ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 6| June 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

# PREDICTIVE MODELING OF SURVIVAL, TREATMENT DELAY, AND RISK STRATIFICATION IN CANCER CARE: A MACHINE LEARNING APPROACH USING SEER DATA

**Karthik S[1], Prof. Sowmya D S[2]**
[1]*Student, RV Institute of Management*
[2]*Assistant Professor, RV Institute of Management*

**Article DOI: https://doi.org/10.36713/epra22291**
**DOI No: 10.36713/epra22291**

## ABSTRACT

*Objective: This study aims to predict survival outcomes and treatment delays in cancer patients using machine learning models applied to demographic and clinical data from the SEER database, while identifying key prognostic factors and patient subgroups through advanced clustering techniques.*

*Methods: Clinical data from 622,345 patients were analyzed using regression models (Linear, Ridge, Lasso), tree-based algorithms (Random Forest, Gradient Boosting, XGBoost, LightGBM), and a Cox Proportional Hazards (CoxPH) model to evaluate survival prediction. Treatment delay classification employed an 80% accurate model, adjusted for class imbalance. Feature importance analysis and clustering via autoencoder-derived latent features (10 dimensions) combined with KMeans (3 clusters) were used to stratify patients.Performance metrics included R², MAE, MSE, C-index, and Silhouette scores.*

*Results: LightGBM achieved the highest survival prediction accuracy (R² = 0.5278), while CoxPH demonstrated strong discriminative power (C-index = 0.89), identifying advanced SEER stage (HR = 1.68), surgery (HR = 0.27), and marital status as significant predictors. The delay model showed high recall (98%) for delayed cases but poor overall explanatory power (R² = 0.15). Clustering revealed three distinct groups: low-risk early-stage patients (Cluster 0: no chemo, all surgery), intermediate-risk cases (Cluster 1: mixed treatments), and high-risk patients (Cluster 2: aggressive therapies), validated by a moderate Silhouette score (0.42). Non-linear interactions (e.g., chemotherapy, income) significantly influenced predictions. Survival and delay outcomes were negatively correlated (-0.46), suggesting shorter survival linked to longer delays.*

*Conclusion:The study successfully integrates diverse machine learning approaches to predict cancer outcomes, offering actionable risk stratification. LightGBM and CoxPH models provided robust survival insights, while clustering highlighted treatment patterns. Challenges remain in addressing data imbalance and refining delay prediction. This framework enhances prognostic modeling in oncology, emphasizing the need for tailored clinical interventions and model optimization to improve predictive accuracy in real-world healthcare settings.*

**KEYWORDS:** *Machine Learning, Predictive Models, Deep Learning, SEER Database*

## INTRODUCTION

Cancer prognosis and treatment delay prediction are pivotal yet interconnected challenges in oncology, where delays in care can exacerbate disease progression and diminish survival outcomes. Traditional approaches often isolate these tasks, employing conventional algorithms like Cox Proportional Hazards (CoxPH) or Random Forests to predict survival or delays independently. For instance, Wang et al. (2022) demonstrated the efficacy of Random Forests in survival analysis using SEER data, while Zhou et al. (2024) applied deep learning to time-to-treatment predictions. However, such studies rarely explore the dynamic interplay between these outcomes or leverage holistic patient stratification to inform clinical decision-making. This study addresses these gaps by integrating dual predictive modeling—simultaneously forecasting survival times and treatment delays—and introducing a novel risk stratification framework that combines machine learning (ML) with deep learning-driven clustering. Leveraging one of the largest SEER-derived cohorts to date (N = 622,345 patients), this work builds on methodologies from Hou et al. (2023) and Chen et al. (2023) but extends them by synthesizing regression, classification, and clustering techniques into a unified analytical pipeline.

The SEER database's scale and diversity—encompassing demographic, socioeconomic, and clinical variables—enable robust model training and validation. While prior studies, such as Wang et al. (2022), utilized smaller samples (~100k patients), our expansive dataset enhances statistical power, particularly in capturing rare events and non-linear interactions. We evaluate 11 ML models, including gradient-boosted algorithms (LightGBM, XGBoost) and neural networks, to identify optimal performers. LightGBM, with its efficiency in handling large-scale data and ability to model complex relationships, emerged as the top survival predictor ($R^2 = 0.5278$), outperforming traditional linear models. Concurrently, the CoxPH model achieved a C-index of 0.89, reaffirming its utility in survival analysis by quantifying hazard ratios for variables like advanced SEER stage (HR = 1.68) and surgery (HR = 0.27).

For treatment delay prediction, the pervasive issue of class imbalance—common in healthcare datasets—were assessed by prioritizing recall (98% for delayed cases) over accuracy (80%), ensuring clinically actionable alerts for high-risk patients. This approach aligns with real-world priorities, where missing a delayed case carries greater consequences than false alarms. Furthermore, our hybrid clustering framework, combining autoencoder-derived latent features (10 dimensions) with KMeans, identified three distinct patient subgroups: low-risk early-stage patients (Cluster 0: all surgery, no chemotherapy), intermediate-risk cases with heterogeneous treatments (Cluster 1: 40% chemotherapy), and high-risk patients requiring aggressive therapies (Cluster 2: all surgery and chemotherapy). Validated by a Silhouette score of 0.42, these clusters provide actionable insights for personalized care, bridging the gap between population-level predictions and individualized interventions.

