Stage-specific survival prediction in gastrointestinal cancers using explainable machine learning on SEER data

Parisa Poorhasani[1], Bogdan Iancu[1,*]

[1]Department of Engineering and Information Technology, Faculty of Science and Engineering, Åbo Akademi University, Turku, 20500, Finland.

Abstract

Disease stage is an important factor in determining cancer survival rates. Yet, many models used today (trained on a large pool of stage-agnostic data) have a tendency to overestimate their ability to predict cancer survivability and mask the differences in survivability among different stages of cancer. Using SEER Program data on gastrointestinal cancers (colorectal, stomach, and liver), we developed and evaluated explainable ML strategies to train separate models for each stage. The four machine learning algorithms tested included Logistic Regression, AdaBoost, Random Forest, and CatBoost. Of these, CatBoost performed best and showed the highest consistency across cancer types and stages. To further improve the transparency and understanding of how the models made predictions for individual cases, two other techniques were used: SHapley Additive exPlanations (SHAP), for determining which variables contributed to the overall model's prediction, and Local Interpretable Model Agnostic Explanations (LIME), for explaining how those contributions were made based on the features of a single case. Both SHAP and LIME were utilized for all nine combinations of cancer and stage. Across both methods, the models relied primarily on established prognostic indicators (age, surgical intervention, tumor extent, and presence of metastasis), and there was a significant shift in the relative contribution of these variables by stage. It was also found that the lack of pathologic evaluation of lymph nodes provided evidence of a regional disease pattern, and that histologic subtype codes became strong predictors in distant-stage cancers, indicating that tumor biology becomes increasingly important as anatomically defined barriers are breached. To generate more accurate predictions and, therefore, improve the quality of care for patients with cancer, as well as to create a platform that enables healthcare providers and their patients to make informed decisions together about cancer treatments, this research suggests that machine learning models should be enhanced with explanations. The inclusion of explanations in machine learning models will allow clinicians to understand the key features that influence survivability. In addition, these models can be used in clinical settings to develop personalized treatment plans and improve patients' quality of life.

Keywords

Cancer survivability prediction, stage-specific survivability, explainable machine learning

1. Introduction

Cancer is a major public health concern and one of the leading causes of death worldwide, with its global incidence projected to increase over the coming decades due to demographic aging, lifestyle-related risk factors, and environmental carcinogens. By 2040, an estimated 28 million additional cases and 16.2 million deaths are expected worldwide [1, 2, 3]. These rising morbidity and mortality rates have led academics and bioinformaticians to develop advanced machine learning (ML) models for predicting cancer survivability [4, 5].

Survival analysis, as a statistical model, is widely used in clinical oncology to provide prognostic information by estimating the likelihood of survival beyond a specified duration [6]. ML-based predictive models are increasingly important for improving cancer care, as they analyze tumor features, treatment, and demographic characteristics to accurately predict patient outcomes, and have advanced considerably in providing individualized prognostic assessments [7].

Accurate prediction of cancer survivability requires understanding how different disease stages influence outcomes, as each stage reflects a different degree of tumor advancement with unique

TKTP 2026: Annual Doctoral Symposium of Computer Science, 8.—9.6.2026 Turku, Finland

*Corresponding author.

pourhasaniparisa@gmail.com (P. Poorhasani); bogdan.iancu@abo.fi (B. Iancu)

0009-0004-3012-7763 (P. Poorhasani); 0000-0001-8801-7250 (B. Iancu)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-02

CEUR-WS.org/Vol-4309/paper07.pdf



Figure 1: ML workflow, including data collection, preprocessing, model training, evaluation, and interpretation using SHAP [8] and LIME [9].

prognostic implications. In this study, we focused on three stages classified by SEER [10]: localized (confined to the primary organ), regional (extension to adjacent lymph nodes, tissues, or organs), and distant (metastasis to remote sites). This stage-specific stratified approach enables more refined modeling of survival outcomes.

Explainable ML is essential for advancing the clinical integration of ML in healthcare, with robustness and interpretability increasingly emphasized as fundamental requirements for responsible ML frameworks [11, 12, 13, 14].

Impact statement: This study shows that training ML models to produce survival estimates for each cancer stage separately leads to more interpretable prognoses. Systematic use of SHAP and LIME on 9 cancer-stage combinations reveals how the influence of established clinical risk factors shifts as the disease advances, e.g., surgery and tumor characteristics in localized disease, lymph node evaluation in regional disease, and histologic subtypes in advanced-stage cancers, changes not identifiable using a stage-agnostic approach. The study achieves high predictive accuracy and demonstrates a method for auditing explainability techniques to ensure ML models produce clinically relevant outputs across cancer stages, crucial for responsibly integrating these technologies into decision support systems.

Organization: The remainder of this paper is organized as follows. Section 2 provides an overview of the materials and methods used in this study. The experimental findings, along with the ML explainability visualizations, are presented in Section 3. Section 4 provides a brief discussion of the most significant factors in survival outcomes. Section 5 summarizes the conclusions of this study and outlines prospective directions for future work.

2. Materials and Methods

The ML workflow adopted in this study commenced with the collection of data from the SEER dataset [10] for colorectal, stomach, and liver cancer. Following data cleaning, one-hot encoding, and splitting by cancer stage, several machine learning models (Logistic Regression [15], AdaBoost [16], Random Forest [17], and CatBoost [18]) were trained using the selected clinical hyperparameters. Hyperparameter tuning was performed using grid search, and the models were evaluated using the appropriate performance metrics. All preprocessing steps, including Z-score normalization, were performed within the training folds and subsequently applied to the validation and test data to prevent data leakage. Finally, to ensure interpretability, SHAP [8] is used for global explanations, whereas LIME [9] provides local, instance-based insights into model predictions.

