

A Comparative Evaluation of Traditional, Machine Learning, and Deep Learning Survival Models for Breast Cancer Survival Prediction.

Abraham Nsiah^{1,2*}, King Solomon Otoo³, Sadick Abubakari², Kenneth Agyare⁴, Lawrence M. Agbota⁴

1. University of Louisville, School of Public Health, Department of Biostatistics, Louisville-USA
2. Ball State University, Department of Statistics, Muncie-USA
3. Department of Mathematical Sciences, University of Texas at El Paso, USA
4. School of Mathematical and Statistical Science, College of Sciences, University of Texas Rio Grande Valley, USA.

* E-mail of the corresponding author: abrahambestnsiah@gmail.com

Abstract

Breast Cancer is the leading cause of death from cancer in women worldwide. In our study, we compared the Cox proportional hazards model, Random Survival Forest, Survival Support Vector Machine, DeepSurv, and XGBoost Survival within a unified analytical framework. The analysis used 1,981 patients with complete overall survival information from the METABRIC dataset, including 1,144 deaths and 837 censored observations. Five-fold cross-validation was used for Cox PH and Random Survival Forest, while 10-fold cross-validation was used for Survival SVM, DeepSurv, and XGBoost Survival. Model performance was evaluated using Harrell's C-index, Uno's C-index, time-dependent AUC, Integrated Brier Score, calibration analysis, decision curve analysis, and explainability methods. Survival SVM achieved the highest Harrell's C-index of 0.7011 and AUC of 0.7541, whereas Random Survival Forest achieved a comparable C-index of 0.6970 and the lowest IBS of 0.1725. Random Survival Forest showed the most favorable prediction error, calibration, and clinical net benefit among models that generated survival probabilities. These findings indicate that Survival SVM provides the strongest discrimination, while Random Survival Forest offers the best overall balance of discrimination, probability accuracy, interpretability, and potential clinical utility for breast cancer survival prediction. Explainability analyses consistently identified age at diagnosis, Nottingham Prognostic Index, tumor size, and positive lymph-node involvement as the most influential predictors of mortality risk.

Keywords: Breast cancer, Random survival Forest, Cox Proportional Model, DeepSurv, Survival Support Vector Machine, XGBoost Survival, Kaplan–Meier Estimator.

DOI: 10.7176/CEIS/17-1-08

Publication date: September 28th, 2026

1. Introduction

Breast Cancer is known to be one of the most common cancers diagnosed in women. It is the leading cause of death from cancer in women worldwide. In the United States, breast cancer is the second most common cause of cancer-related death in women with an estimated 316,950 new cases and 42,170 deaths in 2025 according to American Cancer Society (Siegel *et al.* 2025). The treatment of breast cancer requires a multidisciplinary team of specialists in medical, surgical, and radiation oncology. Because most breast cancers are identified by mammography or by physical examination, primary care physicians are often the first contact for patients with a new breast cancer diagnosis. Primary care physicians should be familiar with the diagnostic workup for breast cancer and the basics of treatment (Trayes & Sarah 2021). But a significant issue in the field of bioinformatics or medical science is the accurate diagnosis of certain important information making the diagnosis of the disease an energetic and tricky job in medicine domain. Early detection of disease has become a crucial problem due to rapid population growth in medical research in recent times. With the rapid population growth, the risk of death incurred by breast cancer is rising exponentially (Islam *et al.* 2020).

Over the years, advances in artificial intelligence have led to the application of Machine Learning (ML), Deep Learning and Statistical Survival Models in breast cancer prognosis. Researchers have achieved rapid progress in the prediction of breast cancer, which has paved the way for the development of innovative healthcare tools. Traditional statistical approaches, particularly the Cox proportional hazards model, have been widely used

because of their interpretability and ability to analyze censored survival data. More recently, machine learning methods such as Random Survival Forests, Survival Support Vector Machines, and gradient boosting algorithms, as well as deep learning models including DeepSurv, have demonstrated considerable potential for capturing complex nonlinear relationships and improving predictive performance (Baidoo & Rodrigo 2025; Buyrukoğlu 2024; Cai *et al.* 2025; Takele & Chen 2026; Tiwari *et al.* 2020). These developments have contributed to the emergence of intelligent clinical decision-support systems for personalized cancer care.

However, important gaps remain in existing literature. Although numerous studies have developed breast cancer survival prediction models using either traditional statistical methods or individual machine learning techniques, relatively few have conducted a comprehensive comparison of traditional, machine learning, and deep learning survival models within a unified analytical framework. Many existing studies evaluate only one or two modeling approaches, use a single performance metric, or provide limited model interpretability. Furthermore, there is limited evidence on how modern survival models compare when assessed using multiple discrimination, calibration, and prediction-error measures while incorporating explainability techniques for clinical interpretation. Consequently, it remains unclear which survival modeling approach offers the best balance between predictive accuracy, calibration, and interpretability for breast cancer prognosis. Therefore, this study presents a comprehensive comparative evaluation of Cox PH, RSF, Survival SVM, XGBoost Survival, and DeepSurv using a unified preprocessing and validation framework with multiple performance metrics and explainable artificial intelligence techniques to identify an accurate and interpretable model for breast cancer survival prediction.

2. Related Works

In recent years, we have seen notable advancements in the prediction of breast cancer disease thanks to the benefits of machine learning approaches, including their capacity to model nonlinear relationships and the caliber of their overall predictions.