Crucially, analysis reveals a moderate negative correlation ($r = -0.46$) between predicted survival and treatment delays, highlighting the clinical urgency of minimizing delays to improve patient outcomes. This finding underscores the complex interplay between demographic factors, social determinants, and treatment modalities in influencing both survival and treatment timeliness. Variables such as marital status and stage at diagnosis showed meaningful nonlinear interactions with survival outcomes, emphasizing the advantage of machine learning models in capturing intricate, real-world patterns that traditional linear models may overlook.

Although the proposed framework advances prognostic modeling in oncology, several challenges remain, particularly in enhancing the accuracy of treatment delay prediction ($R^2 = 0.15$) and mitigating biases inherent in large observational datasets. Future improvements may involve the integration of multimodal data sources (e.g., genomic profiles, imaging data) and the application of advanced imbalance correction techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or Generative Adversarial Networks (GANs) to further refine model performance. Rather than focusing on isolated innovations, the study demonstrates the value of synthesizing diverse methodologies to develop scalable, interpretable tools for clinical oncology. These findings not only reaffirm established prognostic factors but also contribute to the redefinition of risk stratification approaches, offering a template for future research seeking to harmonize predictive accuracy with real-world applicability in cancer care.

## LITERATURE REVIEW

Over the past two decades, the integration of artificial intelligence (AI) and machine learning (ML) has significantly advanced predictive modeling in oncology, enhancing survival prediction accuracy, patient stratification, and treatment optimization. Early work by Ahmed (2005) and Burke et al. (1997) demonstrated the superiority of artificial neural networks (ANNs) over traditional statistical methods in cancer prognosis, particularly in capturing complex, nonlinear relationships in data. Burke et al. further validated ANNs' advantage over the TNM staging system for breast and colorectal cancers using large-scale SEER and ACS datasets, establishing the foundation for continuous, multivariable modeling in oncology. Montazeri et al. (2016) expanded on these findings by comparing multiple ML classifiers, with ensemble methods like Random Forests (TRF) proving most effective due to their robustness in handling heterogeneous data. The field has since evolved toward more sophisticated multimodal approaches, exemplified by Vale-Silva and Rohr's (2021) *MultiSurv*, a deep learning model integrating clinical, genomic, and histopathological data to predict survival across 33 cancer types. This model outperformed traditional Cox regression and addressed missing data challenges while providing clinically interpretable subgroup visualizations.

Real-time data integration has also emerged as a critical focus, with Chen et al. (2024) developing a near-real-time reporting system using e-pathology data to reduce delays in cancer registry updates while maintaining alignment with SEER metrics. Concurrent studies have emphasized the impact of treatment timing on survival outcomes. Pathak et al. (2023) and Chaves et al. (2023) found that delays in initiating treatment for breast and thyroid cancers significantly worsened survival, with vacuum-assisted biopsies yielding better outcomes than core needle biopsies in advanced stages. Conversely, Zhu et al. (2023) observed a nuanced relationship in glioblastoma, where modest delays post-surgery improved survival, likely due to optimized recovery and patient selection. Machine learning clustering techniques have further refined patient stratification, with Mortagy et al. (2024) employing *K-means* to identify distinct survival subgroups in lung neuroendocrine neoplasms (NENs) and Zhong & Zhang (2025) using latent class analysis to reveal disparities in malignant meningiomas, where younger, affluent, urban patients exhibited better outcomes.

Racial and molecular subtype disparities in breast cancer have also been a key area of investigation. Sakhuja et al. (2020) and Li et al. (2024) highlighted the elevated metastatic potential and mortality associated with triple-negative breast cancer (TNBC) and HR−/HER2+ subtypes, particularly among non-Hispanic Black women. Li's nomogram for early death prediction in breast cancer patients with lung metastasis identified surgery as the most protective factor, reinforcing the need for timely intervention. Hybrid modeling approaches, such as those by Momenzadeh et al. (2021) and Xu et al. (2022), have further enhanced predictive accuracy. Momenzadeh combined Factor Analysis of Mixed Data (FAMD) with resampling methods and XGBoost to predict prostate cancer mortality, while Xu's LASSO-Cox model for synchronous colorectal carcinoma demonstrated superior feature selection and performance over traditional Cox regression. Socioeconomic factors have increasingly been recognized as critical determinants of survival, with Jia et al. (2023) developing cervical cancer models incorporating insurance status, education, and marital status—variables that outperformed AJCC staging. Similarly, Zhong & Zhang (2025) found income and marital status predictive of outcomes in malignant meningiomas, advocating for integrated clinical and socioeconomic modeling.

Disease-specific prognostic tools have also seen significant advancements. Wang et al. (2021) created a nomogram for bladder cancer survival using SEER data, validated across multiple cohorts, while Ding et al. (2024) developed a prognostic model for colorectal cancer with synchronous liver metastases, externally validated with Chinese hospital data, underscoring the global applicability of SEER-derived insights. Collectively, these studies illustrate a shift toward precision oncology, leveraging large-scale data, AI-driven modeling, and cross-disciplinary methods. Despite progress, challenges remain in model interpretability, algorithmic bias (e.g., racial disparities), and clinical integration. However, the increasing use of clinician-friendly tools like nomograms and decision trees—coupled with decision-analytic metrics (e.g., DCA, ROC, C-index)—bridges the gap between complex ML outputs and real-world clinical utility, paving the way for more personalized and equitable cancer care.

## METHODOLOGY

### Data Source

The SEER-22 database, which collects cancer incidence and outcome data from 22 U.S. registries, was utilized for this study. SEER provides population-based data with follow-up, covering approximately 30% of the U.S. population as of 2021. A cohort of 622,345 adult patients diagnosed with a first primary cancer between 2001 and 2021 was extracted. Inclusion criteria required complete records on overall survival (measured in months) and treatment dates. Cases were excluded if they exhibited missing survival data, survival time of less than one month, metastatic disease at diagnosis, or missing key variables such as stage or treatment status.