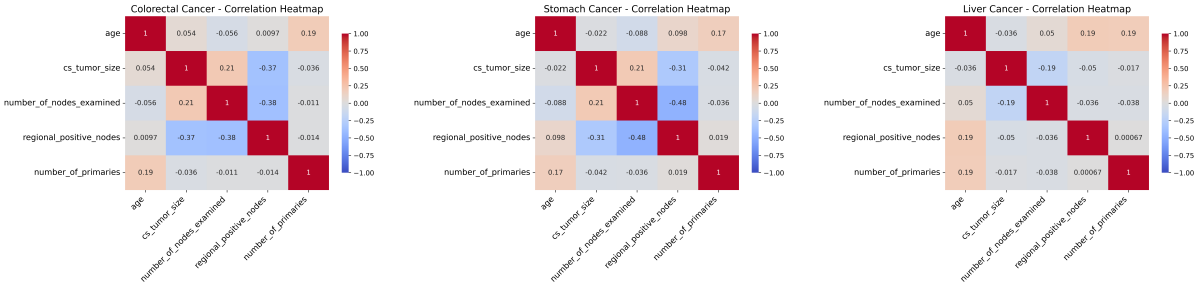
2.1. Data Preparation and Preprocessing

This study used data from the Surveillance, Epidemiology, and End Results (SEER) Program [10] on gastrointestinal cancers, including colorectal, stomach, and liver cancers, diagnosed between 2004 and 2015 (data accessed in March 2025). This publicly available database was selected for its high-quality, non-

Table 1
Overview of the predictor variables applied in the ML models and their clinical descriptions.

Feature/Variable	Description	Feature/Variable	Description
Ordinal Features		**Nominal Features**	
Extension [cs extension code]	Extent of primary tumor spread.	Behavior Code	Indicates tumor aggressiveness.
Grade	Differentiation level of tumor cells.	Histologic Type [ICD-O-3]	Histological classification of tumor.
Lymph Nodes [cs lymph nodes]	Lymph nodes with cancer involvement.	Metastasis at Dx [cs mets at dx]	Distant metastases at diagnosis.
Marital Status	Marital status at diagnosis.	Primary Site	Primary tumor location in the body.
Radiation	Initial radiation therapy type.	Race	Patient racial/ethnic background.
Surgery Code [RX Summ-Surg Prim Site]	Surgery type on the primary site.	Sequence Number	Order of tumor among multiple tumors.
Summary Stage	Overall cancer stage summary.	Sex [Gender]	Gender of the patient as recorded.
Numeric Features			
Age	Age at cancer diagnosis.		
Regional Nodes Examined	Number of lymph nodes examined.		
Regional Nodes Positive	Number of positive lymph nodes.		
Tumor Size [cs tumor size]	Tumor size in millimeters.		

identifiable cases of various malignancies, with the SEER program performing annual outlier detection and quality assurance via the SEER Box Plot Outlier Tool (BOT) [19]. The dataset features included demographic information, tumor characteristics, treatment, and other survival-associated factors (listed in Table 1), selected based on previous studies [7, 20, 21] to ensure clinical relevance and consistency with existing research. The multicollinearity assessment indicated no strong multicollinearity among the analyzed features. Correlation heatmaps were generated for each cancer type (Figure 2) to visualize pairwise correlations between numerical features such as age, tumor size, number of examined lymph nodes, number of positive lymph nodes, and number of primaries, providing initial insights into the data structure. As correlations were generally weak and no strong multicollinearity was detected, no numerical features were removed.



(a) Colorectal Cancer (b) Stomach Cancer (c) Liver Cancer

Figure 2: Correlation heatmaps between numerical features for each cancer type, respectively.

The dataset underwent comprehensive preprocessing, including handling missing values, encoding categorical data, and normalizing continuous variables, encompassing three cancer types totaling 78,150 patient records. Rows with missing values in selected features were removed to ensure data quality. After removing records with missing values and applying the five-year survival criteria, the remaining cases were stratified into localized, regional, and distant stages before model development. The final preprocessed datasets included 53,100 colorectal, 11,350 stomach, and 13,700 liver cancer cases.

Prior to model training, numerical features were standardized using Z-score normalization (StandardScaler) to ensure comparable feature scales and improve model stability. To prevent data leakage, the scaler was fitted exclusively on the training data and subsequently applied to the validation and test data within a machine learning pipeline. To address class imbalance, cost-sensitive learning was employed through class-weighting strategies during model training. Unlike oversampling or undersampling approaches, class weighting preserves the original data distribution while increasing the importance of minority-class instances. This approach was selected to avoid information loss and potential overfitting associated with data resampling techniques.

2.2. Cancer Survivability

This study determined the gold standard for five-year cancer survivability across cancer types using the rationale in [7], as shown in Figure 3. This process aimed to establish a reliable standard for measuring the duration of patient survival after the diagnosis. To calculate this standard, we used three features: vital status record, survival months, and cause of death, which are available in the SEER dataset [10]. Vital status determines whether a patient is alive or dead at a specific time. Survival months measure the duration from diagnosis to death or the last follow-up date. The cause of death helps to focus on deaths caused by cancer.

Five-Year Cancer Survival Standard
if Survival Months ≥ 60 **and** Vital Status Recode = alive $\rightarrow$ **survived**
else if Survival Months < 60 **and** Cause of Death = "Cancer Type" $\rightarrow$ **not survived**
else patient exclusion

Figure 3: Rule-based criteria for defining five-year cancer survival status [7].

