Minimax Theory and its Applications
Volume 11 (2026), No. 3s, 44 to 53

Improving Prognostic Accuracy in Non-Metastatic Gastric Cancer: A Comparative Analysis of Statistical and Machine Learning Survival Models in SEER Cohorts

Mohamed Kalith Oli M[1], Jafar Ali Ibrahim S[2]

[1]School of Bio Sciences & Technology, Vellore Institute of Technology,Vellore, India. Email: mohamedkalith.olim2023@vitstudent.ac.in
[2]Department of QuantumAI, School of Computer Science and Engineering, Vellore Institute of Technology,Vellore, India. Email: jafarali.s@vit.ac.in

Abstract
Gastric cancer (GC) remains a leading cause of global cancer mortality. Accurate prognosis in non-metastatic disease remains challenging when treating with curative intent. This is due to the survival heterogeneity that exists within tumor-node-metastasis (TNM) stages. In this study, six survival models were compared in 20,132 patients using 10-fold cross-validation. Models including Cox proportional hazards, Weibull accelerated failure time, survival support vector machine, random survival forest (RSF), gradient boosting survival analysis (GBSA), and DeepSurv were used. Performance was evaluated using C-index, integrated Brier score (IBS), calibration, decision curve analysis (DCA), and net reclassification improvement (NRI). Over a median 41-month follow-up, 13,272 deaths occurred. Similar discrimination was found among top models; DeepSurv (C-index 0.7542), GBSA (0.7540), and RSF (0.7532) higher than Cox (0.7417). In terms of clinical utility, RSF got the lowest prediction error (IBS 0.1516), followed by GBSA (0.1534). However, DeepSurv indicated poor calibration and limited net benefit at 12 months. NRI analysis indicated that ensemble models (RSF, GBSA) consistently improved reclassification over Cox, whereas DeepSurv exhibited a decline at 12 months. Age, histology, and nodal burden were the primary predictors. Ensemble models, particularly RSF, offer a balance of both accuracy and clinical utility for risk stratification in non-metastatic GC.

Keywords: Gastric cancer; Survival analysis; Machine learning; Random survival forest; SEER database..

Introduction

Gastric cancer (GC) is one of the most common aggressive cancers, ranking fifth in both incidence and mortality worldwide(Bray et al., 2024). East Asia, Eastern Europe, and South America are regions where GC is prevalent due to dietary, environmental, and genetic factors(Sundar et al., 2025). In the last few decades, the incidence of GC has dropped due to the lowered prevalence of H. pylori, dietary changes, decreased smoking, and socioeconomic improvements(Balakrishnan, George, Sharma, & Graham, 2017; Lee et al., 2016; Morgan et al., 2022). Prognostic assessment for patients with non-metastatic GC remains challenging because outcomes vary beyond the TNM stage(Lordick et al., 2022). Patients within the same TNM category might experience different clinical outcomes, highlighting the limits of stage-based prognostic prediction and the need for personalized risk assessment. This is a major clinical problem in non-metastatic disease, where many patients are treated with curative intent(Chen et al., 2022; Mahar et al., 2018). Accurate risk estimates can help to personalize follow-up (surveillance), guide perioperative/adjuvant therapy decisions, and facilitate clinical trial selection(Chen et al., 2022). Prior SEER-based analyses of non-metastatic gastric cancer after gastrectomy have shown that age has a complex association with outcomes, with younger patients exhibiting poorer tumor characteristics yet favorable survival in the non-metastatic subgroup, supporting the need for individualized prognostic modeling(Zeng, Li, Cao, & Zheng, 2023; J. Zhang et al., 2018). Conventional prognostic tools such as TNM classification and Cox proportional hazards regression are widely used. However, Cox regression assumes log-linear covariate effects and proportional hazards over time. Violations of these assumptions can reduce model validity and calibration in applied clinical datasets(Austin, Fang, & Lee, 2022; J. Zhang et al., 2018).

Residual-based diagnostics and time-varying effect checks are used to test these assumptions. Population-based registries, like the Surveillance, Epidemiology, and End Results (SEER) program, provide data for the development and comparison of prognostic models using clinicopathologic variables on a large scale(Duggan,

Anderson, Altekruse, Penberthy, & Sherman, 2016). Recently, machine learning (ML) survival algorithms have been used on datasets like SEER to enhance prognostic prediction in gastric oncology. For example, Zeng et al. developed postoperative survival models that compared Cox, RSF, and DeepSurv. They found that DeepSurv performed well in discrimination, DCA and calibration(Zeng, Song, Li, Cao, & Zheng, 2024). Gong et al. analyzed multiple survival models with locally advanced gastric cancer patients and found that ensemble-based methods resulted in better outcomes supported by DCA(Gong et al., 2025). These studies shows that ensemble and deep learning approaches are increasingly being applied to predict gastric cancer survival outcomes. Also, it suggests that flexible, non-linear survival models may provide additional prognostic value than usual regression in population-based oncology research. However, most comparison studies include only a limited number of models or focus mainly on discrimination without multi-dimensional evaluation. Assessment of survival prediction models should include discrimination, calibration, and clinical usefulness rather than discrimination alone(McLernon et al., 2023). Comprehensive benchmarking of conventional statistical models along diverse ML survival methods within large non-metastatic gastric cancer cohorts remains limited. In this study, six survival models including Cox regression, parametric, kernel-based, ensemble, and neural network were compared across discrimination, integrated prediction error, calibration at clinically relevant time points, reclassification metrics, and decision curve analysis in a non-metastatic gastric cancer cohort using stratified 10-fold cross-validation. For interpretability, permutation importance and SHAP were used to address variable importance and to show how key features contributed to predicted risk.