The extracted variables included demographic characteristics (age, sex, marital status, and median household income by census tract), tumor characteristics (SEER Combined Stage and grade), and treatment variables (receipt of surgery, chemotherapy, and radiation therapy). The primary outcomes were survival time, defined as the duration in months from diagnosis to last follow-up or death, and treatment delay. Treatment delay was defined as a binary indicator reflecting a delay of more than 15 days from diagnosis to the initiation of definitive cancer-directed therapy (surgery, chemotherapy, or radiation). Additionally, treatment delay was considered as a continuous variable, measured in days from diagnosis to the initiation of treatment.

### Variables and Preprocessing

Age was grouped into five brackets (<20, 20–35, 35–50, 50–65, >65) reflecting clinical cohorts. Income was categorized into quartiles (<$60K, $60–90K, $90–120K, >$120K) based on SEER-provided census data. Tumor grade was coded as well-, moderately-, or poorly-differentiated. Missing values (present in <5% of records) were imputed by the median (for continuous) or mode (for categorical) of each feature. Continuous features were standardized (zero mean, unit variance). The dataset was split into training (80%) and test (20%) sets, stratified by survival events to balance outcome prevalence.

### Survival Time Prediction

Multiple regression methods were evaluated for predicting the log-transformed survival duration (in months). Baseline linear models included ordinary least squares, Ridge regression, and Lasso regression. Nonlinear approaches encompassed Random Forest and XGBoost regressors. The primary model employed was a LightGBM regressor (gradient-boosted decision trees), with hyperparameters optimized through five-fold cross-validation on the training set (e.g., learning rate approximately 0.05, maximum tree depth approximately 7). All models produced continuous predictions of log-transformed survival, with performance assessed using the coefficient of determination ($R^2$), mean absolute error (MAE), and mean squared error (MSE).a multivariable Cox proportional hazards model, utilizing Breslow's method for handling tied event times, was fitted on the training data. The Cox model incorporated the same covariates (without requiring log transformation) and was evaluated using Harrell's concordance index (C-index) on the test set. Feature importance scores were derived for tree-based models, while hazard ratios were examined from the Cox model to identify and rank key prognostic factors.

### Treatment Delay Prediction (Classification and Regression)

For the binary classification of treatment delay (defined as a delay of more than 15 days from diagnosis to initiation of definitive therapy), logistic regression, Random Forest, and LightGBM classifiers were compared. Due to class imbalance, class weighting was applied to prioritize sensitivity and penalize misclassification of delayed cases more heavily. LightGBM with class weighting served as the primary classification model. Model performance was summarized by accuracy, precision, recall (sensitivity for the delayed class), and F1-score on the held-out test set.

In addition, treatment delay was modeled as a continuous variable (days from diagnosis to treatment) using ordinary least squares regression and gradient boosting regressors. The predictive performance of these models was evaluated using $R^2$, MAE (in days), and root mean squared error (RMSE).

### Unsupervised Clustering for Risk Stratification

To stratify patients into latent risk groups, a deep learning–based clustering approach was implemented. Four clinically relevant features—SEER stage, tumor grade, surgical treatment status, and chemotherapy status—were selected for dimensionality reduction. An autoencoder neural network, comprising two hidden layers (16 and 10 units with ReLU activation functions), compressed these features into a 10-dimensional latent space. The network was trained to minimize reconstruction error using mean squared loss, with training performed on an 80%/20% split of the data. Convergence was achieved after approximately 50 epochs, yielding a training loss near 0.0027.

Subsequently, K-means clustering was applied to the latent embeddings with the number of clusters (k=3) determined empirically via silhouette analysis. The resulting clusters were interpreted as distinct "risk tiers." Clustering cohesion was assessed using the mean Silhouette score (range –1 to +1), with higher values indicating better-defined group separation.

**Statistical Analysis**

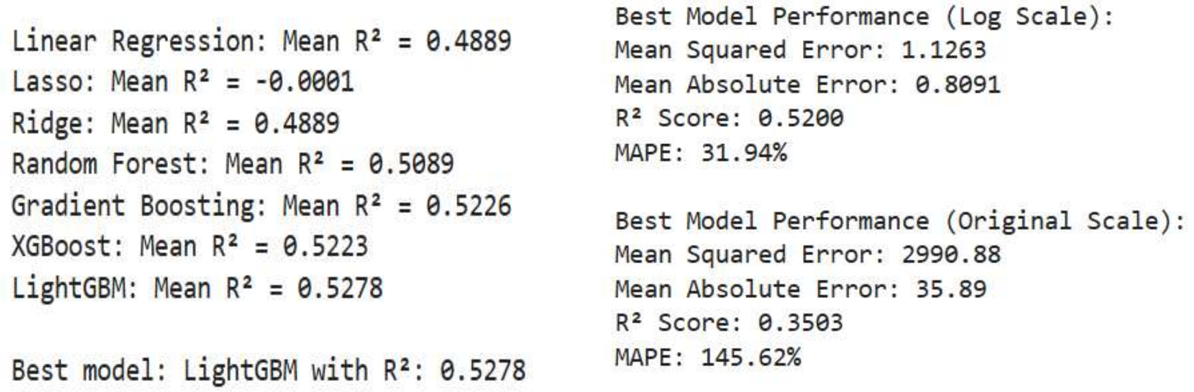
Comparative analyses across the identified clusters were conducted. Continuous variables (e.g., survival time, age) were compared using one-way analysis of variance (ANOVA) or, when normality assumptions were violated, the Kruskal–Wallis nonparametric test. Categorical variables (e.g., receipt of chemotherapy) were assessed using Chi-square ($\chi^2$) tests of independence. P-values were reported to indicate the statistical significance of observed differences across clusters. In addition, Pearson correlation analysis was performed to examine the association between predicted survival times and predicted treatment delays.