The SEER dataset categorizes cancer into three stages: *localized*, *regional*, and *distant*. In the *localized* stage, cancer remains confined to the organ of origin. In the *regional* stage, it disseminates to adjacent locations, such as lymph nodes or tissues proximal to the organ. The *distant* stage occurs when cancer metastasizes to distant organs or tissues within the body. In this study, we developed models to predict and understand cancer at each stage.

2.3. Machine Learning Model Construction and Evaluation

The training phase included multiple ML models for the prediction of survivability including Logistic Regression [15], AdaBoost [16], Random Forest [17], and CatBoost [18].

The selection of algorithms was based on their proven predictive efficacy and capacity to handle the complexity inherent in high-dimensional healthcare data. A previous study [7] developed stage-specific cancer survivability prediction models using traditional classification algorithms such as Logistic Regression and Decision Trees. Their results demonstrated the effectiveness of these methods in multiple cancer types and at different clinical stages. We expanded their methodology by incorporating ensemble and boosting-based models, including AdaBoost and CatBoost, to enhance the predictive performance and better capture complex patterns in the data. The model training regimen is depicted below.

Table 2: Hyperparameter ranges used during grid search for each ML model.

Model	**Param**	**Values**
Logistic Reg.	C	0.001–10
	Class Wt	None, balanced
AdaBoost	n_est	50–200
	LR	0.01–1.0
	Algo	SAMME variants
Random Forest	n_est	100, 200
	Depth	3–7
	Min Split	2, 5
	Min Leaf	1–4
	Class Wt	balanced
CatBoost	Iter	100, 200
	Depth	3–7
	LR	0.03, 0.1
	L2	1, 3, 5
	Weights	[1,1]–[1,5]

2.4. Model Training Regimen

Optimal hyperparameters for each model were determined using GridSearchCV with stratified five-fold cross-validation on the training data. This ensured that preprocessing and model selection were conducted independently within each training fold while preserving the distribution of survival outcomes across training and validation splits. The use of stratified cross-validation allowed efficient utilization of the available data and reduced the risk of overfitting. Additionally, because the proportions of cancer subtypes and stages differ, we set the class weight parameter to "balanced", which adjusts the weights of each class when classes have different numbers of instances. The test set remained completely isolated throughout all phases of this project and was only utilized at the end for testing and evaluating the final models, thereby preventing any form of data leakage or overfitting through exposure of test samples

during either model training or hyperparameter optimization. A fixed random seed (random_state = 42) was applied across all experiments to ensure reproducibility.

2.5. Model Explainability

Machine learning models are often perceived as black boxes, and their decision-making process is not easily understood. To address the explainability of ML models for cancer survival prediction, we employed two prevalent XAI techniques: SHapley Additive exPlanations (SHAP) [8] and Local Interpretable Model-agnostic Explanations (LIME) [9].

SHAP [8] employs concepts from game theory, where the "game" is the model and the "players" are the input features, assigning each feature a Shapley value that quantifies its contribution, with properties such as efficiency, symmetry, and consistency giving it a strong theoretical foundation. Combining values across instances estimates feature importance on a global scale.

LIME [9] explains individual predictions by treating the model as an oracle, requiring only its input-output behavior. It generates perturbed samples around the instance under explanation, weights them by proximity, and fits a simple surrogate model (e.g., a linear model or shallow decision tree) to the weighted data, giving LIME reliable local interpretability. Unlike SHAP's axiomatic determination of exact marginal contributions, LIME relies on a surrogate tailored to the local neighborhood. SHAP provided global explainability by quantifying each feature's contribution (e.g., tumor size, cancer stage) across the dataset, while LIME offered local interpretability for patients with the lowest predicted survival probability, across all stages combined and within each clinical stage (localized, regional, distant) for all studied cancer types.

Although explainable ML improves transparency, it has limitations: Salih et al. [22] noted that using SHAP and LIME in sensitive domains like healthcare requires careful consideration, as misinterpreting their outputs without understanding their assumptions can cause serious errors and unintended consequences for patients.

3. Results

The study cohort for ML model development comprised *n = 53,100* cases of *colorectal cancer*. The mean age at diagnosis was 66.6 (SD ± 14.2: range 8-90) and the median age was 67, with a relatively balanced distribution across three main stages. Localized tumors accounted for 33.1% of cases (n = 17,582), followed closely by regional stage with 35.6% (n = 18,884). Distant-stage cases accounted for 31.3% (n = 16,634), reflecting a broad representation of disease progression. For the *stomach cancer* dataset *(n = 11,350)*, the mean age at diagnosis was 66.2 (SD ± 14.3: range 1-90) and the median age was 67. The stomach dataset exhibited a distinct distribution, with distant-stage cases most frequent at 43.2% (n = 4,903). Regional cases comprised 29.0% (n = 3,292), whereas localized cancers comprised 27.8% (n = 3,155), indicating a tendency toward late-stage diagnoses in this population. The *liver cancer (n = 13,700)* cohort revealed a mean age at diagnosis of 63.2 (SD ± 13.5: range 0-90) and a median age of 62. In this cancer type, localized cases were the most common, representing 49.6% of the total (n = 6,791), followed by regional cases at 33.0% (n = 4,520). Distant-stage cancers were less frequent, accounting for 17.4% (n = 2,389), suggesting a higher proportion of early detection in this group.

The experimental setup involved stratifying the SEER dataset [10] into localized, regional, and distant stages for colorectal, stomach, and liver cancer. The ML models were trained using stratified 5-fold cross-validation and evaluated using metrics such as accuracy, precision, recall, F1 Score, and AUC-ROC. CatBoost exhibited the best performance across most stages and cancer types. The overall predictive performance of the models across cancer types and stages is presented in Table 3.