Takele & Chen (2026) compared survival and classical machine learning models for predicting breast cancer survival in Ethiopia. Methods like Kaplan–Meier estimation, Cox proportional hazards, random survival forests (RSF), DeepSurv, and classical machine learning (SVM, XGBoost, LGBM, and RF) classifiers were used with evaluation metrics such as AUC, C-index, and Integrated Brier Score (IBS). Their findings demonstrated that random survival forest and random forest achieved the highest performance (C-index: 0.754; IBS: 0.091) and (0.729 ± 0.006) , respectively, outperforming the other models under consideration. The Shapley Additive Explanations (SHAP) analysis for the RSF model identified age, tumour size, metastasis, stage, comorbidities, and marital status as the most important predictors of breast cancer survival.

Baidoo & Rodrigo (2025) investigated data-driven survival modeling approaches to predict breast cancer survival, including survival-specific methods such as the Cox Proportional Hazards (CPH) model, Random Survival Forests (RSF), and Cox Proportional Deep Neural Networks (DeepSurv), as well as machine learning models like Random Forests (RF), XGBoost, Support Vector Machines (SVM) with an RBF Kernel, and LightGBM. The performance of survival-specific and ML models was evaluated using the Concordance index (C-index), Integrated Brier Score (IBS), mean accuracy, and mean AUC. The CPH model achieved a C-index of 0.71 ± 0.015 and an IBS of 0.08 ± 0.006 , while RSF demonstrated slightly better discriminatory power with a C-index of 0.72 ± 0.0117 . DeepSurv performed comparably, with a C-index of 0.71 ± 0.0095 and an IBS of 0.09 ± 0.0008 . Both Cox and RSF models achieved the lowest IBS (0.08), indicating accurate survival probability predictions over time. For ML models, RF achieved a mean AUC of 0.74 ± 0.0021 , and XGBoost with a mean AUC 0.69 ± 0.0183 , reflecting fair discriminatory ability but not accounting for censoring in survival data. The SHAP analysis for the top performing models highlighted the extent of lymph node involvement, Regional Node Positive (number of affected lymph nodes), tumor grade (cell abnormality and growth rate), progesterone status, and age as key predictors of breast cancer survival outcomes.

Also, Cai *et al.* (2025) developed a survival prediction model for patients with breast cancer using multiple methods. Their study analyzed 8,119 patients with HER2-positive, HR-negative breast cancer from the SEER database who were randomly allocated to training, validation, and test cohorts in a 7:1:2 ratio. Predictive models were developed using five feature sets and three algorithms (Cox PH, RSF, DeepSurv), with feature selection optimized via Concordance index (C-index). Evaluation revealed that the C-index of the DeepSurv models constructed using the training set is greater than 0.8 performed better than both the RSF and Cox PH models. Based on the training set, the DeepSurv model predicted the highest ROC-AUCs of 0.91, 0.863, and 0.855 for 1-, 3-, and 5-year overall survival (OS), respectively. The RSF model achieved the highest AUCs, specifically 0.876, 0.861, and 0.845, for 1-, 3-, and 5-year overall survival in the test group. Also, the calibration graphs indicate

that of the three models forecasting overall survival at 1, 3, and 5 years, the RSF model demonstrated the greatest level of agreement between predictions and actual observations, trailed by the DeepSurv model but, there was poor agreement between Cox PH model predictions and observed data.

The study cohort of Jin *et al.* (2025) included 306 patients with breast cancer, with most aged 40-60 years (204/306, 66.7%). RSF and Cox regression models were compared using the time-dependent area under the curve (AUC), the concordance index (C-index), and risk stratification. In the internal cohort, the RSF achieved significantly higher time-dependent AUCs compared to Cox regression at 1-year (0.811 vs 0.763), 3-year (0.834 vs 0.783), and 5-year (0.810 vs 0.771) intervals (overall C-index: 0.803, 95% CI 0.747-0.859, vs 0.736, 95% CI 0.673-0.799). External validation confirmed robust generalizability: the duke cohort showed 1-, 3-, and 5-year AUCs of 0.912, 0.803, and 0.776, respectively, while the SEER cohort maintained consistent performance with AUCs of 0.771, 0.729, and 0.702, respectively. Risk stratification using the RSF identified 25.8% (79/306) high-risk patients and a significantly reduced survival time ($P<.001$). Notably, the RSF maintained improved net benefits across decision thresholds in decision curve analysis (DCA). The RSF model showed promising performance across different molecular subtypes in all datasets. Also, based on the RSF predicted scores, patients were stratified into high- and low-risk groups, with notably poorer survival outcomes observed in the high-risk group compared to the low-risk group.

Kaindal & Venkataramana (2025) evaluated a combination of parametric statistical models and machine learning algorithms to identify the most influential prognostic factors affecting the survival of patients. Two commonly used parametric models, log-gaussian regression and logistic regression, were applied to assess the relationship between survival and a set of clinical variables, including age at diagnosis, tumor grade, primary tumor site, marital status, American Joint Committee on Cancer (AJCC) stage, race, and receipt of radiation therapy or chemotherapy. Among these, the neural network model exhibited the highest predictive accuracy. The random forest model achieved the best balance between model fit and complexity, as indicated by its lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. Across all models, five variables consistently emerged as significant predictors of survival: age, tumor grade, AJCC stage, marital status, and radiation therapy use. Their findings highlight the importance of combining traditional survival analysis techniques with machine learning approaches to enhance predictive accuracy and support evidence-based personalized treatment planning in breast cancer care.