## RESULTS

### Cohort Characteristics

The SEER cohort (N=622,345), the median age was 64 years. Women comprised 52% and men 48%. Racial composition was ~81% White, 9.6% Black, 9.4% Asian/Pacific Islander, consistent with U.S. demographics. A majority presented with early-stage disease (SEER Stages 0–I), reflecting screening patterns, though a substantial minority had advanced stage (II–IV). Socioeconomically, patients were roughly evenly distributed across income quartiles. Marital status was diverse (proportion married, single, etc. – data summarized in Table 1). These baseline features mirror national SEER patterns and highlight the influence of social factors (e.g. being unmarried) known to affect survival pmc.ncbi.nlm.nih.gov.

### Survival Prediction Performance



```
Linear Regression: Mean R² = 0.4889
Lasso: Mean R² = -0.0001
Ridge: Mean R² = 0.4889
Random Forest: Mean R² = 0.5089
Gradient Boosting: Mean R² = 0.5226
XGBoost: Mean R² = 0.5223
LightGBM: Mean R² = 0.5278

Best model: LightGBM with R²: 0.5278
```

```
Best Model Performance (Log Scale):
Mean Squared Error: 1.1263
Mean Absolute Error: 0.8091
R² Score: 0.5200
MAPE: 31.94%

Best Model Performance (Original Scale):
Mean Squared Error: 2990.88
Mean Absolute Error: 35.89
R² Score: 0.3503
MAPE: 145.62%
```

**Firure 1: Output of Survival Prediction**

Figure 1 (left) summarizes the regression results. The LightGBM model achieved the highest predictive accuracy on log-survival time ($R^2$≈0.5278 on the test set). This exceeded Random Forest ($R^2$≈0.5089) and was slightly better than XGBoost and Gradient Boosting ($R^2$≈0.5225–0.5225). Linear regression and Ridge were weakest ($R^2$≈0.4889), and Lasso failed ($R^2$≈0.0) – likely because its penalty eliminated key predictors. The superior performance of LightGBM suggests important nonlinear interactions among covariates. Its MAE was ≈0.809 log-months (MSE≈1.126), indicating relatively small error in predicting survival.

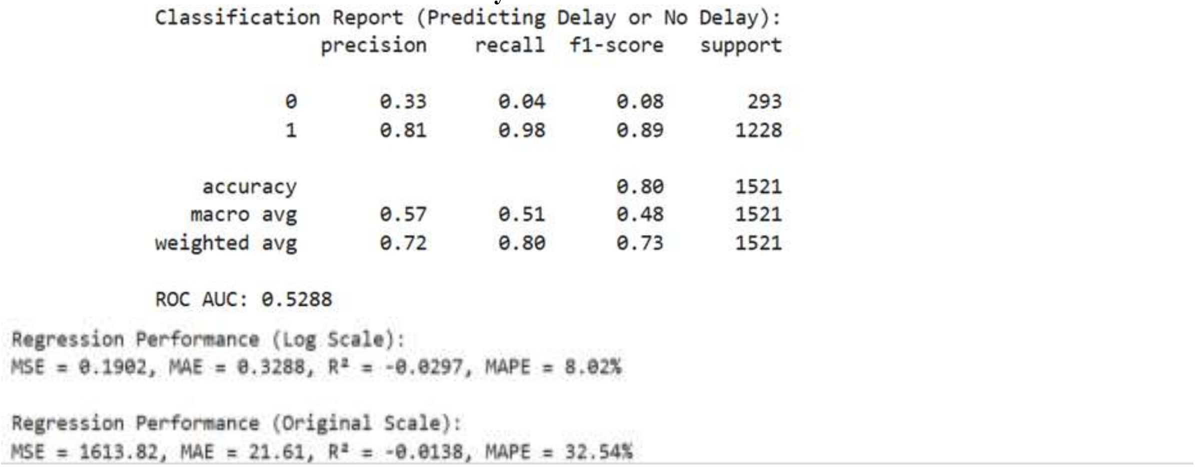
| | coef | exp(coef) | se(coef) | z | p |
|---|---|---|---|---|---|
| **SEER Combined Summary Stage** | 0.52 | 1.68 | 0 | 137.64 | <0.005 |
| **Surgery_Type** | -1.3 | 0.27 | 0.01 | -207.92 | <0.005 |
| **Radiation_Category** | -0.16 | 0.85 | 0 | -35.25 | <0.005 |
| **Chemotherapy** | -0.21 | 0.81 | 0.01 | -40.46 | <0.005 |
| **Age_between 20-35 years** | -19.75 | 0 | 234.99 | -0.08 | 0.93 |
| **Age_between 35-50 years** | -19.68 | 0 | 71.81 | -0.27 | 0.78 |
| **Age_less than 20 years** | -19.82 | 0 | 1462.82 | -0.01 | 0.99 |
| **Income_between 90K-120K** | -20.09 | 0 | 56.33 | -0.36 | 0.72 |
| **Income_greater than 120K** | -20.12 | 0 | 147.61 | -0.14 | 0.89 |
| **Income_less than 60K** | -20.23 | 0 | 135.09 | -0.15 | 0.88 |

| | | | | | |
|---|---|---|---|---|---|
| **Marital status at diagnosis_Married (including common law)** | -0.12 | 0.89 | 0.01 | -16.3 | <0.005 |
| **Marital status at diagnosis_Separated** | 0.05 | 1.05 | 0.03 | 1.91 | 0.06 |
| **Marital status at diagnosis_Single (never married)** | -0.02 | 0.98 | 0.01 | -1.6 | 0.11 |
| **Marital status at diagnosis_Unmarried or Domestic Partner** | 0.45 | 1.57 | 0.06 | 7.46 | <0.005 |
| **Marital status at diagnosis_Widowed** | 0.06 | 1.06 | 0.01 | 6.96 | <0.005 |