Model performance declines in late-stage disease across the three cancer categories, affecting the models' ability to accurately identify those who will not survive. Although in distant colorectal cancer, the top-performing model (CatBoost) has a recall of 0.486, an F1 score of 0.451, with an AUC of 0.860, in distant liver cancer, the CatBoost model performs slightly better, with an overall AUC of 0.811 and a recall of 0.545. Distant stomach cancer is a partial exception – the CatBoost model achieves an AUC of

Table 3
Performance of ML models for survival prediction across cancer types and stages.

Cancer	Stage	Logistic Regression					Random Forest					AdaBoost					CatBoost				
		Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC
Colorectal	Localized	0.858	0.862	0.987	0.920	0.805	0.859	0.865	0.982	0.920	0.817	0.857	0.870	0.972	0.918	0.813	0.864	0.873	0.977	0.922	0.821
	Regional	0.677	0.642	0.908	0.752	0.758	0.700	0.658	0.924	0.769	0.791	0.710	0.726	0.774	0.749	0.797	0.722	0.682	0.912	0.780	0.809
	Distant	0.886	0.323	0.482	0.386	0.821	0.913	0.416	0.406	0.411	0.839	0.928	0.590	0.092	0.160	0.845	0.912	0.422	0.486	0.451	0.860
Stomach	Localized	0.887	0.893	0.945	0.918	0.937	0.901	0.919	0.936	0.927	0.951	0.887	0.907	0.927	0.917	0.952	0.895	0.880	0.976	0.926	0.956
	Regional	0.758	0.527	0.733	0.613	0.824	0.822	0.632	0.767	0.693	0.865	0.830	0.731	0.552	0.629	0.877	0.830	0.650	0.756	0.699	0.880
	Distant	0.976	0.741	0.541	0.625	0.953	0.963	0.512	0.595	0.550	0.942	0.962	0.500	0.676	0.575	0.938	0.968	0.558	0.784	0.652	0.955
Liver	Localized	0.751	0.582	0.816	0.679	0.842	0.772	0.614	0.794	0.693	0.854	0.785	0.641	0.761	0.696	0.860	0.781	0.634	0.768	0.694	0.868
	Regional	0.905	0.524	0.530	0.527	0.826	0.920	0.635	0.482	0.548	0.846	0.899	0.495	0.554	0.523	0.839	0.919	0.614	0.518	0.562	0.860
	Distant	0.977	0.500	0.545	0.522	0.899	0.975	0.462	0.545	0.500	0.890	0.981	0.600	0.545	0.571	0.880	0.987	0.857	0.545	0.667	0.811

0.955, yet the highest F1 score is only 0.652. There is no strong indication that any particular algorithm failed at the distant stage; models show a similar trend, which suggests that the difficulty lies in the data and the task being predicted, rather than in model choice or hyperparameter optimization.

There are several possible explanations for the model's performance degradation, and each of these could have implications for clinicians. We present them as follows. Class imbalance often worsens at the distant stage, as five-year survival rates for metastatic gastrointestinal cancers are poor [23], leading to very few members of the "survived" class. This imbalance is greatest for distant liver cancer, since its dataset is by far the smallest of all nine configurations (n = 2,389), with a particularly sparse survivors class. While standard class weighting techniques can counteract this imbalance to some extent, they generally cannot adequately correct it when the minority class is too small to allow sufficient representation. Therefore, while the model learns the dominant pattern well, given the good accuracy scores, it finds it very difficult to determine what separates the rare patients who survive metastatic disease from the much larger group of non-survivors.

The feature space also becomes less discriminatory in advanced stages of the disease. Clinical features present in localized and regional stages provide clear prognostic divisions [24], clearly associated with subsequent survival outcomes. However, this relationship breaks down in later disease states, where all patients have spread to distant sites, limiting the range of a potential prognosis. Factors that differentiate between distant-stage survivors and non-survivors increasingly include variables not represented within the SEER database, such as responses to systemic chemotherapies, molecular and genetic tumor profiles, performance status, comorbidities, availability of specialist oncological care, and specifics about both site(s) and volume of metastases.

3.1. Explainability and Interpretability of the CatBoost Model

As CatBoost consistently demonstrated superior predictive performance across all three cancer types and stages, explainability analyses were conducted using this model. Figure 4 shows the top five SHAP-ranked predictors for each cancer type and clinical stage. In localized disease, initial treatment ('Surgery Code') and tumor burden ('Tumor Size' and 'Extension'), together with age, dominate. In regional disease, age and tumor spread consistently appeared among the top features across all three cancers, whereas lymph node indicators, namely 'Regional Nodes Positive' and 'Regional Nodes Examined', were prominent mainly in colorectal and stomach cancers. Radiation was also one of the top features for these two cancers. In distant disease, histologic subtypes (SEER codes 8936 for stomach; 9133 and 8970 for liver) emerged beyond Age and Surgery Code, while nodal variables remained important in colorectal cancer.

Figure 5 shows the presence of LIME-based key features across colorectal, stomach, and liver cancers, stratified by disease stage. Several general patterns have emerged. Age consistently appeared across nearly all cancers and stages as an influential factor, reaffirming its well-established role as a strong prognostic indicator of cancer survival. The surgical code, particularly its absence (value = 0), was prominently highlighted in stomach and liver cancers at multiple stages, underscoring the clinical importance of surgical treatment in the prediction of survival.

Among the lymph node related variables, the category 'Regional Nodes Positive' with the value of 98, which in the SEER dataset denotes that no pathological examination of lymph nodes was performed, emerged exclusively in the regional stage for colorectal and stomach cancers and was absent in other stages. This is clinically coherent, as nodal status is key to regional disease, and absent evaluation (Code 98) may reflect diagnostic or therapeutic limitations linked to poorer outcomes.

