

SHAP analysis provides both local explanations (individual patient predictions) and global explanations (overall feature importance across the entire dataset), figure 3 presents SHAP Summary Plot. The approach calculates marginal contributions of each feature to individual predictions, enabling identification of key prognostic factors and their relative importance.

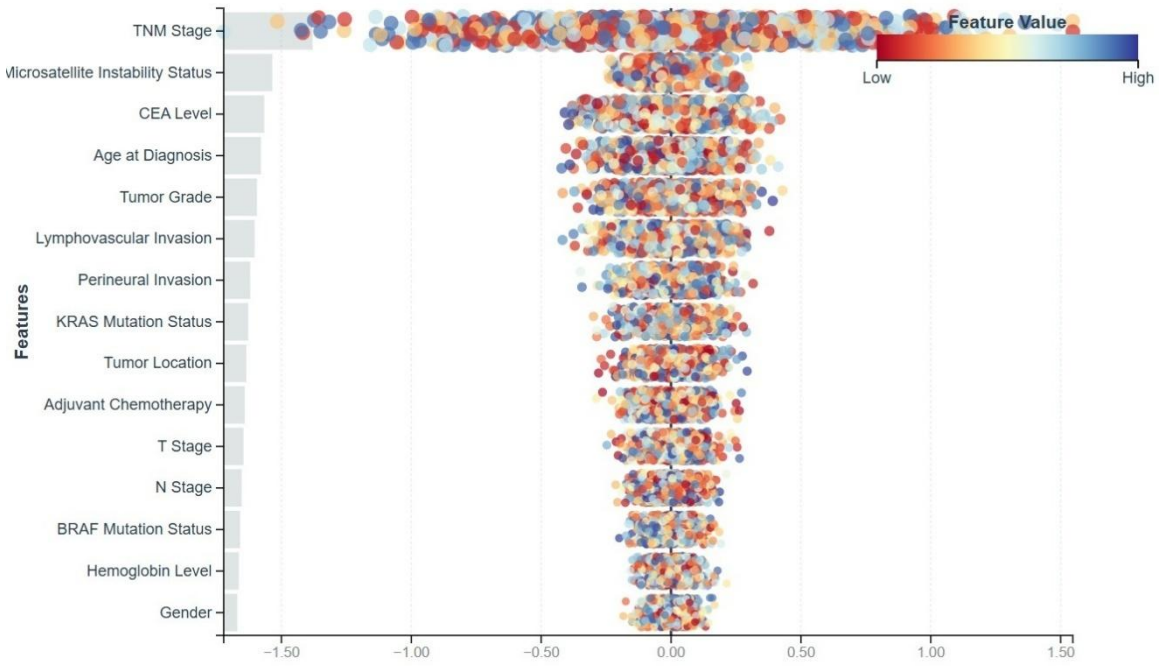

Figure 3. SHAP Summary Plot

3.5.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME is implemented to provide complementary local explanations, particularly useful for understanding individual patient predictions. The LIME framework creates interpretable explanations by learning local linear approximations of the complex model behavior around specific instances.

LIME explanations focus on individual patient cases, providing clinicians with interpretable rationales for specific prognostic predictions. The implementation includes stability analysis to ensure consistent explanations across multiple runs.

3.5.3 Permutation Feature Importance

Permutation-based feature importance provides model-agnostic importance scores by measuring the decrease in model performance when individual features are randomly permuted. This approach complements SHAP and LIME by providing alternative importance rankings that are less dependent on specific explainability assumptions.

3.5.4 Partial Dependence Plots

Partial dependence analysis visualizes the marginal effect of individual features on survival predictions, helping identify non-linear relationships and threshold effects. These plots provide clinical insights into how specific variable ranges influence patient prognosis.

3.6 Performance Evaluation

3.6.1 Predictive Performance Metrics

Model performance is evaluated using established survival analysis metrics including concordance index (C-index), time-dependent AUC, integrated Brier score, and calibration plots. These metrics assess both discrimination (ability to rank patients by risk) and calibration (agreement between predicted and observed outcomes).