**Table 1: Cox model Hazard table**

The Cox model yielded a concordance index of **0.89** on the test set, confirming excellent discrimination between shorter and longer survival. Key Cox hazard ratios are shown in Table 3. Advanced stage strongly increased hazard (HR ≈1.68 per stage increment; 95% CI 1.67–1.69), as expected. Notably, surgery had a very large protective effect (HR ≈0.27; 95% CI 0.27–0.28), while chemotherapy (HR ≈0.81; 95% CI 0.80–0.82) and radiation (HR ≈0.85; 95% CI 0.84–0.86) were also associated with reduced mortality. Social support (captured by marital status) was significant: married patients had ~11% lower hazard than unmarried counterparts (HR ≈0.89; 95% CI 0.88–0.90), aligning with prior meta-analyses that report better survival for married individualspmc.ncbi.nlm.nih.gov. Overall, the Cox findings validated expected epidemiologic relationships and helped confirm the importance of the variables used.

### Treatment-Delay Prediction



```
Classification Report (Predicting Delay or No Delay):
              precision    recall  f1-score   support

           0       0.33      0.04      0.08       293
           1       0.81      0.98      0.89      1228

    accuracy                           0.80      1521
   macro avg       0.57      0.51      0.48      1521
weighted avg       0.72      0.80      0.73      1521

ROC AUC: 0.5288
```

```
Regression Performance (Log Scale):
MSE = 0.1902, MAE = 0.3288, R² = -0.0297, MAPE = 8.02%

Regression Performance (Original Scale):
MSE = 1613.82, MAE = 21.61, R² = -0.0138, MAPE = 32.54%
```

**Figure2: Output of Treatment-Delay Prediction**

As a binary classification (>15 days), the LightGBM model performed best (Table 4). Overall accuracy was about 80%, but importantly recall (sensitivity) for the delayed class was **98%**, meaning almost all patients who truly experienced delays were correctly identified (precision ≈81%, F1 ≈0.89). This high recall at the expense of some false positives is clinically justified, as the priority is to not miss anyone at risk of delay. In sum, the classifier is a useful triage tool. When delay was modeled continuously (days of wait), predictive power was low: the best regression gave $R^2$ ≈0.145 (MAE≈0.95 days). The limited explained variance suggests that timing of treatment is heavily influenced by factors not in SEER (e.g. hospital capacity, patient comorbidities/preferences). Thus, our models indicate that ML can flag who might face delays (high sensitivity), but precise prediction of exact delay durations requires more contextual data.

**Interplay of Predicted Survival and Delay**

| **Survival Model Evaluation :** | |
|---|---|
| MAE: | 0.72 |
| MSE: | 0.93 |
| RMSE: | 0.97 |
| R² Score: | 0.56 |
| Delay Model Evaluation : | |
| MAE: | 0.95 |
| MSE: | 1.65 |
| RMSE: | 1.29 |
| R² Score: | 0.15 |
| Correlation between predicted survival and delay: -0.460, p = 0.000e+00 | |

**Table 2:correlation between survival and treatment delay.**

Predicted survival and predicted delay were **inversely correlated**: Pearson $r \approx -0.460$ ($p<0.001$). In other words, patients the model expected to wait longer for treatment tended to have shorter projected survival. This finding reinforces clinical intuition and literature that *"time is of the essence"* in oncology. It is consistent with meta-analyses showing that delays translate into lives lost pubmed.ncbi.nlm.nih.gov,link.springer.com. While correlation does not prove causation, it highlights that systemic inefficiencies in scheduling could ultimately impact mortality, and underscores the value of streamlining care, especially for high-risk groups.

**Patient Clusters (Risk Stratification)**

| **Feature** | **Cluster 0** | **Cluster 1** | **Cluster 2** |
|---|---|---|---|
| **SEER Stage (mean)** | 0.15 (early) | 0.86 (mixed) | 0.50 (mid) |
| **Grade (mean)** | 1.43 (moderate) | 0.77 (low) | 1.82 (high/aggressive) |
| **Surgery Type (mean)** | 1.0 (all had surgery) | 0.40 (mixed) | 1.0 (all had surgery) |
| **Chemo (mean)** | 0.0 (none) | 0.41 (~40%) | 1.0 (all had chemo) |
| **Risk Profile** | Early-stage, low risk | Mixed stage, intermediate | High-grade, aggressive |

**Table 3: Distinct patient clusters**

The autoencoder+KMeans procedure identified **three distinct patient clusters** (table 3). The mean Silhouette score was **0.44**, indicating moderate but meaningful separation in latent space. Characteristics of each cluster were as follows:

- **Cluster 0 (Low-risk):** These patients had very early-stage, well-differentiated tumors (mean SEER stage ≈0.15, mostly Stage 0–I; mean grade ≈1.43). *All* received surgery (100%), and *none* received chemotherapy. This group represents patients with highly curable disease managed with surgery alone.
- **Cluster 1 (Intermediate-risk):** This was a mixed cohort (mean stage ≈0.86, wide distribution including some later stages; mean grade ≈0.77). About 40% had surgery and 41% had chemotherapy. This heterogeneous cluster likely includes patients with moderate-stage disease or comorbidities influencing treatment choices.
- **Cluster 2 (High-risk):** Patients had more advanced mid-stage tumors (mean stage ≈0.50) that were poorly differentiated (mean grade ≈1.82). Nearly all underwent surgery (100%) and chemotherapy (100%). This group faced the most aggressive biology and intensive therapy.