Another notable group of variables comprised histologic subtype codes (e.g., 8240 and 8936 for stomach; 9133 and 8970 for liver), which were predominantly observed in metastatic (distant-stage) disease, suggesting that tumor biology and subtype heterogeneity exert a greater influence on survival in the advanced stages. Among the indirect clinical factors, marital status appeared only once, in the localized stage of liver cancer patients, indicating a limited, stage-dependent role in survival prediction.

The most clinically relevant result from the LIME analysis was the repeated dominance of SEER Code 98 'Regional Nodes Positive', indicating insufficient pathological examination of lymph nodes, in regional-stage colorectal and stomach cancer, with a predictive strength comparable to traditional pathologic characteristics. While this represents a simple coding artefact, the model's treatment of it as an extremely influential predictor is accurate. It has become a powerful composite indicator of several unmeasured aspects related to surgery (inoperability, functional status issues, palliative treatment goals, etc.) as well as potential barriers to receiving appropriate health care. This demonstrates the unique ability of explainable ML models to identify factors with predictive power not because they directly measure clinical conditions, but because they represent values for unseen ones. Clinicians may be aware of the significance of absent nodal evaluation, but may not be able to quantify its association with increased mortality relative to other factors; SHAP and LIME provide a quantitative representation of this previously implicit relationship.

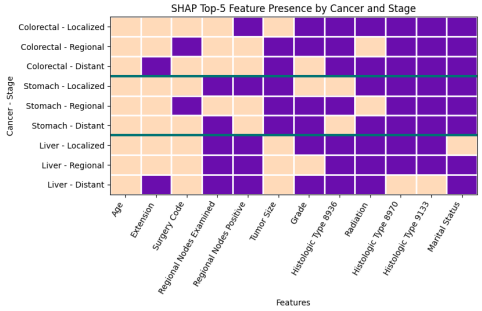


Figure 4: Binary heatmap of top five SHAP features across cancer types (Colorectal, Stomach, and Liver) and stages (Localized, Regional, and Distant). Orange cells indicate that the feature was among the top five most important features for the corresponding cancer-stage group.

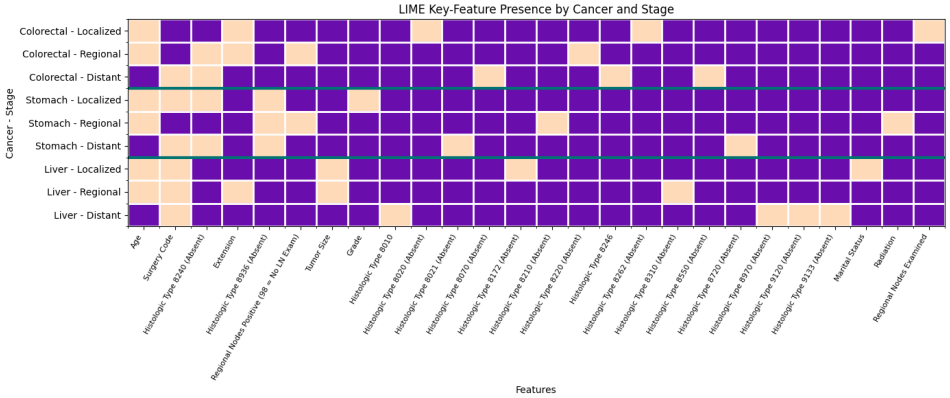
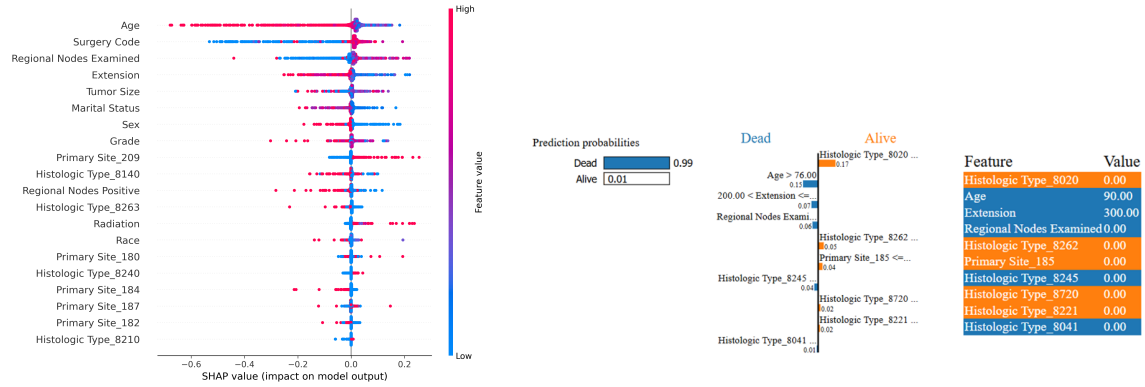


Figure 5: Binary heatmap showing the presence of top-5 LIME features across cancer types (Colorectal, Stomach, and Liver) and stages (Localized, Regional, and Distant). Orange cells indicate that the feature was among the top-5 most important features for the corresponding cancer-stage group.



(a) SHAP beeswarm summary plot for the prediction of survivability of colorectal cancer at localized stage using CatBoost. Advanced age and greater extension drive the prediction toward *Dead*, with small countervailing effects from selected histologic codes.

(b) LIME explainability for a single instance (patient). *Regional Nodes Examined* reported as zero, together with age and extension, are the strongest local drivers toward *Dead*.

Figure 6: Case-level explanations for a representative patient with *colorectal* cancer at the *localized* stage : (a) SHAP summarizes global impact directions, whereas (b) LIME provides a local, instance-specific rationale.