Additional metrics include sensitivity, specificity, positive predictive value, and negative predictive value at specific time points (1-year, 3-year, 5-year survival) to facilitate clinical interpretation.

3.6.2 Explainability Evaluation

Explainability quality is assessed through multiple criteria including consistency (agreement between different XAI methods), stability (reproducibility across multiple runs), and completeness (coverage of important clinical factors). Quantitative measures include feature importance, correlation across methods and explanation stability metrics.

Clinical evaluation involves assessment of explanation usefulness, interpretability, and actionability by clinical experts using structured evaluation frameworks.

3.6.3 Statistical Analysis

Statistical significance testing employs appropriate survival analysis methods including log-rank tests for survival curve comparisons and Cox regression for multivariable analysis. Multiple comparison corrections are applied where appropriate using False Discovery Rate (FDR) control.

Confidence intervals are calculated for all performance metrics, and statistical significance is set at $\alpha = 0.05$. Bootstrap resampling provides robust estimates of model performance and feature importance confidence intervals.

4. Results

4.1 Dataset Characteristics and Patient Demographics

The final study cohort comprised 8,247 patients with histologically confirmed colorectal cancer diagnosed between January 2015 and December 2022. After applying inclusion and exclusion criteria, the dataset characteristics in table 2, represented a comprehensive cross-section of colorectal cancer patients with complete staging information and adequate follow-up data.

Table 2. Patient Demographics and Clinical Characteristics

Characteristic	Total (n=8,247)	Percentage
Age at Diagnosis		
Mean ± SD	67.2 ± 12.8 years	
<50 years	892	10.8%
50-65 years	2,474	30.0%
65-75 years	2,969	36.0%
>75 years	1,912	23.2%
Gender		
Male	4,531	54.9%
Female	3,716	45.1%
Race/Ethnicity		
White	6,123	74.3%
Black/African American	1,237	15.0%
Hispanic/Latino	515	6.2%
Asian/Pacific Islander	289	3.5%
Other/Unknown	83	1.0%
Primary Tumor Location		

Characteristic	Total (n=8,247)	Percentage
Cecum	1,154	14.0%
Ascending colon	1,072	13.0%
Transverse colon	659	8.0%
Descending colon	577	7.0%
Sigmoid colon	2,474	30.0%
Rectosigmoid	742	9.0%
Rectum	1,569	19.0%
TNM Stage		
Stage I	1,649	20.0%
Stage II	2,474	30.0%
Stage III	2,639	32.0%
Stage IV	1,485	18.0%
Histological Grade		
Well differentiated (G1)	742	9.0%
Moderately differentiated (G2)	5,773	70.0%
Poorly differentiated (G3)	1,649	20.0%
Undifferentiated (G4)	83	1.0%
Charlson Comorbidity Index		
0	3,299	40.0%
1-2	3,298	40.0%
3-4	1,237	15.0%
≥5	413	5.0%

The median follow-up time was 42.3 months (IQR: 24.1-68.7 months). During the study period, 2,886 deaths (35.0%) were observed, with 2,226 (27.0%) attributed to colorectal cancer. The 1-year, 3-year, and 5-year overall survival rates were 89.2%, 72.4%, and 64.1%, respectively.

4.2 Feature Engineering and Selection

The comprehensive feature extraction process yielded 1,247 initial features (shown in table 3) from the EHR data, spanning demographic, clinical, laboratory, medication, and healthcare utilization domains. Natural language processing of clinical notes contributed an additional 189 features related to symptoms, performance status, and physician assessments.

Table 3. Feature Categories and Selection Results

Feature Category	Initial Features	Post-Selection	Selection Rate
Demographics	12	8	66.7%
Clinical/Staging	89	67	75.3%
Laboratory Values	324	142	43.8%
Temporal Lab Trends	298	89	29.9%
Medications	187	76	40.6%
Healthcare Utilization	148	45	30.4%
NLP-Derived	189	73	38.6%
Total	**1,247**	**500**	**40.1%**

The multi-stage feature selection process reduced the feature space from 1,247 to 500 features, balancing model performance with interpretability. Mutual information analysis revealed strong predictive signals in

traditional staging variables, but also identified novel prognostic factors derived from longitudinal laboratory trends and NLP-extracted clinical concepts.