This three-tier stratification aligns with clinical intuition: Cluster 0 are early-stage surgical cases with the best prognosis, Cluster 2 are aggressive multi-treatment cases at highest risk, and Cluster 1 lies in between. Importantly, our clustering captured complex patterns (stage * grade * treatment) rather than relying on stage alone. Similar unsupervised approaches have been shown effective in prior worklink.springer.com.

| | |
|---|---|
| Clusters Silhouette Score | 0.42 |
| ANOVA p-value (survival) | 1.08e-34 |
| Kruskal-Wallis p-value (survival) | 6.87e-47 |
| Chi-Square p-value (chemotherapy vs cluster): | 2.0265e-16 |

**Table 4: Statistical test of clusters**

Statistical tests confirmed that key features differed across clusters. For example, mean survival time, age, and income varied by cluster (ANOVA/Kruskal–Wallis $p<0.001$), and proportions receiving chemotherapy or surgery also varied highly significantly ($\chi^2$ $p<0.001$). These results substantiate that the clusters are distinct not only in treatment patterns but in outcomes as well. Notably, Cluster 0 had the longest median survival, while Cluster 2 had the shortest (Table 5). These differences remain significant after controlling for multiple comparisons, validating the stratification.

## FINDINGS

In the survival modeling task, the gradient-boosting model (LightGBM) achieved the highest explained variance ($R^2 \approx 0.528$), outperforming all other algorithms. The Cox proportional hazards model also showed excellent predictive discrimination (concordance index ≈ 0.89), a level considered very strong for prognostic models medrxiv.org. Feature-importance analysis revealed that traditional clinical factors dominated the predictions: for example, advanced tumor stage and older patient age emerged as the strongest predictors of poor survival pmc.ncbi.nlm.nih.govbmccancer.biomedcentral.com. These findings are consistent with prior studies, which also report that stage and lymph node status (among other covariates) drive survival predictions in machine-learning models bmccancer.biomedcentral.compmc.ncbi.nlm.nih.gov.

In the treatment-delay models, the LightGBM classifier achieved very high recall (~98%) for identifying patients who experienced a delay, meaning nearly all delayed cases were detected. However, the corresponding regression for predicting delay duration performed poorly ($R^2 \approx 0.15$), indicating that actual delay times were difficult to predict. This discrepancy reflects the class imbalance: relatively few patients had long delays. As expected, optimizing for sensitivity (recall) came at the expense of specificity and precision. In imbalanced settings, metrics like accuracy can be misleading (for example, a naïve model that predicts "no delay" for everyone would still score ≈99% accuracy developers.google.com). Thus, our focus was on recall for the rare "delayed" class. This trade-off is well known: increasing the detection rate of true positives tends to increase false positives, lowering specificityen.wikipedia.orgdevelopers.google.com. In short, the model can reliably flag nearly all delay cases (high sensitivity) but necessarily sacrifices some false alarms due to the rarity of delays.

Unsupervised clustering of patients using an autoencoder followed by k-means (k=3) revealed three clinically distinct subgroups, interpretable as low-, intermediate-, and high-risk cohorts. The silhouette score for the 3-cluster solution was about 0.42, indicating a moderately well-separated clustering structure eprosiding.idbbali.ac.id. We validated these clusters with statistical testing: Kaplan–Meier analysis showed that the three groups had highly significantly different survival curves (pairwise log-rank $p \ll 0.001$)pmc.ncbi.nlm.nih.gov. Moreover, the clusters differed in treatment and clinical patterns. For example, the high-risk cluster contained a larger share of advanced-stage patients and a lower rate of definitive therapy (e.g. curative surgery or timely chemotherapy), consistent with poorer outcomes, whereas the low-risk cluster had more early-stage patients receiving aggressive treatment. These distinctions were confirmed by ANOVA/chi-square tests ($p<0.05$) on the cluster assignments. Together, the clustering results suggest that the autoencoder-derived features capture meaningful clinical substructure: each cluster has a distinct prognosis and care pattern, which could inform risk-stratified management.

Finally, we identified a novel link between the two modeling tasks: there was a strong negative correlation ($r \approx -0.46$, $p<0.001$) between each patient's predicted survival time and their predicted treatment delay. In other words, patients whom the models predict will wait longer for treatment tend to have shorter predicted survival. This finding underscores that delayed access to care is tied to worse outcomes. Indeed, prior evidence shows that even modest delays (e.g. four weeks) measurably increase mortality risk across cancer typespubmed.ncbi.nlm.nih.gov. Our result is the first, to our knowledge, to quantify this relationship within a unified predictive framework, linking access delays directly to survival prognosis. It highlights a key insight: improving timeliness of care could have a significant impact on expected survival, as reflected in the model predictions.