Methods

Study Design and Data Source

In this retrospective study, the clinical data were retrived from the Surveillance, Epidemiology and End Results program (SEER Research Data, 17 Registries, Nov 2024 Submission (2000-2022) using SEER*Stat software (version 9.0.42.2)(Institute, 2018). The cohort included patients diagnosed between 2010 and 2015 with gastric cancer defined by ICD-O-3 primary site codes C16.0-C16.9, malignant behavior, and known age. The extracted variables included: (1) demographic characteristics, including age at diagnosis and sex; (2) tumor characteristics, including historic stage, AJCC T and N categories, tumor size, and histology; (3) nodal characteristics, including the number of regional lymph nodes examined and positive regional lymph nodes; (4) treatment-related variables, including primary-site surgery, radiotherapy, and chemotherapy; and (5) outcome data, including overall survival time and vital status.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: 1) Histologically confirmed gastric carcinoma; 2) Diagnosed with M0 disease (non-metastatic) at diagnosis; 3) Available overall survival time and vital status. Non-metastatic disease (M0) were defined using Summary Stage 2000 (1998–2017) by restricting cases to localized and regional stage at diagnosis (distant stage excluded). Exclusion criteria included: 1) Missing or zero follow-up time; 2) Inconsistent records after cleaning.

Data preprocessing

Data for each non-metastatic gastric cancer patient, including age, sex, tumor details, lymph node counts, histology were preprocessed for modeling. Histology codes were converted into textual descriptions and categorized into three groups: intestinal, diffuse, and other. The lymph node ratio (LNR) were continuous and calculated as the ratio of positive to examined lymph nodes. When the number of examined nodes were zero or missing, LNR were set to missing and either imputed within-fold or explicitly flagged using a binary indicator variable (“LNR_missing”). Age at diagnosis and tumor size were both considered continuous variables. Treatment-related variables for primary-site surgery, radiotherapy, and chemotherapy were included as binary predictors. In multivariable modeling, missing numeric values were imputed by the median values, and categorical variables were converted using one-hot encoding. Imputation and one-hot encoding were performed within each cross-validation fold using the training set only to prevent data leakage.

Construction of Prognostic Models

Six survival models, including statistical and machine learning methods, were used in a stratified 10-fold cross-validation to support model development and evaluation. The same demographic, clinicopathological, nodal, and treatment-related predictors, including primary-site surgery, radiotherapy, and chemotherapy, were used across all six prognostic models.

- Semi-parametric (Baseline): A penalized Cox Proportional Hazards (CoxPH) model was fitted using an L2 penalizer of 0.1 to improve stability and to avoid overfitting(Cox, 1972).
- Parametric: A Weibull Accelerated Failure Time (AFT) model were utilized with 0.1 penalization(Wei, 1992).
- Linear machine learning: A ranking-based survival support vector machine were trained with rank_ratio = 1.0, max_iter = 2000, tol = 1×10^{-5}, and random_state = 42(Van Belle, Pelckmans, Van Huffel, & Suykens, 2011).
- Tree-based ensemble (bagging): The random survival forest (RSF) was trained with 400 trees, min_samples_split = 10, min_samples_leaf = 10, max_features = $\sqrt{p}$, n_jobs = 1, and random_state = 42(Ishwaran, Kogalur, Blackstone, & Lauer, 2008).
- Tree-based ensemble (boosting): Gradient boosting survival analysis (GBSA) were trained with 400 estimators, learning_rate = 0.05, max_depth = 3, subsample = 1.0, and random_state = 42(Hothorn, 2005).
- Neural network: DeepSurv were implemented using a multilayer perceptron with two hidden layers (64 and 32 units), batch normalization, and dropout = 0.0, optimized with Adam (learning rate 1×10^{-3}) for 100 epochs with batch size = 256 (random seed = 42(Katzman et al., 2018).

Event status was used to stratify cross-validation. In addition to the usual CoxPH baseline, the top-performing models based on Harrell's C-index were chosen for additional calibration analysis and clinical utility tests.

Model Validation and Statistical Analysis

Stratified 10-fold cross-validation based on event status were performed to calculate out-of-fold predictions, which is used to evaluate the predictive performance of the models. Preprocessing, like median imputation for continuous variables and one-hot encoding for categorical features, were performed within each fold on the training set to avoid data leakage.