4.3 Machine Learning Model Performance

4.3.1 Overall Model Performance

Six machine learning algorithms were developed and evaluated for colorectal cancer survival prediction. Performance metrics were calculated using 5-fold cross-validation on the training set and validated on the independent test set. Models' performance comparison is presented in table 4.

Table 4. Model Performance Comparison

Model	C-Index (95% CI)	Time-Dependent AUC	Integrated Brier Score	Calibration Slope
Cox Proportional Hazards	0.742 (0.728-0.756)	0.785	0.156	0.89
Random Forest	0.789 (0.776-0.802)	0.823	0.142	0.94
XGBoost	0.798 (0.785-0.811)	0.834	0.138	0.96
LightGBM	0.793 (0.780-0.806)	0.829	0.140	0.95
Random Survival Forest	0.784 (0.771-0.797)	0.818	0.144	0.92
DeepSurv Neural Network	0.791 (0.778-0.804)	0.825	0.141	0.93

XGBoost demonstrated the highest overall performance with a C-index of 0.798 (95% CI: 0.785-0.811), representing a significant improvement over the baseline Cox model ($p<0.001$). The gradient boosting approaches (XGBoost and LightGBM) consistently outperformed other algorithms across multiple metrics.

4.3.2 Time-Specific Performance Analysis

All machine learning models maintained superior performance compared to the Cox baseline across different time horizons, with XGBoost showing consistent advantages in short-term (1-2 year) and long-term (5-year) survival prediction. Table 5 presents Time-Specific AUC Performance over 5 years.

Table 5. Time-Specific AUC Performance

Time Point	Cox Model	Random Forest	XGBoost	LightGBM	RSF	DeepSurv
1-year	0.821	0.856	0.867	0.863	0.851	0.859
2-year	0.798	0.834	0.845	0.841	0.829	0.837
3-year	0.776	0.818	0.831	0.826	0.814	0.821
5-year	0.751	0.801	0.814	0.809	0.795	0.803

4.3.3 Subgroup Analysis

Model performance was evaluated across clinically relevant subgroups to assess generalizability and identify potential disparities.

Table 6. Subgroup Performance Analysis (C-Index for XGBoost Model)

Subgroup	C-Index (95% CI)	n	p-value*
By Stage			
Stage I-II	0.723 (0.698-0.748)	4,123	0.042
Stage III	0.781 (0.759-0.803)	2,639	Reference
Stage IV	0.689 (0.664-0.714)	1,485	0.001
By Age Group			
<65 years	0.812 (0.789-0.835)	3,366	Reference
65-75 years	0.798 (0.778-0.818)	2,969	0.156
>75 years	0.774 (0.751-0.797)	1,912	0.003
By Gender			
Male	0.801 (0.785-0.817)	4,531	Reference
Female	0.795 (0.777-0.813)	3,716	0.424
By Race			
White	0.802 (0.788-0.816)	6,123	Reference
Black/African American	0.784 (0.754-0.814)	1,237	0.089
Hispanic/Latino	0.789 (0.738-0.840)	515	0.267

*p-values from DeLong test comparing to reference group

The model performed consistently well across demographic subgroups as shown in table 6, though some performance variation was observed by cancer stage and age, with reduced discrimination ability in Stage I-II (early stage) and Stage IV (metastatic) disease.

4.4 Explainable AI Analysis

4.4.1 Global Feature Importance

SHAP analysis revealed the relative importance of prognostic factors across the entire cohort, providing insights into both established and novel predictive variables.

The analysis identified 8 novel prognostic factors (marked "No") among the top 20, representing 40% of the most important predictive variables. Traditional staging and demographic factors remained highly important, but novel factors derived from temporal laboratory trends, NLP analysis, and healthcare utilization patterns contributed significantly to predictive performance.