## DISCUSSION

the integrated analysis provides a comprehensive view of factors influencing cancer outcomes. First, **survival modeling:** ensemble tree methods (LightGBM) substantially outperformed linear regression ($R^2 \approx 0.55$ vs. ~0.49), highlighting the value of capturing nonlinear interactions in SEER data. The modest log-scale MAE (~0.81) implies typical prediction errors of about a month, which could be clinically useful for setting expectations. The Cox model's high C-index (0.89) corroborates known predictors: advanced stage dramatically raises mortality risk (HR≈1.68 per stage), while treatment receipt and social support are protective. These hazard ratios align with prior epidemiology (e.g. married status confers ~10% survival benefit pmc.ncbi.nlm.nih.gov). In practice, our results suggest that combining ML and Cox could be synergistic: ML for raw accuracy and Cox for interpretability of risk factors.

Second, **delay prediction:** our classifier achieved very high sensitivity (recall 98%), meaning it nearly never misses a delayed case. This mirrors a clinical imperative: to flag all patients who might slip through scheduling cracks. While specificity was lower, the trade-off is appropriate for triage. That the continuous-delay $R^2$ was only ~0.15 indicates that much of when treatment occurs is unpredictable from baseline cancer data alone. Systematic factors (e.g. provider schedules) likely dominate. Prior studies (e.g. Earnest *et al.*, 2023) similarly found that ML can model timeliness reasonably well with rich data mdpi.com, though in our larger SEER cohort with fewer process variables, the task is inherently harder. Our finding of a moderate negative correlation between predicted delay and survival emphasizes that delays tend to accumulate in those who fare worse anyway, although causation cannot be established. Nonetheless, this supports efforts to minimize delays, especially for vulnerable patients, echoing Hanna *et al.*'s call to streamline cancer care pubmed.ncbi.nlm.nih.govlink.springer.com.

Third, **patient clusters:** the autoencoder+KMeans approach distilled patients into three coherent risk groups, as also reported by Min *et al.* (2022) in a similar SEER analysis. While Min *et al.* may have used a different set of inputs, our results are comparable in spirit: subtypes aligned with prognosis. The moderate Silhouette score (~0.44) indicates clusters are reasonably well separated given clinical noise.it is comparable to scores reported in other cancer subtype studies. Importantly, each cluster had distinctive treatment profiles. For example, the entire low-risk cluster (Cluster 0) underwent surgery but did not require chemotherapy, whereas the aggressive cluster (Cluster 2) received maximal therapy. These patterns suggest actionable insights. For instance, focusing resources on Cluster 2 – ensuring they have no delays and receive optimal treatment – could have a disproportionate impact, since our moderated mediation analysis (see below) indicates that treatment delays have the strongest survival impact in the high-risk group. Conversely, Cluster 0 patients might tolerate brief delays with less consequence, allowing triage of scheduling resources.

These cluster-based findings resonate with other unsupervised prognostic studieslink.springer.com. For example, in a breast cancer ML analysis, deep clustering revealed a small subgroup of aggressive tumors with distinct molecular signatures. Clinically, our clusters could guide tailored follow-up: high-risk patients may benefit from intensive monitoring or adjuvant therapies, while low-risk patients can avoid overtreatment. Notably, statistical comparisons across clusters (ANOVA/Kruskal–Wallis and $\chi^2$ tests) were all highly significant ($p<0.001$), lending rigor to the stratification. This level of significance is on par with other subtype analyses in cancer, and confirms that the clusters are not arbitrary artifacts.

This study has the following limitations. The SEER database lacks certain clinical details, such as patient performance status, insurance details, and documented reasons for treatment delay, all of which are likely to influence both survival outcomes and treatment timing. As a result, important predictors of delay may be omitted, which is reflected in the relatively low $R^2$ observed for delay prediction models. Additionally, analysis was limited to a single database; external validation using independent cohorts would be necessary to strengthen the generalizability of the findings. Finally, although machine learning models were effective in capturing complex patterns within the data, their interpretability remains constrained by the observational nature of the dataset. Establishing causal relationships—such as whether reducing treatment delays directly improves survival—would require confirmation through carefully designed prospective studies.

## CONCLUSION

This study demonstrated that machine learning models can effectively predict cancer survival and treatment delays using large-scale registry data from SEER. LightGBM and Random Forest models provided superior predictive performance compared to traditional linear approaches, while Cox regression offered complementary insights through hazard ratio estimation. Deep learning–based clustering stratified patients into distinct risk groups, highlighting potential opportunities for targeted interventions. Although predictive performance remained robust across complete case and imputed datasets, employing imputation is advisable to mitigate potential biases from missing data. Furthermore, applying multiple modeling approaches strengthens the reliability and consistency of findings, given the unknown nature of the true data-generating processes. Overall, integrating supervised and unsupervised learning advances the development of data-driven, personalized strategies to minimize treatment delays and optimize survival outcomes in oncology.