3. Materials and Methods Used

3.1 Data Source and Information

The dataset analyzed in this paper represents Breast cancer data from the Cbioportal for Cancer Genomics. The significance of this study stems from the critical need to improve clinical outcomes through adequate preventive and therapeutic interventions for breast cancer patients. The dataset is made up of a combined dataset of patients and sample data called clinical dataset. This dataset contains 2509 instances across 36 distinct features. Each instance in the dataset corresponds to a unique record of hospitalized patients (females) diagnosed to determine the present of breast cancer. These records encapsulate comprehensive details about the patients' hospital stays consisting of laboratory tests and medication administration during their admission.

1. Response Variable

In this research, the (OS Month, OS Status) variables have been designated as the target or response variables for analysis. Overall survival was defined as the time, in months, from the specified study origin, such as diagnosis or surgery, to death from any cause or the last documented follow-up. The overall survival event indicator was coded as 1 for patients who died and 0 for patients who were alive or censored at the end of follow-up. Of the 2,509 patients in the original dataset, 1,981 had complete overall survival endpoint information, meaning that both survival time and survival status were available. Among these patients, 1,144 deaths were observed, while 837 patients were censored. The remaining 528 patients had both overall survival time and survival status missing. These patients were excluded from analyses in which overall survival was the outcome because neither an event time nor a censoring time was available. Overall survival outcomes were not imputed because imputing both survival time and event status would require generating unobserved outcome information and could bias survival estimates, hazard ratios, and model-performance measures. The analysis was conducted using 1,981 patients with complete overall survival time and status information.

2. Predictor Variables

Prior to model development, several variables were excluded to prevent data leakage and improve model reliability. Patient identifiers (PATIENT_ID, SAMPLE_ID) and imported row indices were removed because they contained no prognostic information. The variable VITAL_STATUS was excluded because it directly reflects the survival outcome and would introduce information leakage. Recurrence-free survival variables (RFS_MONTHS and RFS_STATUS) were also removed since the primary endpoint of this study was overall

survival. Variables with no variability (single-category variables: SEX, CANCER_TYPE, SAMPLE_TYPE) were discarded because they provided no discriminatory information. The dataset used for the analysis contains 1981 instances across 27 distinct features. Missing continuous variables were imputed using the median of the training data, while categorical variables were imputed using the most frequent category. Continuous predictors were standardized, and categorical predictors were transformed using one-hot encoding. All preprocessing parameters were estimated using the training dataset only and subsequently applied to the test dataset to prevent information leakage. The data were divided into 80% training and 20% independent test sets using a stratified split based on the event indicator to preserve the proportion of deaths and censored observations.

3.2 Model Tunning and Validation

To improve model robustness and reduce overfitting, 5-fold stratified cross-validation was employed for the CPH and RSF models, whereas 10-fold stratified cross-validation was used for the Survival SVM, DeepSurv, and XGBoost Survival models. Hyperparameter tuning was performed using grid search, and the optimal model for each approach was selected based on the highest mean validation Harrell's C-index before being retrained on the full training dataset and evaluated on the independent test dataset. For the RSF model, the tuning process considered 300 trees, minimum node sizes of 5, 10, and 20, minimum split sizes of 10 and 20, and max_features (mtry) equal to 50% of the predictors, with the optimal model using 300 trees, min_samples_split = 10, min_samples_leaf = 5, max_features = 0.5, and the default log-rank splitting rule. The Cox model was optimized over a range of ridge regularization parameters ($\alpha = 0.001, 0.01, 0.1, 1, 10, \text{ and } 100$), the Survival SVM was tuned using combinations of alpha (0.001, 0.01, 0.1, 1.0) and rank_ratio (0.5, 1.0), the XGBoost Survival model was optimized by maximum tree depth (2 and 3), learning rate (0.02 and 0.05), subsample ratio (0.8 and 1.0), column sampling ratio (0.8), and minimum child weight (1 and 5). Each candidate model was trained using 1000 boosting iterations (n_estimators = 1000) with L2 regularization ($\lambda = 1.0$) and L1 regularization ($\alpha = 0$), and the model with the highest mean validation Harrell's C-index was selected, while the DeepSurv model was tuned by optimizing its neural network architecture and training parameters using early stopping. Model performance was subsequently evaluated on the independent test dataset using Harrell's concordance index (C-index), Uno's C-index, time-dependent area under the curve (AUC), Integrated Brier Score (IBS), calibration curves, decision curve analysis, and explainable artificial intelligence techniques, including SHAP and feature importance analyses. All analyses were performed in Python 3.12 using Google Colaboratory, with the lifelines, scikit-survival, scikit-learn, XGBoost, PyTorch, PyCox, TorchTuples, SHAP, and Matplotlib libraries, and a fixed random seed (2026) was used to ensure reproducibility.

3.3 Methods Used

In our study, we compared the Cox proportional hazards model, Random Survival Forest, Survival Support Vector Machine, DeepSurv, and XGBoost Survival as the methods used for our analysis.

1. Kaplan–Meier Estimator

It is a nonparametric estimator of the survival function that appropriately accounts for right-censored observations and does not require a specific probability distribution for survival times (Goel *et al.* 2010; Kaplan & Meier 1958; Rich *et al.* 2010). The Kaplan–Meier estimator was used to describe the overall survival experience of the study population. The survival function is defined as

$$S(t) = P(T > t) \quad (1)$$

where T denotes the survival time. The Kaplan–Meier estimate of the survival function is

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \quad (2)$$

where t_j represents the j th ordered event time, d_j is the number of deaths occurring at time t_j , and n_j is the number of patients at risk immediately before time t_j .