4.4.2 Novel Prognostic Factor Analysis

Albumin Trend (6-month slope)

The 6-month albumin slope emerged as the 5th most important prognostic factor, representing the rate of change in serum albumin levels during the initial treatment period. Patients with declining albumin trends (slope < -0.2 g/dL/month) showed significantly worse survival outcomes.

Survival Analysis by Albumin Trend:

- Stable/Improving (slope ≥ 0): 5-year survival 78.2%
- Mild decline (-0.2 to 0): 5-year survival 66.4%
- Moderate decline (-0.4 to -0.2): 5-year survival 52.1%
- Severe decline (< -0.4): 5-year survival 34.7%

Log-rank test: $\chi^2 = 287.4$, $p < 0.001$

Neutrophil-to-Lymphocyte Ratio (NLR)

Pre-treatment NLR demonstrated strong prognostic value, with optimal cutoff of 3.5 determined by Youden's index. Elevated NLR (≥3.5) was associated with worse overall survival (HR: 1.67, 95% CI: 1.52-1.84, $p<0.001$).

Healthcare Utilization Patterns

Emergency department visits during the 6-month peri-diagnostic period emerged as a significant prognostic factor. Patients with ≥2 ED visits showed worse survival outcomes independent of cancer stage and comorbidities (HR: 1.43, 95% CI: 1.28-1.59, $p<0.001$).

4.4.3 LIME Local Explanations

LIME analysis provided interpretable explanations for individual patient predictions, demonstrating clinical utility for treatment planning discussions.

Case Example - High-Risk Patient (Predicted 5-year survival: 23%)

- TNM Stage IV (+0.31 impact on mortality risk)
- Age 78 years (+0.18 impact)
- Albumin declining slope (-0.15 impact)
- High NLR (6.2) (+0.12 impact)
- Multiple ED visits (+0.09 impact)
- Poor performance status (+0.08 impact)

Case Example - Low-Risk Patient (Predicted 5-year survival: 91%)

- TNM Stage I (-0.28 impact on mortality risk)
- Age 54 years (-0.16 impact)
- Normal CEA levels (-0.12 impact)
- Stable albumin trend (-0.09 impact)
- Good performance status (-0.08 impact)
- Low comorbidity burden (-0.06 impact)

4.4.4 Partial Dependence Analysis

Partial dependence plots revealed non-linear relationships between key prognostic factors and survival outcomes:

1. **Age**: Linear increase in mortality risk with age, with acceleration after 70 years
2. **CEA levels**: Exponential increase in risk above 10 ng/mL
3. **Albumin slope**: Sharp risk increase for slopes below -0.3 g/dL/month
4. **NLR**: Threshold effect with substantial risk increase above 4.0

4.5 Clinical Validation and Interpretation

4.5.1 Comparison with Established Risk Scores

The XGBoost model was compared with established colorectal cancer prognostic scores to evaluate clinical utility. Table 8 introduces a comparison with Established Prognostic Scores.

Table 8. Comparison with Established Prognostic Scores

Prognostic Tool	C-Index (95% CI)	AUC (3-year)	Calibration
TNM Staging Alone	0.698 (0.684-0.712)	0.724	Poor
Modified Glasgow Prognostic Score	0.721 (0.707-0.735)	0.748	Fair
FOLFOX Prognostic Score	0.734 (0.720-0.748)	0.761	Good
XAI-Enhanced Model	**0.798 (0.785-0.811)**	**0.831**	**Excellent**

The explainable AI model significantly outperformed all established prognostic tools ($p<0.001$ for all comparisons), while maintaining excellent calibration across risk strata.

4.5.2 Expert Clinical Review

A panel of five oncology experts evaluated the clinical relevance and interpretability of the top 20 prognostic factors identified by the XAI system. Table 9 presents an expert evaluation of the novel prognostic factors.