For discrimination, Harrell's concordance index (C-index) was calculated for each fold and averaged over all folds(McLernon et al., 2023). The 95% confidence intervals for the mean C-index were computed across the 10 folds as mean ± 1.96 × (SD/$\sqrt{10}$). The IBS were calculated to evaluate prediction error throughout a 5-year follow-up period(Graf, Schmoor, Sauerbrei, & Schumacher, 1999). For graphical calibration, patients were grouped into deciles according to predicted event risk at 12, 36, and 60 months. Within each decile, the mean predicted event probability was compared with the corresponding observed Kaplan–Meier event probability. Clinical utility was determined using DCA by plotting the net benefit against the threshold probability (0.01-0.50), and NRI were calculated in a time-dependent manner using risk categorization based on tertiles, with 95% confidence intervals from 1,000 bootstrap resamples(Pencina, D'Agostino, & Steyerberg, 2011; Vickers & Elkin, 2006). Analyses were implemented in Python using scikit-survival, lifelines, and pycox.

Model Interpretability

Permutation importance and SHAP (SHapley Additive exPlanation) were used to analyze interpretability. By measuring the decrease in C-index after randomly shuffling each feature, permutation importance were calculated, which helps to identify the features impact on models. SHAP values were used to quantify how each feature influenced individual predictions. Additionally, KernelExplainer were applied using 36-month predicted death risk ($1 - S(36)$) as the target, and global summary plots and individual explanations were generated to visualize non-linear relationships and feature interactions

Results

Patient Cohort and Baseline Characteristics

After applying the eligibility criteria, a total of 20,132 patients with non-metastatic gastric cancer were included in the final cohort. Baseline demographic and clinicopathologic characteristics are summarized in Table 1. There were 12,085 men (60.0%) and 8,047 women (40.0%). Within the cohort, 32.9% of patients were aged ≥ 75 years, and 9.7% were < 50 years old. 53.3% of the tumors were localized, while 46.7% were regional. Intestinal-type adenocarcinoma is the common histological subtype (57.1%), followed by other histologies (23.6%) and diffuse-type tumors (19.2%). During a median follow-up of 41 months, 13,272 deaths occurred (65.9%).

Table 1: Baseline demographic and clinicopathologic characteristics of the cohort (n = 20,132). Treatment variables were represented as binary indicators (1 and 0).

Variable	Entire Cohort

Age (years)	67.00 (57.00-77.00) | 67.54 ± 13.41
Tumor size (mm)	36.00 (20.00-60.00) | 44.80 ± 43.97
Examined lymph nodes	3.00 (0.00-17.00) | 10.64 ± 15.84
Positive lymph nodes	0.00 (0.00-4.00) | 4.87 ± 13.69
Lymph node ratio	0.00 (0.00-0.29) | 0.19 ± 0.29
Sex: Male	12085 (60.0%)
Sex: Female	8047 (40.0%)
SEER stage: Localized	10727 (53.3%)
SEER stage: Regional	9405 (46.7%)
AJCC T stage: T3	6188 (30.7%)
AJCC T stage: T2	3272 (16.3%)
AJCC T stage: T1A	2191 (10.9%)
AJCC T stage: T4A	1855 (9.2%)
AJCC T stage: T1B	1757 (8.7%)
AJCC T stage: T1NOS	1457 (7.2%)
AJCC T stage: T1	1127 (5.6%)
AJCC T stage: T4B	830 (4.1%)
AJCC T stage: TX	813 (4.0%)
AJCC T stage: T4	451 (2.2%)
AJCC T stage: TIS	189 (0.9%)
AJCC T stage: T0	2 (0.0%)
AJCC N stage: N0	12117 (60.2%)
AJCC N stage: N1	3979 (19.8%)
AJCC N stage: N2	1888 (9.4%)
AJCC N stage: N3A	999 (5.0%)
AJCC N stage: N3B	469 (2.3%)
AJCC N stage: N3	369 (1.8%)
AJCC N stage: NX	311 (1.5%)
Histologic subtype: Intestinal	11500 (57.1%)
Histologic subtype: Other	4759 (23.6%)
Histologic subtype: Diffuse	3873 (19.2%)
Primary-site surgery: 1	14,535 (72.2%)
Primary-site surgery: 0	5,597 (27.8%)
Chemotherapy: 1	9,062 (45.0%)
Chemotherapy: 0	11,070 (55.0%)
Radiotherapy: 1	5,483 (27.2%)
Radiotherapy: 0	14,649 (72.8%)

Model Discrimination Performance

In stratified 10-fold cross-validation, machine learning models achieved higher discrimination than the statistical approaches in this cohort. DeepSurv (mean C-index 0.7542), GBSA (0.7540), and RSF (0.7532) showed comparable discrimination and exceeded Cox regression (0.7417). The Weibull AFT and survival SVM achieved C-indices of 0.7414 and 0.7427, respectively (Table -2). Based on these results, DeepSurv, GBSA, and RSF were carried forward for further analyses. Also, the CoxPH model were retained as the reference model for comparison.

Table 2. Model discrimination performance across stratified 10-fold cross-validation. Values are mean C-index across folds with 95% confidence intervals (mean ± 1.96 × SD/√10).