3.2. Patient-Level Explanations

To illustrate patient-level explanations, we present a representative case of localized colorectal cancer, a 90-year-old patient who had an Extension value of 300 in the SEER coding system; the tumor was in the colon but there was insufficient information regarding the exact location. There was no documented lymph node involvement (Regional Nodes Examined = 0), and therefore, the model assigned a very high probability of mortality.

Figure 6a shows the SHAP summary plot for the *localized stage of colorectal cancer*, highlighting the key features that influence the model predictions. The most important factors included 'Age', 'Surgery Code', 'Regional Nodes Examined', 'Extension', and 'Tumor Size'. Lower SHAP values pushing the prediction toward "Dead" suggest that older age (i.e., 'Age') increases the risk of death, which is consistent with previous studies [25]. Undergoing surgical treatment (indicated by a higher 'Surgery Code') and having a greater number of lymph nodes examined (i.e., 'Regional Nodes Examined') are associated with better survival outcomes, as reported in earlier studies [26, 27]. In contrast, larger 'Tumor Size' and higher 'Extension' values, represented by red points on the left side of the plot, were associated with worse outcomes. These observations align with the findings of previous studies [28, 29].

The LIME visualisation (Figure 6b) reveals three key factors that contributed to the 'Dead' classification: the patient was 90 years of age, had a high 'Extension' value, and had no lymph node evaluation performed. These three factors are individually plausible based upon clinical evidence. A 90 year-old patient represents a significant independent predictor of mortality for virtually every type of cancer [24]. It is most likely that the absence of any lymph node examination indicates a decision to pursue less aggressive diagnostic or surgical approaches; this would be compatible with either palliative or conservative management strategies in older adults, as opposed to providing evidence of favorable nodal status. In addition, although the 'Extension' value of 300 indicates that the cancer is localized, it is possible that this value reflects some level of uncertainty in making a definitive diagnosis, rather

Table 4: Comparison of numerical features for survivor and non-survivor groups for each cancer type.

Cancer	Age	Nodes Exam.	Nodes Pos.	Tumor Size	p-Value
Survivors					
Colorectal	63.2	16.0	18.9	32.5	<0.001
Stomach	60.5	11.1	43.3	27.9	<0.001
Liver	55.6	2.5	84.1	36.2	<0.001
Non-survivors					
Colorectal	70.0	13.8	37.8	38.9	<0.001
Stomach	68.5	7.1	70.0	31.1	<0.001
Liver	64.9	6.8	96.3	47.8	<0.001

than being indicative of favorable tumor biology. Collectively, these factors support the conclusion that the patient likely received non-curative therapy.

As such, LIME provides no contradictory findings; instead, it presents a clinically reasonable scenario. Although there was a slight positive impact towards 'Alive', this was due to the fact that the patient did not possess two particularly aggressive histological subtypes: undifferentiated carcinoma (Type 8020) and adenocarcinoma in a papillary adenoma (Type 8262); this is supportive of clinical expectations, since patients who do not have one or both of these high-risk subtypes tend to have a slightly better prognosis. However, this protective effect was insufficient to counteract the primary risk factors previously identified.

It is important to note that these local LIME findings are entirely consistent with the global SHAP patterns presented in Section 3.1. Specifically, age, 'Extension' values and lymph node examination represented the primary risk factors for localized colorectal cancer in all patients. Therefore, consistency between LIME-based (individual-level) and SHAP-based (population-level) explanations demonstrates that the model uses appropriate logic both consistently and appropriately when predicting mortality for individual cases; this is an important consideration for developing confidence in utilizing the model in a clinical setting.

3.3. Sensitivity Analysis of Regional Nodes Positive (Code 98)

A common concern is that the model's use of Regional Nodes Positive = 98 may reflect a data artifact (from unaccounted for confounding variables) rather than a legitimate clinical measure. As mentioned earlier in Section 3.1, code 98 is a surrogate indicator for many variables, including operability status, level of function status, palliation intent, and availability/accessibility of health care services. It would therefore be expected that this code would be strongly associated with mortality. To determine whether the predictive power of the CatBoost models remains independent of the inclusion/exclusion of patients with code 98, additional CatBoost models were trained using only patients without this code. As shown in Table 5, model performance decreased slightly for most colorectal and stomach cancer stages, whereas mixed effects were observed for liver cancer. Importantly, Regional Nodes Positive remained among the most influential predictors in both SHAP and LIME analyses after retraining. These findings suggest that the predictive importance of Regional Nodes Positive is not entirely attributable to code 98 and that the variable retains prognostic value even after exclusion of these cases, although its impact varies across cancer types and stages. Because excluding code 98 cases alters cohort composition beyond the regional stage where the code was most LIME-salient, some of the larger shifts (e.g., stomach-localized) partly reflect changes in the retained sample rather than the removal of the code 98 signal itself.

Table 5: Sensitivity analysis of CatBoost models after excluding patients with Regional Nodes Positive coded as 98. Positive Δ values indicate performance improvement after exclusion, whereas negative values indicate degradation.

Cancer	Stage	Acc (Orig.)	Acc (w/o 98)	ΔAcc	AUC (Orig.)	AUC (w/o 98)	ΔAUC
Colorectal	Localized	0.864	0.848	-0.016	0.821	0.772	-0.049
	Regional	0.722	0.710	-0.012	0.809	0.801	-0.008
	Distant	0.912	0.835	-0.077	0.860	0.810	-0.050
Stomach	Localized	0.895	0.808	-0.087	0.956	0.793	-0.163
	Regional	0.830	0.834	+0.004	0.880	0.883	+0.003
	Distant	0.968	0.884	-0.084	0.955	0.836	-0.119
Liver	Localized	0.781	0.906	+0.125	0.868	0.906	+0.038
	Regional	0.919	0.800	-0.119	0.860	0.878	+0.018
	Distant	0.987	0.844	-0.143	0.811	0.920	+0.109

Δ = Performance after excluding code-98 cases minus original.