2. Cox Proportional Hazards Model

The Cox proportional hazards model was used as the conventional statistical benchmark for evaluating the association between the predictor variables and overall survival. The Cox model is semiparametric because it does not require a parametric form for the baseline hazard function, although it assumes that the effects of the predictors are multiplicative and proportional overtime (Cox 1984; Harrell Jr 2015; Kleinbaum 1998; Therneau & Grambsch 2000). For a patient with predictor vector X , the Cox model is expressed as

$$h(t | X) = h_o(t) \exp(\beta^T X) \quad (3)$$

where $h(t | X)$ is the hazard function at time t , $h_o(t)$ is the unspecified baseline hazard function, and $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ is the vector of regression coefficients.

3. Random Survival Forest

Random Survival Forest (RSF) is a nonparametric ensemble-learning extension of random forests for right-censored survival data (Breiman 2001; Ishwaran *et al.* 2008). Unlike the Cox model, RSF does not require the proportional hazards assumption and can capture nonlinear predictor effects and complex interactions. The RSF algorithm constructs an ensemble of B survival trees using bootstrap samples of the training data. At each internal node, a random subset of “mtry” predictors is considered, and the split that produces the greatest separation between the survival experiences of the two child nodes is selected. The log-rank splitting rule was used in this study.

4. DeepSurv

DeepSurv is a deep neural-network extension of the Cox proportional hazards model (Katzman *et al.* 2018). It replaces the linear Cox risk function with a nonlinear function estimated by a multilayer feed-forward neural network. This enables the model to capture complex and nonlinear relationships among clinical, pathological, molecular, and treatment-related predictors. For a patient with predictor vector X_i , DeepSurv estimates a nonlinear log-risk function

$$\eta_i = h_\theta(X_i) \quad (4)$$

where $h_\theta(\cdot)$ represents the neural network and θ denotes its trainable parameters.

The DeepSurv hazard function is

$$h(t/X_i) = h_o(t) \exp[h_\theta(X_i)] \quad (5)$$

Model training was performed by minimizing the negative Cox partial log-likelihood,

$$L_{DeepSurv} = - \sum_{i: \delta_i=1} \left[h_\theta(X_i) - \log \left[\sum_{j \in R(T_i)} \exp[h_\theta(X_j)] \right] \right] \quad (6)$$

When regularization is included, the objective

becomes

$$L_{reg} = L_{DeepSurv} + \lambda \|\theta\|_2^2 \quad (7)$$

where λ controls the strength of $L2$ regularization.

5. Survival Support Vector Machine

The Survival Support Vector Machine (Survival SVM) extends support vector learning to censored time-to-event outcomes by formulating survival prediction as a ranking problem (Pölsterl 2020; Pölsterl *et al.* 2015; Van Belle *et al.* 2011). Rather than directly estimating the survival function, Survival SVM learns a risk score that correctly orders patients according to their observed survival outcomes. For two comparable patients i and j , where patient i experienced the event before patient j , the desired ranking relationship is

$$f(X_i) > f(X_j) \quad (8)$$

where $f(X)$ is the model-estimated risk score. For a linear Survival SVM,

$$f(X) = w^T X \quad (9)$$

where w is the vector of model coefficients.

6. XGBoost Survival

Extreme Gradient Boosting Survival is a tree-based boosting method adapted to censored survival data using the Cox partial-likelihood objective (Chen & Guestrin 2016; Friedman 2001). It sequentially adds regression trees to correct prediction errors from previous iterations while incorporating regularization, shrinkage, row subsampling, and feature subsampling. At boosting iteration m , the predicted log-risk for patient i is updated as

$$\hat{\eta}_i^{(m)} = \hat{\eta}_i^{(m-1)} + v f_m(X_i) \quad (10)$$

where $\hat{\eta}_i^{(m-1)}$ is the prediction from the previous iteration, $f_m(\cdot)$ is the newly added decision tree, and v is the learning rate.

3.4 Model Performance Evaluation and Explainability

Model performance was assessed using measures of discrimination, prediction error, calibration,

and explainability.

1. Harrell's Concordance Index

Harrell's concordance index measures the probability that, for a randomly selected comparable pair of patients, the patient with the shorter observed survival time receives the higher predicted risk score (Harrell Jr *et al.* 1996). A value of 0.5 represents random ranking, values between 0.5 – 1 represent good model performances, whereas a value of 1 represents perfect discrimination.

2. Uno's Concordance Index

Uno's concordance index was included as a censoring-adjusted measure of discrimination (Uno *et al.* 2011). It applies inverse-probability-of-censoring weights to reduce bias caused by heavy censoring. It complements Harrell's C-index by reducing potential bias from censoring and provides additional evidence about how well each model correctly ranks patients according to survival risk.

3. Time-Dependent Area Under the Curve

Time-dependent AUC was used to evaluate the ability of each model to distinguish patients who experienced death by time t from those who survived beyond time t . It is defined as

$$AUC(t) = P[\eta_i > \eta_j | T_i \leq t, T_j > t] \quad (11)$$

where η_i and η_j are the predicted risk scores for a case and a control at time t , respectively (Heagerty *et al.* 2000; Heagerty & Zheng 2005). Time-dependent AUC values were evaluated across follow-up, with specific estimates reported at 60 and 120 months.