Model	**Mean C-index**	**SD**	**95% CI**
DeepSurv	0.7542	0.0075	0.7496 ± 0.7588
GBSA	0.754	0.0061	0.7503 ± 0.7578
RSF	0.7532	0.0067	0.7490 ± 0.7573
SVM	0.7427	0.0081	0.7376 ± 0.7477
Cox	0.7417	0.0081	0.7367 ± 0.7468
Weibull	0.7414	0.0082	0.7363 ± 0.7464

Calibration and Prediction Error

Calibration plots comparing predicted survival probabilities with observed Kaplan–Meier estimates were assessed at 12, 36, and 60 months as shown in Figure 1. Most models were well calibrated, except for DeepSurv at 12 months, which showed underprediction in the lowest-risk decile (very low predicted risk despite observed events). At 36 and 60 months, calibration were generally good across models, suggesting stable longer-horizon risk estimation. The integrated Brier score (IBS) were used to calculate prediction error over the 60-month horizon. Tree-based models showed the lowest prediction error, with RSF achieving the lowest IBS (0.1516), followed by GBSA (0.1534). DeepSurv showed a higher IBS (0.1644), and Cox had an IBS of 0.1597. These results were consistent with the discrimination ranking observed for the ensemble models.

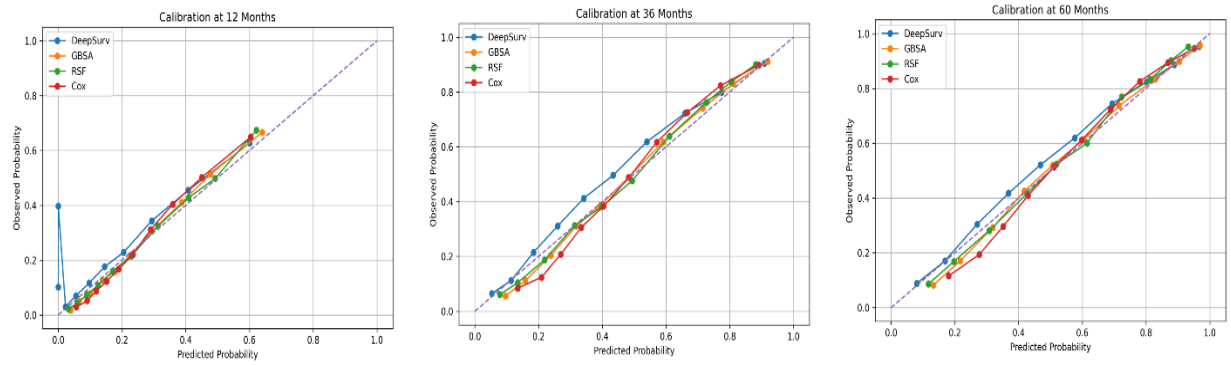


Fig 1. Calibration plots comparing predicted and observed event probabilities at (A) 12 months, (B) 36 months, and (C) 60 months for DeepSurv (blue), GBSA (orange), RSF (green) and Cox (red). DeepSurv shows underprediction at 12 months in the lowest-risk range, where calibration is generally similar across models at 36 and 60 months.

Risk Stratification and Clinical Utility

Kaplan-Meier curves stratified by predicted risk showed clear separation as shown in figure 2. Risk groups were defined using a median split of the 36-month predicted death risk, yielding low-risk and high-risk groups. In time-dependent reclassification analyses, NRI were computed using tertile-based risk categories (33rd and 66th percentile cutpoints) at 12, 36, and 60 months.

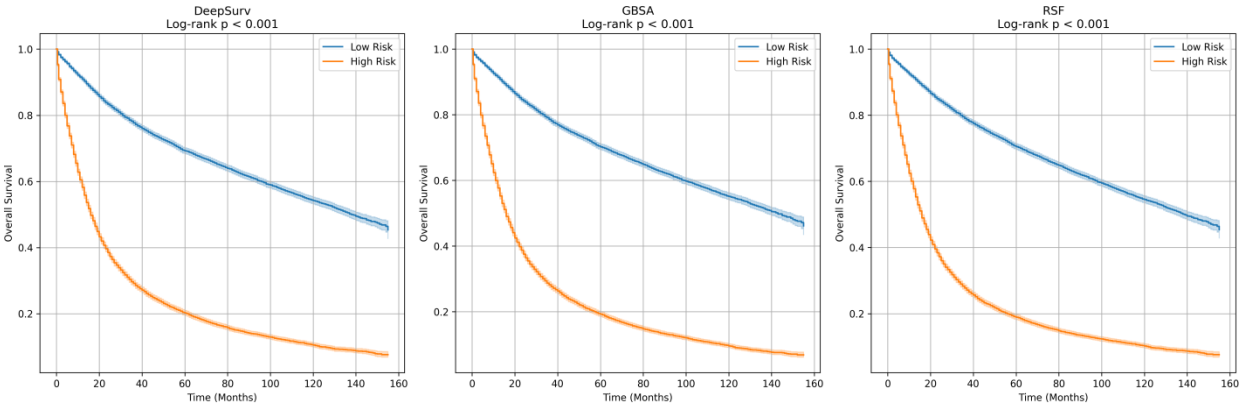


Fig 2. Kaplan-Meier overall survival curves for low-risk and high-risk groups defined by a median split of the 36-month predicted death risk for (A) DeepSurv, (B) GBSA, and (C) RSF. Shaded bands indicate uncertainty around the survival estimates.