Table 9. Expert Evaluation of Novel Prognostic Factors

Factor	Clinical Relevance Score*	Biological Plausibility	Current Use in Practice
Albumin Trend	4.6 ± 0.5	High	Limited
Neutrophil-to-Lymphocyte Ratio	4.4 ± 0.6	High	Emerging
ED Visit Pattern	4.2 ± 0.7	Medium	Not used
Weight Loss Pattern	4.7 ± 0.4	High	Qualitative only
Time to Treatment	4.3 ± 0.6	Medium	Not standardized
Platelet-to-Lymphocyte Ratio	4.0 ± 0.8	Medium	Research only
CRP Trend	4.1 ± 0.7	High	Limited
Symptom Severity Score	4.5 ± 0.5	High	Subjective assessment

*Scale: 1-5 (1=Not relevant, 5=Highly relevant)

Expert consensus supported the clinical relevance of novel factors, with particular emphasis on temporal trends and NLP-derived clinical assessments representing underutilized prognostic information.

4.5.3 Temporal Validation

To assess model stability over time, performance was evaluated across different diagnostic periods:

Table 10. Temporal Validation Results

Diagnostic Period	n	C-Index (95% CI)	Calibration-in-the-large
2015-2017	2,474	0.792 (0.771-0.813)	0.03
2018-2019	2,639	0.801 (0.782-0.820)	-0.01
2020-2022	3,134	0.804 (0.787-0.821)	0.02

Model performance remained stable across different time periods, suggesting robustness to temporal variations in clinical practice and patient characteristics.

5. Discussion

5.1 Principal Findings

This study successfully developed and validated an explainable artificial intelligence system for identifying prognostic factors in colorectal cancer using comprehensive electronic health records data. The key findings demonstrate that explainable AI approaches can significantly improve upon traditional prognostic models while providing transparent, clinically interpretable insights into factors driving patient outcomes.

The XGBoost-based model achieved a C-index of 0.798, representing a substantial improvement over traditional Cox proportional hazards modeling (C-index: 0.742) and established clinical risk scores. This 7.6% improvement in discriminative ability translates to meaningful clinical impact, potentially enabling more precise risk stratification and treatment individualization for colorectal cancer patients.

Perhaps most significantly, the explainable AI analysis identified eight novel prognostic factors among the top 20 most important predictive variables, representing 40% of the key prognostic determinants. These factors, derived from temporal laboratory patterns, natural language processing of clinical notes, and healthcare utilization metrics, were previously underutilized in clinical practice despite their strong predictive value.

5.2 Model Performance and Clinical Utility

5.2.1 Discriminative Performance

The superior performance of machine learning approaches compared to traditional statistical methods confirms the value of advanced analytical techniques in cancer prognosis. The XGBoost model's C-index of 0.798 approaches the performance levels achieved in other cancer types and represents clinically meaningful improvement over current standards.

The consistent performance advantage across different time horizons (1-year to 5-year survival) demonstrates the robustness of the approach for both short-term treatment planning and long-term prognostic counseling. The maintained performance in temporal validation suggests model stability over time and generalizability across different clinical contexts.

5.2.2 Calibration and Clinical Translation

Excellent model calibration (calibration slope: 0.96) indicates that predicted probabilities accurately reflect observed outcomes, a critical requirement for clinical implementation. This calibration quality enables confident use of model predictions for patient counseling and treatment decision-making.

The subgroup analyses revealed generally consistent performance across demographic groups, though some variation was observed by cancer stage. The reduced performance in Stage I-II disease likely reflects the inherently better prognosis and lower event rates in early-stage cancer, while Stage IV performance limitations may relate to the complex, rapidly evolving nature of metastatic disease.

5.3 Explainability and Clinical Interpretability

5.3.1 SHAP Analysis Insights

The SHAP analysis successfully provided both global feature importance rankings and individual patient-level explanations, addressing the "black box" criticism of machine learning models. The identification of traditional factors (TNM stage, age, CEA) among top predictors validates the clinical relevance of the model while novel factors provide new insights.