4. Brier Score and Integrated Brier Score

The time-dependent Brier score evaluates the squared difference between the observed survival status and the predicted survival probability while accounting for censoring (Gneiting & Raftery 2007; Graf *et al.* 1999). It is defined as

$$BS(t) = \frac{1}{n} \sum_{i=1}^n w_i(t) [I(T_i > t) - \hat{S}(t|X_i)]^2 \quad (12)$$

where $w_i(t)$ represents the inverse-probability-of-censoring weight. The Integrated Brier Score summarizes prediction error over a specified time interval,

$$IBS = \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} BS(t) dt \quad (13)$$

Lower Brier scores indicate more accurate survival probability predictions. The IBS was not computed for Survival SVM because the model did not directly estimate survival probability functions.

5. Calibration Analysis

Calibration was assessed by comparing predicted survival probabilities with Kaplan–Meier observed survival probabilities at 60 and 120 months. Patients were grouped according to their predicted probabilities, and the mean prediction within each group was compared with the corresponding observed survival estimate.

6. Decision Curve Analysis

Decision Curve Analysis was used to evaluate the clinical utility of the models at 60 and 120 months. The net benefit at threshold probability p_t was calculated

as

$$NB(p_t) = \frac{TP_{p_t}}{n} - \frac{FP_{p_t}}{n} \left(\frac{p_t}{1 - p_t} \right) \quad (14)$$

where TP_{p_t} and FP_{p_t} are the numbers of true-positive and false-positive decisions at threshold p_t , respectively (Vickers & Elkin 2006). The net benefit of each model was compared with the strategies of treating all patients and treating no patients. Survival SVM was excluded because of its risk scores were not calibrated absolute event probabilities.

7. Shapley Additive Explanations (SHAP)

SHAP values are based on Shapley values from cooperative game theory. The SHAP value for predictor j can be written as

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (15)$$

where F is the complete set of predictors, S is a subset not containing predictor j , and $f(S)$ represents the prediction obtained using the predictors in subset S (Lundberg & Lee 2017). It values were used to measure the contribution of individual features to the model's predictions. Positive SHAP values increased predicted mortality risk, whereas negative values decreased predicted risk.

4. Results and Discussion

In our analysis, a total of five survival models, namely the Cox Proportional Hazards (CPH) model, Random Survival Forest (RSF), Survival Support Vector Machine (Survival SVM), DeepSurv, and XGBoost Survival, were developed and evaluated using the training and independent test datasets.

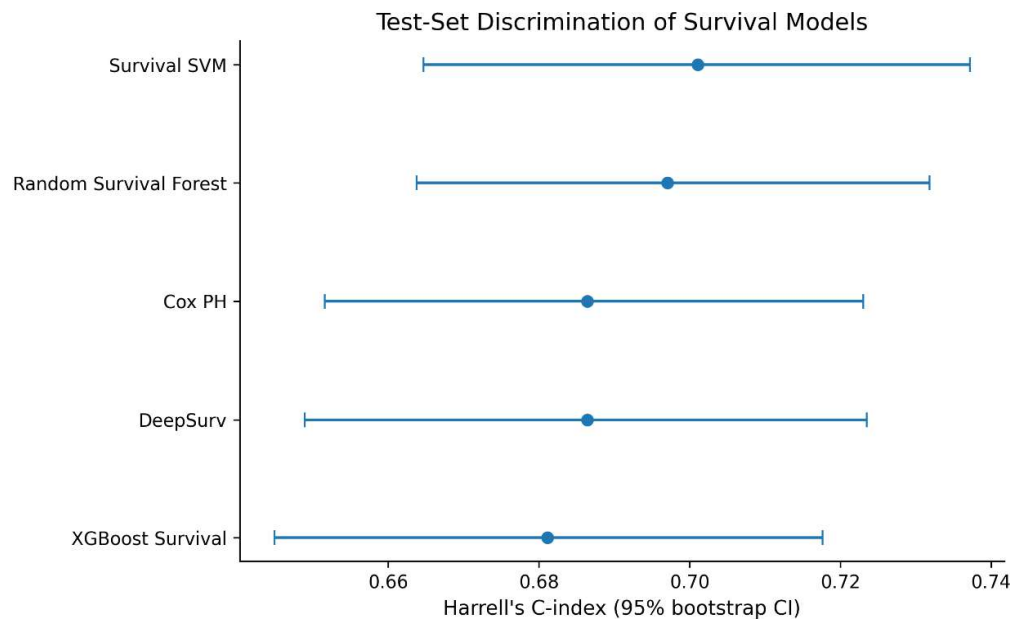


Figure 1: Kaplan–Meier overall survival curve for the breast cancer cohort.

Figure 1 shows the overall survival experience of patients during follow-up. The solid blue line represents the estimated survival probability, while the shaded region denotes the 95% confidence interval. The stepwise decreases correspond to observed death events, and the widening confidence interval toward the end of follow-up reflects the decreasing number of patients remaining at risk. The estimated survival probability decreased steadily throughout the follow-up period, indicating a progressive accumulation of mortality events over time. Survival remained high during the early months of follow-up but declined gradually as follow-up increased. By approximately 180 months, the estimated survival probability had fallen to about 45%, and by the end of follow-up (approximately 360 months), only a small proportion of patients remained alive. The widening 95% confidence intervals observed during the later follow-up period reflect the reduced number of patients remaining under observation, resulting in greater uncertainty in the survival estimates.