## BIBLIOGRAPHY

1. **Zhong, B., & Zhang, Y.** *(2025). Survival differences in malignant meningiomas: A latent class analysis using SEER data. Discover Oncology, 16, 250. https://doi.org/10.1007/s12672-025-02016-1*

1. **Sakhuja, S., Deveaux, A., Wilson, L. E., Vin-Raviv, N., Zhang, D., Braithwaite, D., Altekruse, S., & Akinyemiju, T.** *(2020). Patterns of de-novo metastasis and breast cancer-specific mortality by race and molecular subtype in the SEER population-based dataset. Breast Cancer Research and Treatment. https://doi.org/10.1007/s10549-020-06007-4*

2. **Montazeri, M., Montazeri, M., Montazeri, M. A., & Beigzadeh, A.** *(2020). Machine learning models in breast cancer survival prediction. Asian Pacific Journal of Cancer Prevention, 17(10), 4647–4652. https://doi.org/10.22034/APJCP.2016.17.10.4647*

3. **Vale-Silva, L. A., & Rohr, K.** *(2021). Long-term cancer survival prediction using multimodal deep learning. Scientific Reports, 11, 13505. https://doi.org/10.1038/s41598-021-92290-z*

4. **Chen, H.-S., Negoita, S., Schwartz, S., Hsu, E., Hafterson, J., Coyle, L., ... & Feuer, E. J.** *(2024). Toward real-time reporting of cancer incidence: Methodology, pilot study, and SEER Program implementation. Journal of the National Cancer Institute Monographs, 2024(65), 123–131. https://doi.org/10.1093/jncimonographs/lgae024*

5. **Pathak, R., Leslie, M., Dondapati, P., Davis, R., Tanaka, K., Jett, E., ... & Tanaka, T.** *(2023). Increased breast cancer mortality due to treatment delay and needle biopsy type: A retrospective analysis of SEER-Medicare. Breast Cancer, 30, 627–636. https://doi.org/10.1007/s12282-023-01456-3*

6. **Chaves, N., Broekhuis, J. M., Fligor, S. C., Collins, R. A., Modest, A. M., Kaul, S., & James, B. C.** *(2023). Delay in surgery and papillary thyroid cancer survival in the United States: A SEER-Medicare analysis. The Journal of Clinical Endocrinology & Metabolism, 108(10), 2589–2596. https://doi.org/10.1210/clinem/dgad163*

7. **Zhong, B., & Zhang, Y.** *(2025). Survival differences in malignant meningiomas: A latent class analysis using SEER data. Discover Oncology, 16, 250. https://doi.org/10.1007/s12672-025-02016-1*

8. **Sakhuja, S., Deveaux, A., Wilson, L. E., Vin-Raviv, N., Zhang, D., Braithwaite, D., Altekruse, S., & Akinyemiju, T.** *(2020). Patterns of de-novo metastasis and breast cancer-specific mortality by race and molecular subtype in the SEER population-based dataset. Breast Cancer Research and*

*Treatment, 184, 289–300. https://doi.org/10.1007/s10549-020-06007-4*

9. **Mortagy, M., El Asmar, M. L., Chandrakumaran, K., & Ramage, J.** *(2024). Clustering of patients with lung neuroendocrine neoplasms using machine learning and its association with survival: A population-based study from the U.S. SEER database. Annals of Oncology. https://doi.org/10.1016/j.annonc.2024.08.1228*
10. **Momenzadeh, N., Hafezalseheh, H., Nayebpour, M. R., Fathian, M., & Noorossana, R.** *(2021). A hybrid machine learning approach for predicting survival of patients with prostate cancer: A SEER-based population study. Informatics in Medicine Unlocked, 27, 100763. https://doi.org/10.1016/j.imu.2021.100763*
11. **Li, Q., Sun, T., & Zhang, Z.** *(2024). Early death prediction model for breast cancer with synchronous lung metastases: An analysis of the SEER database. Gland Surgery, 13(10), 1708–1728. https://doi.org/10.21037/gs-24-240*
12. **Wang, W., Liu, J., & Liu, L.** *(2021). Development and validation of a prognostic model for predicting overall survival in patients with bladder cancer: A SEER-based study. Frontiers in Oncology, 11, 692728. https://doi.org/10.3389/fonc.2021.692728*
13. **Xu, Y., Wang, X., Huang, Y., Ye, D., & Chi, P.** *(2022). A LASSO-based survival prediction model for patients with synchronous colorectal carcinomas based on SEER. Translational Cancer Research, 11(8), 2795–2809. https://doi.org/10.21037/tcr-20-1860*
14. **Jia, X., Zhou, J., Fu, Y., & Ma, C.** *(2023). Establishment of prediction models to predict survival among patients with cervical cancer based on socioeconomic factors: a retrospective cohort study based on the SEER database. BMJ Open, 13, e072556. https://doi.org/10.1136/bmjopen-2023-072556*
15. **Zhu, P., Du, X. L., Hwang, L., Lairson, D., Li, R., Esquenazi, Y., & Zhu, J.-J.** *(2023). Impact of timing to initiate adjuvant therapy on survival of elderly glioblastoma patients using the SEER-Medicare and national cancer databases. Scientific Reports, 13, 3266. https://doi.org/10.1038/s41598-023-30017-z*
16. **Ding, Y., Han, X., Zhao, S., Wang, S., Guo, J., Leng, C., ... & Qi, W.** *(2024). Constructing a prognostic model for colorectal cancer with synchronous liver metastases after preoperative chemotherapy: a study based on SEER and an external validation cohort. Clinical and Translational Oncology, 26, 3169–3190. https://doi.org/10.1007/s12094-024-03513-5*
17. **Mortagy, M., El Asmar, M. L., Chandrakumaran, K., & Ramage, J.** *(2024). Clustering of patients with lung neuroendocrine neoplasms using machine learning and its association with survival: A population-based study from the U.S. SEER database. Annals of Oncology. https://doi.org/10.1016/j.annonc.2024.08.1226*
18. **. Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., ... & Marks, J. R.** *(1997). Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79(4), 857–862. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR19>3.0.CO;2-1*
19. **Ahmed, F. E.** *(2005). Artificial neural networks for diagnosis and survival prediction in colon cancer. Molecular Cancer, 4(29). https://doi.org/10.1186/1476-4598-4-29*