4. Discussion

We can distinguish between two claims of statistical discrimination and actionable use. Statistical discrimination is demonstrated by our models' ability to accurately predict survival outcomes for patients with localized and regional cancers, with AUC values up to 0.956. Model performance degrades for patients with distant-stage cancers. Both SHAP and LIME show that the models rely on clinically validated and well-established prognostic features. Therefore, this suggests the models have learned

clinically valid structures, a test of the models' validity, but not a demonstration of clinical applicability. Our explanations were overwhelmingly confirmatory; age, surgical intervention, nodal assessment, and the extent of the tumor all contributed to the model's predictions. These provide clinicians with reassurance regarding their diagnostic and therapeutic decisions, but they do not provide additional information on which to base those decisions. One potential exception to this may be the association between the absence of nodal evaluation (code 98) and increased mortality. Although this association is the strongest candidate for providing actionable information, it is also likely to be influenced by many other unmeasured confounders. As such, we cannot draw any conclusions regarding treatment based on this variable.

To determine the factors influencing survival outcomes, we examined the feature distributions between the survivor and non-survivor groups within the dataset. For the comparison between survivors and non-survivors, the Mann–Whitney U test was applied to numerical variables ('Age', 'Regional Nodes Examined', 'Regional Nodes Positive', and 'Tumor Size'), using a two-tailed alternative hypothesis.

Table 4 demonstrates statistically significant differences in the mean values of key clinical features between survivors and non-survivors in all cancer types studied. These findings are consistent with the model interpretation results obtained using SHAP and LIME. Non-survivors were generally older, had more positive lymph nodes, and presented with larger tumors than survivors. In contrast, survivors had more lymph nodes examined, suggesting that more comprehensive diagnostic and therapeutic procedures were performed in their cases. This concordance between statistical trends and model interpretability enhances confidence in the clinical relevance of the model's predictions.

5. Conclusion and Future Work

In this study, survival prediction for patients with colorectal, stomach, and liver cancers was performed using stage-specific modeling (localized, regional, and distant). The results showed that model performance varied across stages, and that stage-wise analysis more accurately reflected heterogeneity in patient survival. For example, in colorectal cancer, although the models performed well in the early stages, their sensitivity decreased noticeably in the advanced stage. Similar patterns were observed in stomach and liver cancers, where differences between early and advanced stages were only clear through stage-specific modeling. These findings indicate that stage-based analysis prevents overly optimistic estimates and provides a more accurate and clinically meaningful picture of survival patterns.

Beyond predictive performance, this study emphasizes the explainability of the model. Using SHAP and LIME, we showed that the models relied primarily on well-established prognostic features such as age, surgery, tumor size, extension, and lymph node involvement. These findings align with prior evidence, where older age has consistently been associated with increased mortality risk [25], whereas surgical treatment and a higher number of lymph nodes examined have been linked to improved survival [26, 27]. Thus, the primary value of explainability lies not in discovering novel predictors but in confirming that machine learning models learn clinically relevant information. This is essential for ensuring transparency and trust in clinical applications. Simultaneously, some less obvious patterns, such as the prognostic role of "code 98" for lymph nodes (indicating the absence of pathological evaluation) or the scattered influence of rare histologic subtypes, illustrate how explainable ML methods can also uncover subtle, stage-dependent signals. As our sensitivity analysis shows, such signals are partly proxies for unmeasured clinical conditions and warrant cautious interpretation and further investigation rather than direct clinical action.

Overall, the findings of this study show that stage-specific modeling combined with explainability methods provides more accurate and clinically meaningful survival predictions. Together, these aspects strengthen the role of explainable machine learning as a trustworthy decision-support tool in oncology research. Nevertheless, as emphasized by Salih et al. [22], interpretability methods such as SHAP and LIME are based on mathematical approximations; if misinterpreted or applied without awareness of their assumptions, they may lead to errors and unintended outcomes. Therefore, the use of explainability in

medicine should always be accompanied by human oversight and clinical judgment to ensure the safe and responsible deployment of ML. Finally, because this study was based on the SEER dataset, further validation is required in independent and more diverse cohorts to confirm the generalizability of our findings. Furthermore, the interpretability of ML models is a critical factor in establishing clinicians' trust and enabling their effective incorporation into clinical decision support systems.

Formal evaluation of both SHAP and LIME for explanation by oncologists is outside of the scope of this study. Future studies will need to evaluate the explanatory power of various XAI approaches through an expert's lens (the combination of multiple explainable AI approaches significantly reduces the likelihood of providing incorrect explanations in situations with high liability, such as treatment planning). The counterfactual explanation method can provide actionable guidance by determining the minimal changes in a patient's characteristics necessary to increase predicted survival. Gradient-based explanations, such as Integrated Gradients, are likely to produce more consistent and reproducible attribution.

Acknowledgment

The authors gratefully acknowledge Prof. Ahmad P. Tafti, PhD, FAMIA, at the University of Pittsburgh, for his support and comments on this research. The authors thank the SEER program for providing access to these datasets.

Declaration on Generative AI

The authors used Grammarly to improve spelling, word choice, readability, and clarity. The authors have reviewed any AI-edited content as necessary and assume full responsibility for the accuracy of the published work.

References

[1] B. Liu, H. Zhou, L. Tan, K. T. H. Siu, X.-Y. Guan, Exploring treatment options in cancer: tumor treatment strategies, Signal transduction and targeted therapy 9 (2024) 175.