Time-dependent NRI analyses showed that the ensemble models improved risk reclassification compared with Cox regression (Table 3).

Table 3. Time-dependent net reclassification improvement (NRI) versus CoxPH at 12, 36, and 60 months. NRI were computed using tertile-based risk categories (33rd and 66th percentile cutpoints). Values are mean NRI with 95% bootstrap confidence intervals (1,000 resamples).

Timepoint	Comparison	Mean_NRI	CI_2.5%	CI_97.5%
12	RSF vs Cox	0.0482	0.0349	0.0614

12	DeepSurv vs Cox	-0.1746	-0.1941	-0.1545
12	GBSA vs Cox	0.0308	0.0200	0.0417
36	RSF vs Cox	0.0399	0.0280	0.0524
36	DeepSurv vs Cox	-0.0053	-0.0168	0.0068
36	GBSA vs Cox	0.0289	0.0192	0.0385
60	RSF vs Cox	0.0306	0.0195	0.0417
60	DeepSurv vs Cox	0.0128	0.0015	0.0245
60	GBSA vs Cox	0.0307	0.0209	0.0407

RSF had the largest improvement, particularly at 12 months (NRI = 0.0482) and 36 months (NRI = 0.0399), followed by GBSA (36-month NRI = 0.0289). DeepSurv showed less consistent reclassification, including a negative NRI at 12 months (-0.1746). Decision curve analysis showed greater net benefit for the tree-based ensemble models (RSF and GBSA) across threshold probabilities from 0.01 to 0.50 as shown in Figure 3.

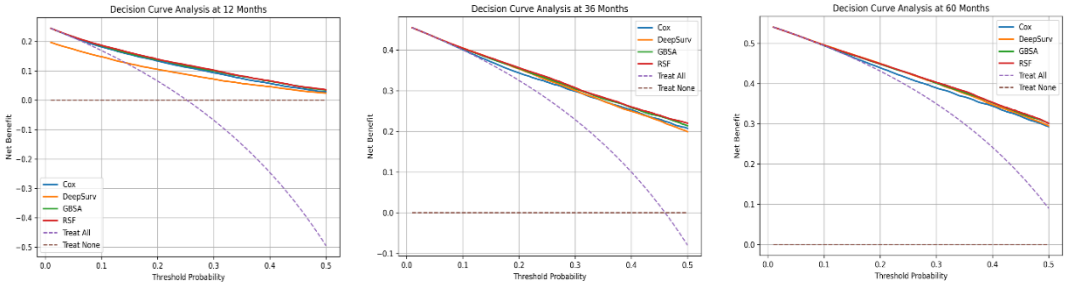


Fig 3. Decision curve analysis showing net benefit across threshold probabilities (0.01–0.50) at (A) 12 months, (B) 36 months, and (C) 60 months for DeepSurv, GBSA, RSF, Cox. The red dashed line indicates the "treat-all" strategy and the purple dashed line indicate the "treat-none" strategy. At 12 months, DeepSurv shows reduced net benefit relative to the ensemble models across portions of the threshold range.

Feature Importance

Age, histologic subtype, and nodal characteristics (such as nodal burden and lymph node ratio) were found to be the significant predictors of outcome based on permutation relevance shown in figure 4.

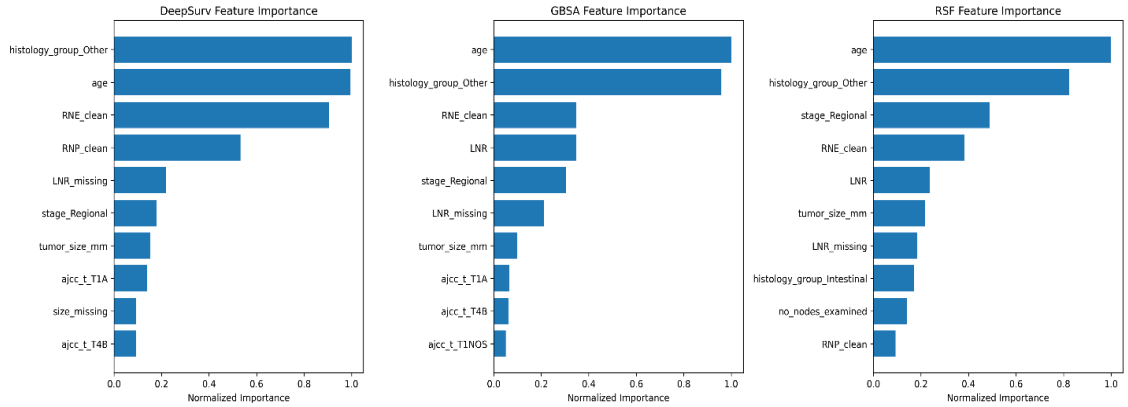


Fig 4. Permutation feature importance for DeepSurv, GBSA, and RSF, quantified as the decrease in C-index after permuting each feature. Importance values are normalized within each model (x-axis). Higher values indicate greater contribution to predictive performance.