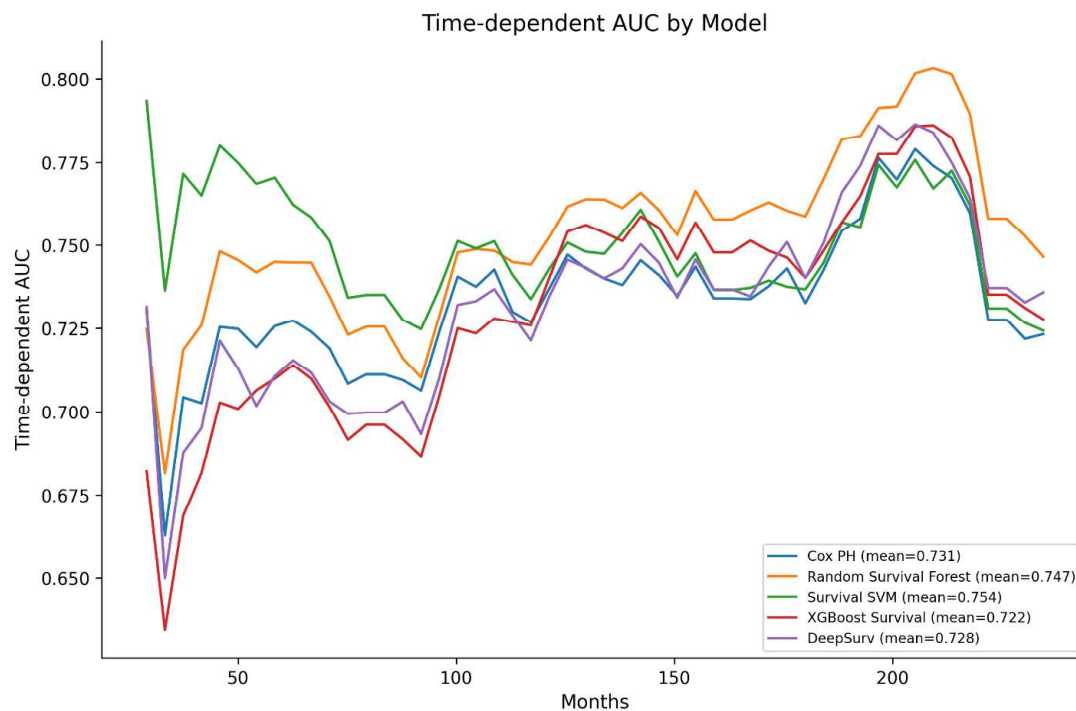


Figure 2: Forest plot of hazard ratios from the multivariable stratified Cox proportional hazards model.

Figure 2 presents the adjusted hazard ratios from the stratified Cox proportional hazards model. Points represent adjusted hazard ratios, horizontal lines indicate 95% confidence intervals, and the dashed vertical line denotes a hazard ratio of 1. Age was included as a stratification variable; therefore, no single hazard ratio was estimated for age. Higher Nottingham Prognostic Index (NPI) was associated with an increased risk of death, while advanced tumor stage, particularly Stage IV, showed the largest estimated hazard ratio. Most other variables had confidence intervals that crossed the null value (HR = 1), indicating limited evidence of an independent association with overall survival after multivariable adjustment. Age was treated as a stratification variable and therefore does not have a corresponding hazard ratio estimate.

Table 1: Comparison of performance of the survival model

Model	Harrell's C	Uno's C	AUC (60 months)	AUC (120 months)	Mean Dynamic AUC	IBS	Harrell's C 95% CI (Lower)	Harrell's C 95% CI (Upper)
Survival SVM	0.7011	0.6875	0.7660	0.7429	0.7541	NaN	0.6647	0.7372
Random Survival Forest	0.6970	0.6894	0.7513	0.7523	0.7518	0.1725	0.6638	0.7318
Cox Proportional Hazards	0.6864	0.6799	0.7313	0.7369	0.7342	0.1759	0.6516	0.7230
DeepSurv	0.6864	0.6828	0.7142	0.7346	0.7247	0.1761	0.6489	0.7235
XGBoost Survival	0.6811	0.6765	0.7167	0.7401	0.7288	0.2005	0.6449	0.7177

Figure 3 and Table 1 compare the discrimination performance of the five survival models on the independent test set. Survival SVM achieved the highest predictive performance, with the largest Harrell's C-index (0.7011; 95% CI: 0.6647–0.7372), followed closely by the Random Survival Forest (0.6970; 95% CI: 0.6638–0.7318). The Cox proportional hazards and DeepSurv models showed similar discrimination (Harrell's C = 0.6864), while

XGBoost Survival produced the lowest Harrell's C-index (0.6811). Consistent with these findings, Survival SVM also achieved the highest time-dependent AUC at 60 months (0.7660) and the highest mean dynamic AUC (0.7541). Among the models for which the Integrated Brier Score (IBS) could be estimated, the Random Survival Forest yielded the lowest IBS (0.1725), indicating the best overall prediction accuracy, whereas IBS was not computed for Survival SVM because it does not estimate individual survival probability functions required for Brier score calculation.