[2] C. Mattiuzzi, G. Lippi, Current cancer epidemiology, Journal of epidemiology and global health 9 (2019) 217–222.

[3] X.-W. Tang, J. Jiang, S. Huang, X.-M. Shi, H. Xu, J. Xu, J.-Y. Peng, W. Zhang, L. Shi, X.-L. Zhong, et al., Long-term trends in cancer incidence and mortality among us children and adolescents: A seer database analysis from 1975 to 2018, Frontiers in Pediatrics 12 (2024) 1357093.

[4] I. El Badisy, Z. BenBrahim, M. Khalis, S. Elansari, Y. ElHitmi, F. Abbass, N. Mellas, K. El Rhazi, Risk factors affecting patients survival with colorectal cancer in morocco: survival analysis using an interpretable machine learning approach, Scientific Reports 14 (2024) 3556.

[5] L. Buk Cardoso, V. Cunha Parro, S. Verzinhasse Peres, M. P. Curado, G. A. Fernandes, V. Wünsch Filho, T. Natasha Toporcov, Machine learning for predicting survival of colorectal cancer patients, Scientific reports 13 (2023) 8874.

[6] P. Deepa, C. Gunavathi, A systematic review on machine learning and deep learning techniques in cancer survival prediction, Progress in Biophysics and Molecular Biology 174 (2022) 62–71.

[7] E. S. H. Pour, R. J. Kate, Stage-specific survivability prediction models across different cancer types, in: AMIA Annual Symposium Proceedings, volume 2017, 2018, p. 1421.

[8] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems (2017) 4765–4774.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[10] National Cancer Institute, Surveillance, epidemiology, and end results (seer) program, 2025. URL: https://seer.cancer.gov/, accessed: 2024-12-05.
[11] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, R. Sharma, Explainable ai for healthcare 5.0: opportunities and challenges, IEEE Access 10 (2022) 84486–84517.
[12] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, J. Qadir, Explainable, trustworthy, and ethical machine learning for healthcare: A survey, Computers in Biology and Medicine 149 (2022) 106043.
[13] S. Amirian, L. A. Carlson, M. F. Gong, I. Lohse, K. R. Weiss, J. F. Plate, A. P. Tafti, Explainable ai in orthopedics: Challenges, opportunities, and prospects, in: 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), IEEE, 2023, pp. 1374–1380.
[14] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, P. Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC medical informatics and decision making 20 (2020) 1–9.
[15] D. W. Hosmer, S. Lemeshow, R. Sturdivant, Applied Logistic Regression, 3rd ed., Wiley, Hoboken, NJ, 2013.
[16] Y. Freund, R. E. Schapire, et al., Experiments with a new boosting algorithm, in: icml, volume 96, Citeseer, 1996, pp. 148–156.
[17] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[18] O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 6638–6648.
[19] H.-S. Chen, B. Liu, J. Cucinelli, Detecting outliers with sample size adjustment in the SEER Box Plot Outlier Tool, Technical Report 2018-01, Surveillance Research Program, NCI, 2018.
[20] R. J. Kate, R. Nadig, Stage-specific predictive models for breast cancer survivability, International journal of medical informatics 97 (2017) 304–311.
[21] J. Kim, H. Shin, Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data, Journal of the American Medical Informatics Association 20 (2013) 613–618.
[22] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, G. Menegaz, A perspective on explainable artificial intelligence methods: Shap and lime, Advanced Intelligent Systems 7 (2025) 2400304.
[23] H.-M. Hu, H.-J. Tsai, H.-Y. Ku, S.-S. Lo, Y.-S. Shan, H.-C. Chang, Y. Chao, J.-S. Chen, S.-C. Chen, C.-J. Chiang, et al., Survival outcomes of management in metastatic gastric adenocarcinoma patients, Scientific reports 11 (2021) 23142.
[24] D. Han, Z. Zhang, J. Deng, H. Du, Survival analysis of gastric malignancy patients: identifying key prognostic factors, Translational Cancer Research 14 (2025) 1928–1941.
[25] W. Lin, X. Pan, C. Zhang, B. Ye, J. Song, Impact of age at diagnosis of bladder cancer on survival: a surveillance, epidemiology, and end results-based study 2004-2015, Cancer Control 30 (2023) 10732748231152322.
[26] H. Kim, L. Shen, J. Jeon, Y. D. Han, D. H. Han, M. Jung, S. J. Shin, S. C. You, N. K. Kim, B. S. Min, et al., Number of lymph nodes examined as a prognosis factor in patients with stage ii or iii colon cancer, Clinical Colorectal Cancer (2025).
[27] N. D. Arhin, C. Shen, C. E. Bailey, L. K. Matsuoka, A. T. Hawkins, A. N. Holowatyj, K. K. Ciombor, M. B. Hopkins, T. M. Geiger, A. E. Kam, et al., Surgical resection and survival outcomes in metastatic young adult colorectal cancer patients, Cancer medicine 10 (2021) 4269–4281.
[28] M. Ghorbanpour, S. Salarvand, S. Salarvand, F. Shahsavari, M. Shirkhoda, P. A. Shakib, R. Ghalehtaki, Depth of invasion and extranodal extension: the influential factors to predict survival rate of patients with oral tongue squamous cell carcinoma, Bmc Cancer 24 (2024) 1256.
[29] K. Zhang, A. Su, X. Wang, W. Zhao, L. He, T. Wei, Z. Li, J. Zhu, Y.-W. Chen, Non-linear correlation between tumor size and survival outcomes for parathyroid carcinoma: A seer population-based cohort study, Frontiers in Endocrinology 13 (2022) 882579.