These results were supported by SHAP analysis of the RSF model, which also revealed non-linear correlations between important predictors and individual risk assessments represented in figure 5.

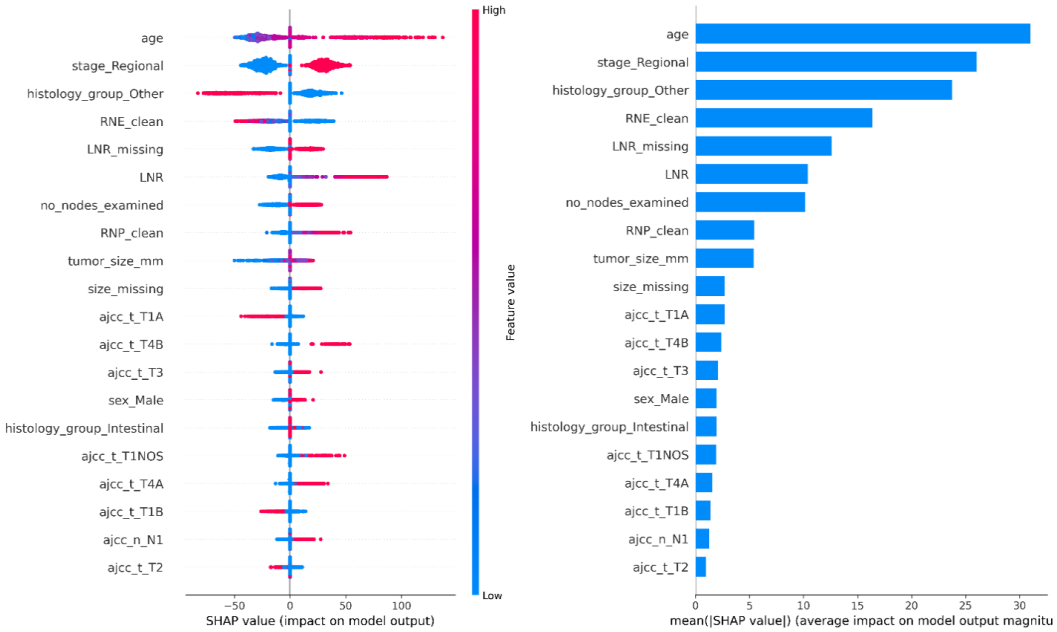


Fig 5A). SHAP beeswarm plot showing the distribution and direction of feature effects on predicted risk, where each point represents an individual and color denotes feature value (low to high). (B) Bar plot of mean absolute SHAP values showing global feature importance. Positive SHAP values indicate increased predicted risk, whereas negative values indicate decreased predicted risk.

Discussion

In this large SEER-based cohort of 20,132 patients with non-metastatic gastric cancer, we compared six survival models with a unified stratified 10-fold cross-validation and multiple evaluation techniques. With a median follow-up of 41 months, 65.9% of patients died (13,272 out of 20,132). This shows how important it is to accurately classify prognostic factors in M0 cases. Metastatic gastric cancer has been widely studied, but making predictions about non-metastatic (potentially curable) disease is harder because there are variations within same TNM stage, likely reflecting differences in tumor biology, nodal burden, and inherent disease aggressiveness(Zubarayev, Min, & Son, 2019).Personalized risk assessments may help for individualized surveillance, perioperative and adjuvant treatment decisions and enhancing the selection of higher-risk patients for clinical studies (Lordick et al., 2022). Similar survival modeling frameworks have been reported in other malignancies using SEER and comparable registries, including colorectal and breast cancer(Li, 2024; Q. Zhang et al., 2025).

In this study, the top performing models showed similar discrimination; DeepSurv, GBSA, and RSF showed similar mean C-indices (0.7542, 0.7540, and 0.7532, respectively), exceeding Cox regression (0.7417). This close clustering of the top C-indices has practical implications: while discrimination is necessary for risk ranking, it is not sufficient for selecting a model intended to provide reliable absolute risk estimates or to support threshold-based clinical decisions(Steyerberg, 2019). Calibration and prediction error have provided clear differentiation between approaches. At 36 and 60 months, calibration were stable, but DeepSurv underestimated early risk at 12 months in the lowest-risk decile despite observed events. This shows that a high C-index does not mean that short-term probability estimates are always reliable, which are often the most relevant in clinical settings. IBS over the 60-month horizon clearly favoured the tree-based ensembles (RSF 0.1516, GBSA 0.1534 vs DeepSurv 0.1644 and Cox 0.1597). The error rates for Cox regression and DeepSurv were higher.