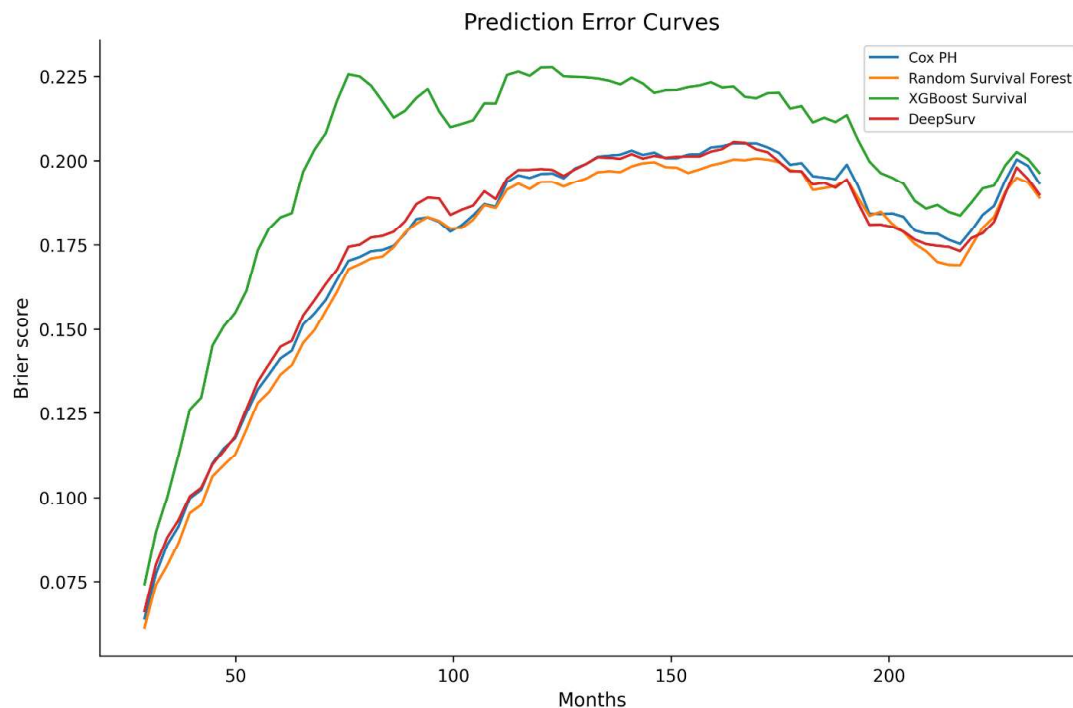


Figure 3: Comparison of test-set discrimination performance across survival models.

4.1 Time-Dependent AUC by Model

Figure 4 compares the time-dependent AUC of the five survival models throughout follow-up. Higher AUC values indicate better discrimination between patients with different survival outcomes over time. Overall, Survival SVM achieved the highest mean dynamic AUC (0.754), followed closely by the Random Survival Forest (0.752), indicating superior discrimination over time. The Cox proportional hazards, DeepSurv, and XGBoost Survival models showed comparatively lower but relatively stable performance, with mean dynamic AUC values of 0.734, 0.725, and 0.729, respectively. Although the AUC values fluctuated across follow-up, all models maintained good discriminatory ability, with most AUC estimates exceeding 0.70. These findings further support the superior overall discrimination performance of the Survival SVM and Random Survival Forest models.

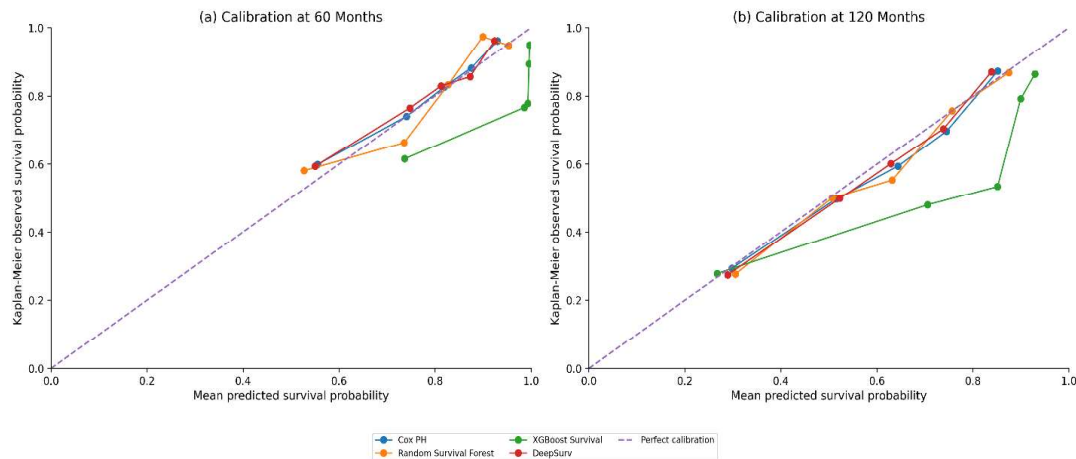


Figure 4: Time-dependent AUC of the five survival models across the follow-up period.

4.2 Brier Score

Lower Brier scores indicate better prediction accuracy over time. Figure 5 compares the prediction error of the survival models using the time-dependent Brier score. The Random Survival Forest consistently achieved the lowest prediction error throughout most of the follow-up period, indicating the best overall prediction accuracy. The Cox proportional hazards and DeepSurv models showed similar performance with only slightly higher Brier scores than the Random Survival Forest. In contrast, XGBoost Survival exhibited the highest prediction error across most time points, suggesting comparatively poorer predictive accuracy. These findings are consistent with the integrated Brier score results, where the Random Survival Forest produced the lowest IBS among the models capable of estimating survival probabilities.