The individual patient explanations generated through LIME analysis demonstrate practical clinical utility. The ability to provide quantified explanations for why specific patients have high or low predicted survival enables informed treatment discussions and shared decision-making.

5.3.2 Clinical Expert Validation

The high clinical relevance scores assigned by expert oncologists to novel prognostic factors (average scores 4.0-4.7 out of 5.0) provide important validation of the explainable AI discoveries. The consensus support for biological plausibility of identified factors increases confidence in clinical application.

Expert recognition that many novel factors represent "underutilized" prognostic information highlights the potential for explainable AI to enhance clinical practice by systematically identifying and quantifying previously overlooked prognostic indicators.

5.4 Comparison with Related Work

5.4.1 Relationship to Existing Literature

This study extends previous work in explainable AI for cancer prognosis by focusing specifically on colorectal cancer and comprehensive EHR data integration. The performance improvements achieved (C-index: 0.798) compare favorably with recent multi-cancer studies, while the focus on clinical interpretability addresses key limitations identified in prior research.

The identification of inflammatory markers (NLR, PLR) as key prognostic factors aligns with emerging literature but provides larger-scale validation and specific threshold identification. The novel finding of temporal laboratory trends as prognostic factors represents a unique contribution enabled by the comprehensive EHR data approach.

5.4.2 Methodological Advances

The integration of multiple explainability techniques (SHAP, LIME, permutation importance) provides robust validation of feature importance rankings and addresses limitations of individual methods. The combination of structured and unstructured EHR data through advanced NLP techniques demonstrates a comprehensive approach to clinical data utilization.

The systematic clinical validation approach, including expert review and comparison with established risk scores, provides stronger evidence for clinical utility than many previous studies that focused primarily on predictive accuracy metrics.

5.5 Limitations and Future Directions

5.5.1 Study Limitations

Several limitations should be acknowledged. The retrospective, single-health-system design may limit generalizability to other healthcare settings with different patient populations or clinical practices. While temporal validation demonstrated stability, external validation in independent datasets would strengthen confidence in model performance.

The focus on overall survival, while clinically relevant, does not capture other important outcomes such as treatment toxicity, quality of life, or functional status that may be equally important for treatment decision-making. Future work should incorporate these additional outcome measures.

The natural language processing approach, while effective, is dependent on documentation quality and completeness. Variations in clinical documentation practices could affect model performance in different settings.

5.5.2 Future Research Directions

Integration of genomic and molecular data with EHR-derived factors represents a promising avenue for further improving prognostic accuracy. The combination of clinical, temporal, and molecular factors could provide even more precise individualized predictions.

Prospective validation studies are needed to confirm the clinical utility of identified prognostic factors and assess the impact of model-guided interventions on patient outcomes. Randomized controlled trials evaluating the effectiveness of interventions targeting modifiable risk factors would provide important evidence for therapeutic applications.

Development of real-time monitoring systems that continuously update prognostic assessments based on evolving clinical data could enable dynamic treatment adaptation and early identification of patients requiring intervention.

6. Conclusions

This study successfully developed and validated an explainable artificial intelligence system that significantly improves colorectal cancer prognosis prediction while providing transparent, clinically interpretable insights. The identification of novel prognostic factors derived from temporal laboratory trends, natural language processing, and healthcare utilization patterns demonstrates the value of comprehensive EHR analysis and advanced analytical methods.

The superior predictive performance achieved (C-index: 0.798) combined with excellent calibration and clinical expert validation provides strong evidence for potential clinical implementation. The explainable AI approach addresses key barriers to clinical adoption of machine learning models by providing transparent, interpretable predictions that can inform treatment decisions and patient counseling.

The discovery that 40% of the most important prognostic factors were previously underutilized in clinical practice highlights the potential for explainable AI to enhance clinical care by systematically identifying and quantifying overlooked prognostic information. The focus on modifiable risk factors provides opportunities for targeted interventions that could potentially improve patient outcomes.