DCA supported the ensemble-based models, showing greater net benefit than DeepSurv or Cox across threshold probabilities of 0.01-0.50 at 12, 36, and 60 months. NRI analyses using tertile risk categories and 1,000 bootstrap resamples, were compared with cox regression. RSF yielded largest reclassification benefit at 12 months (mean NRI 0.0482; 95% CI 0.0349-0.0614) and 36 months (0.0399; 95% CI 0.0280-0.0524), followed by GBSA (0.0308 and 0.0289, respectively). At 60 months, both ensembles performed well (RSF 0.0306; GBSA 0.0307). DeepSurv showed negative reclassification at 12 months (- 0.1746; 95% CI -0.1941 to -0.1545), near-neutral reclassification at 36 months, and only modest improvement at 60 months (0.0128). Direct comparison favours ensemble over DeepSurv at all time points. These findings state that for clinical risk

modeling considering discrimination alone is insufficient models are needed to demonstrate strong calibration, reclassification, and decision-analytic utility.

Permutation importance and SHAP confirmed that age, histologic subtype, nodal burden-related factors, and lymph node ratio as important predictors. These variables are well established in gastric cancer prognosis, which supports the model's validity(Grigorie et al., 2025). SHAP patterns showed non-linear effects of nodal and tumor features. It suggests that ML models can detect complex patterns that a standard linear model would likely miss. Anyhow, positive nodes, examined nodes, and lymph node ratio are closely related to one another, using these nodes together in a regression model can make the individual coefficients harder to interpret reliably.

We chose RSF as the main model for our prognostic prediction framework. The goal is to give personalized survival probabilities at specific time points, rather than just ranking risk. Because of this, we prioritized stable calibration and low prediction error, along with discrimination. RSF showed strong discrimination, the lowest IBS, and stable calibration at 12, 36, and 60 months. This made it the most balanced choice for estimating probabilities. GBSA had similar results on some measures and reclassified well, but RSF offered the best overall balance for deployment. DeepSurv performed well in discrimination, but its early calibration was unstable and its 12-month reclassification was weak. This made it less suitable for short-term risk estimates. RSF offers a good balance of accuracy and stability for public use as a prognostic calculator, but it still needs external validation and recalibration before it can be used in clinical settings.

We chose RSF as the main model for our prognostic prediction framework. The goal is to give personalized survival probabilities at specific time points, rather than just ranking risk. Because of this, we prioritized stable calibration and low prediction error, along with discrimination. RSF showed strong discrimination, the lowest IBS, and stable calibration at 12, 36, and 60 months. GBSA got similar results on several parameters and showed that it could effectively reclassify. DeepSurv exhibited strong discrimination; however, its early calibration was unstable and its 12-month reclassification performance was weak, making it less suitable for short-term risk estimation. RSF provides a good balance of accuracy and stability for public use as a prognostic calculator, but it still needs external validation and recalibration before it can be used in clinical settings.

This study has both strengths and limitations. Strengths include identical preprocessing and evaluation across all models, within-fold prevention of data leakage, and thorough analysis of discrimination, calibration, reclassification, and decision-analytic utility. Some of the limitations need to addressed which include the absence of cancer-specific survival, lack of molecular variables in SEER, and the need for external validation in independent cohorts. Future work should incorporate competing-risk modeling and prospective evaluation of risk-stratified treatment pathways.

References

1. Austin, P. C., Fang, J., & Lee, D. S. (2022). Using fractional polynomials and restricted cubic splines to model non-proportional hazards or time-varying covariate effects in the Cox regression model. *Statistics in Medicine*, *41*(3), 612–624. https://doi.org/10.1002/sim.9259

2. Balakrishnan, M., George, R., Sharma, A., & Graham, D. Y. (2017). Changing Trends in Stomach Cancer Throughout the World. *Current Gastroenterology Reports*, *19*(8), 36. https://doi.org/10.1007/s11894-017-0575-8

3. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/10.3322/caac.21834

4. Chen, D., Fu, M., Chi, L., Lin, L., Cheng, J., Xue, W., … Yan, J. (2022). Prognostic and predictive value of a pathomics signature in gastric cancer. *Nature Communications*, *13*(1), 6903. https://doi.org/10.1038/s41467-022-34703-w

5. Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

6. Duggan, M. A., Anderson, W. F., Altekruse, S., Penberthy, L., & Sherman, M. E. (2016). The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology. *American Journal of Surgical Pathology*, *40*(12), e94–e102. https://doi.org/10.1097/PAS.0000000000000749

7. Gong, Z., Zhou, L., He, Y., Deng, Y., Zhou, J., Wang, W., … Ma, M. (2025). Explainable Machine

Learning Model for Predicting Postoperative Survival in Patients With Locally Advanced Gastric Cancer. *Cancer Medicine*, *14*(22). https://doi.org/10.1002/cam4.71408

8. Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*(17–18), 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

9. Grigorie, T. R., Verdea, C., Chiriac, T. D., Gramaticu, I. M., Iliesiu, A., Popescu, G. A., … Alexandrescu, S. T. (2025). Improving Prognostic Stratification in Gastric Cancer: The Role of Lymph Node Staging Systems. *Medicina*, *62*(1), 85. https://doi.org/10.3390/medicina62010085

10. Hothorn, T. (2005). Survival ensembles. *Biostatistics*, *7*(3), 355–373. https://doi.org/10.1093/biostatistics/kxj011

11. Institute, N. C. (2018). Surveillance, epidemiology, and end results (SEER) program. *Cancer Statistics, SEER Data & Software, Registry Operations*.

12. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3). https://doi.org/10.1214/08-AOAS169

13. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1), 24. https://doi.org/10.1186/s12874-018-0482-1

14. Lee, Y.-C., Chiang, T.-H., Chou, C.-K., Tu, Y.-K., Liao, W.-C., Wu, M.-S., & Graham, D. Y. (2016). Association Between Helicobacter pylori Eradication and Gastric Cancer Incidence: A Systematic Review and Meta-analysis. *Gastroenterology*, *150*(5), 1113-1124.e5. https://doi.org/10.1053/j.gastro.2016.01.028

15. Li, L.-W. (2024). Development and validation of a random survival forest model for predicting long-term survival of early-stage young breast cancer patients based on the SEER database and an external validation cohort. *American Journal of Cancer Research*, *14*(4), 1609–1621. https://doi.org/10.62347/OJTY4008

16. Lordick, F., Carneiro, F., Cascinu, S., Fleitas, T., Haustermans, K., Piessen, G., … Smyth, E. C. (2022). Gastric cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, *33*(10), 1005–1020. https://doi.org/10.1016/j.annonc.2022.07.004

17. Mahar, A. L., Zagorski, B., Kagedan, D., Dixon, M., El-Sedfy, A., Vasilevska-Ristovska, J., … Coburn, N. (2018). Evaluating TNM stage prognostic ability in a population-based cohort of gastric adenocarcinoma patients in a low-incidence country. *Canadian Journal of Public Health*, *109*(4), 480–488. https://doi.org/10.17269/s41997-018-0102-1

18. McLernon, D. J., Giardiello, D., Van Calster, B., Wynants, L., van Geloven, N., van Smeden, M., … Taylor, J. (2023). Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes: Practical Guidance for Cox Proportional Hazards Models. *Annals of Internal Medicine*, *176*(1), 105–114. https://doi.org/10.7326/M22-0844

19. Morgan, E., Arnold, M., Camargo, M. C., Gini, A., Kunzmann, A. T., Matsuda, T., … Soerjomataram, I. (2022). The current and future incidence and mortality of gastric cancer in 185 countries, 2020–40: A population-based modelling study. *EClinicalMedicine*, *47*, 101404. https://doi.org/10.1016/j.eclinm.2022.101404

20. Pencina, M. J., D'Agostino, R. B., & Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, *30*(1), 11–21. https://doi.org/10.1002/sim.4085

21. Steyerberg, E. W. (2019). *Clinical Prediction Models*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-16399-0

22. Sundar, R., Nakayama, I., Markar, S. R., Shitara, K., van Laarhoven, H. W. M., Janjigian, Y. Y., & Smyth, E. C. (2025). Gastric cancer. *The Lancet*, *405*(10494), 2087–2102. https://doi.org/10.1016/S0140-6736(25)00052-2

23. Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. K. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, *53*(2), 107–118. https://doi.org/10.1016/j.artmed.2011.06.006

24. Vickers, A. J., & Elkin, E. B. (2006). Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

25. Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, *11*(14–15), 1871–1879. https://doi.org/10.1002/sim.4780111409

26. Zeng, J., Li, K., Cao, F., & Zheng, Y. (2023). Development and validation of survival prediction model for gastric adenocarcinoma patients using deep learning: A SEER-based study. *Frontiers in Oncology*, *13*. https://doi.org/10.3389/fonc.2023.1131859

27. Zeng, J., Song, D., Li, K., Cao, F., & Zheng, Y. (2024). Deep learning model for predicting postoperative survival of patients with gastric cancer. *Frontiers in Oncology*, *14*. https://doi.org/10.3389/fonc.2024.1329983

28. Zhang, J., Gan, L., Xu, M., Huang, M., Zhang, X., Gong, Y., … Guo, W. (2018). The prognostic value of age in non-metastatic gastric cancer after gastrectomy: a retrospective study in the U.S. and China. *Journal of Cancer*, *9*(7), 1188–1199. https://doi.org/10.7150/jca.22085

29. Zhang, Q., Xu, R., Zhen, W., Bai, X., Li, Z., Zhang, Y., … Li, X. (2025). Competing risk and random survival forest models for predicting survival in post-resection elderly stage I–III colorectal cancer patients. *Scientific Reports*, *15*(1), 24269. https://doi.org/10.1038/s41598-025-05824-1

30. Zubarayev, M., Min, E.-K., & Son, T. (2019). Clinical and molecular prognostic markers of survival after surgery for gastric cancer: tumor-node-metastasis staging system and beyond. *Translational Gastroenterology and Hepatology*, *4*, 59–59. https://doi.org/10.21037/tgh.2019.08.05