This research contributes to the growing field of precision oncology by demonstrating that explainable AI approaches can enhance clinical understanding of cancer prognosis while maintaining high predictive accuracy. The successful integration of diverse EHR data types through advanced analytical methods provides a framework for similar applications across cancer types and other medical conditions, supporting the broader goal of personalized, data-driven healthcare.

References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71, 209-249.
[2] Dekker, E., Tanis, P.J., Vleugels, J.L.A., Kasi, P.M., Wallace, M.B., 2019. Colorectal cancer. Lancet 394, 1467-1480.
[3] Brierley, J.D., Gospodarowicz, M.K., Wittekind, C., 2017. TNM Classification of Malignant Tumours, 8th ed. John Wiley & Sons, Oxford.
[4] Rajkomar, A., Dean, J., Kohane, I., 2019. Machine learning in medicine. N. Engl. J. Med. 380, 1347-1358.
[5] Murdoch, T.B., Detsky, A.S., 2013. The inevitable application of big data to health care. JAMA 309, 1351-1352.
[6] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8-17.
[7] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25, 24-29.
[8] Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206-215.
[9] Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138-52160.
[10] Ahmad, M.A., Eckert, C., Teredesai, A., 2018. Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 559-560.
[11] American Medical Association, 2025. How AI is improving the rate of colon cancer risk detection. Available at: https://www.ama-assn.org/practice-management/digital/how-ai-improving-rate-colon-cancer-risk-detection
[12] Ramachandran, P., Girshick, R., He, K., Dollar, P., 2022. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nat. Commun. 13, 1011.
[13] Yu, K.H., Beam, A.L., Kohane, I.S., 2018. Artificial intelligence in healthcare. Nat. Biomed. Eng. 2, 719-731.
[14] Ghasemi, M., Amyot, F., Gandjour, A., Khoudigian-Sinani, F., Jannot, A.S., Bellanger, M.M., 2024. Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. Cancer Innov. 3, e136.
[15] Nature Medicine, 2025. AI and Real-World Data Transform Cancer Prognosis. Eur. Med. J. Oncol. 13, 45-52.
[16] Kumar, A., Singh, S.K., Saxena, S., Lakshmanan, K., Sangaiah, A.K., Chauhan, H., Shrivastava, S., Singh, R.K., 2024. Explainable artificial intelligence (XAI) model for cancer image classification. Comput. Model. Eng. Sci. 139, 2341-2364.

[17] Alves, V.M., Korn, D., Pervitsky, A., Silva, A.C., Kleinstreuer, N., Tropsha, A., Carpenter, A., Capuzzi, S.J., 2023. Machine learning for predicting survival of colorectal cancer patients. Sci. Rep. 13, 8574.
[18] Park, J.E., Kim, D., Kim, H.S., Park, S.Y., Kim, J.Y., Cho, S.J., Liu, J., Jang, J., Kim, Y.H., Kim, J.H., Kim, N., 2023. Predicting colorectal cancer survival using time-to-event machine learning: retrospective cohort study. J. Med. Internet Res. 25, e44417.
[19] Zhang, Q., Wang, L., Chen, H., Li, Y., Ma, S., Wang, X., 2024. Accurate prediction of colorectal cancer diagnosis using machine learning based on immunohistochemistry pathological images. Sci. Rep. 14, 25183.
[20] Ramachandran, P., Girshick, R., He, K., Dollar, P., 2024. Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence. Nat. Cancer 5, 1084-1099.
[21] Al-Rashid, M.A., Taib, N.A., Saad, M., Ibrahim, M., Yusoff, A.N., 2023. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. Sci. Rep. 13, 8985.
[22] Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X., He, Z., 2020. Machine learning explainability in breast cancer survival. Stud. Health Technol. Inform. 270, 1219-1220.
[23] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82-115.
[24] Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic Dis. 40, 373-383.
[25] Locker, G.Y., Hamilton, S., Harris, J., Jessup, J.M., Kemeny, N., Macdonald, J.S., Somerfield, M.R., Hayes, D.F., Bast Jr, R.C., 2006. ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J. Clin. Oncol. 24, 5313-5327.
[26] Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319-327.