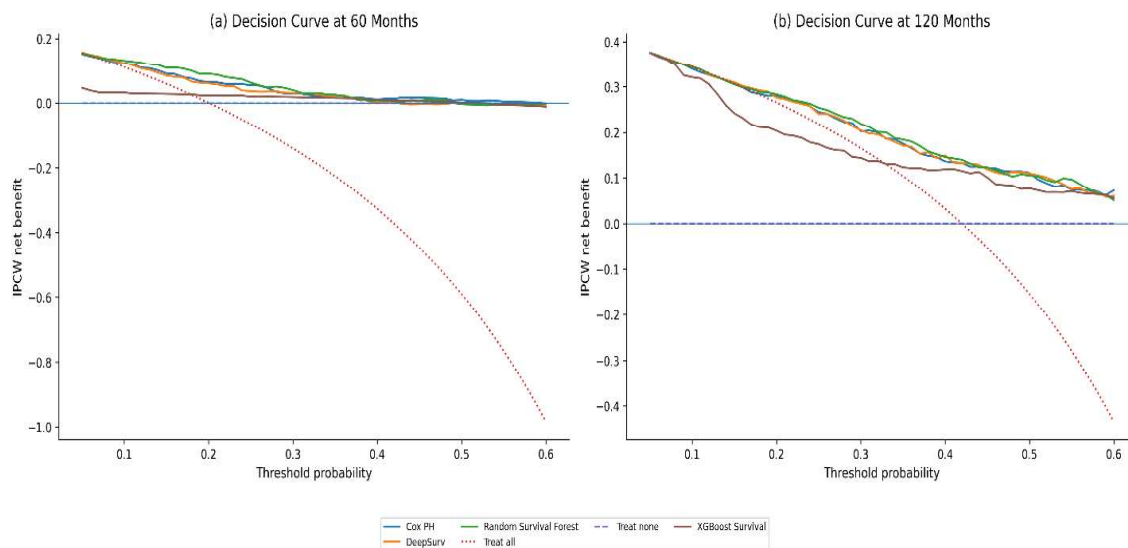


Figure 5: Prediction error (Brier score) curves for the survival models.

4.3 Calibration Plot

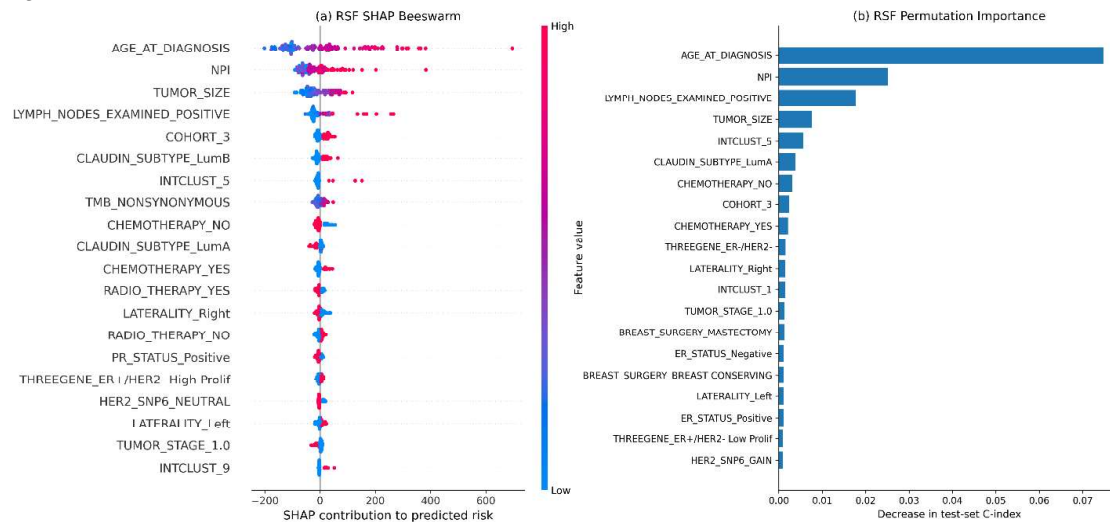


Figure 6: Calibration plots of the survival models at (a) 60 months and (b) 120 months.

Figure 6 compares the calibration performance of the survival models at 60 and 120 months. The dashed diagonal line represents perfect calibration, where predicted survival probabilities equal the observed Kaplan–Meier survival probabilities. At both time points, the Cox proportional hazards, Random Survival Forest, and DeepSurv models showed good agreement between predicted and observed survival probabilities, with calibration curves lying close to the reference line of perfect calibration. In contrast, XGBoost Survival deviated more noticeably from the diagonal, particularly at higher predicted survival probabilities, indicating poorer calibration. Calibration was generally maintained from 60 to 120 months for the Cox PH, Random Survival Forest, and DeepSurv models, demonstrating stable probability estimation over time. Overall, these results indicate that the Random Survival Forest, Cox PH, and DeepSurv models provided more reliable survival probability estimates than the XGBoost Survival model.

4.4 Decision Curve Analysis

Figure 7 presents the clinical utility of the survival models at 60 and 120-months using decision curve analysis. Higher net benefit indicates greater potential clinical utility across a range of threshold probabilities. The "Treat all" and "Treat none" strategies are included as reference curves. At both time points, the Cox proportional hazards, Random Survival Forest, and DeepSurv models consistently provided higher net benefit than the treat-all and treat-none strategies across most clinically relevant threshold probabilities. The Random Survival Forest generally achieved the highest net benefit over a broad range of thresholds, indicating the greatest potential clinical usefulness. In contrast, XGBoost Survival yielded lower net benefit than the other models across most threshold probabilities. Overall, the decision curve analysis suggests that the Cox PH, Random Survival Forest, and DeepSurv models offer greater clinical value for supporting individualized prognosis than XGBoost Survival.

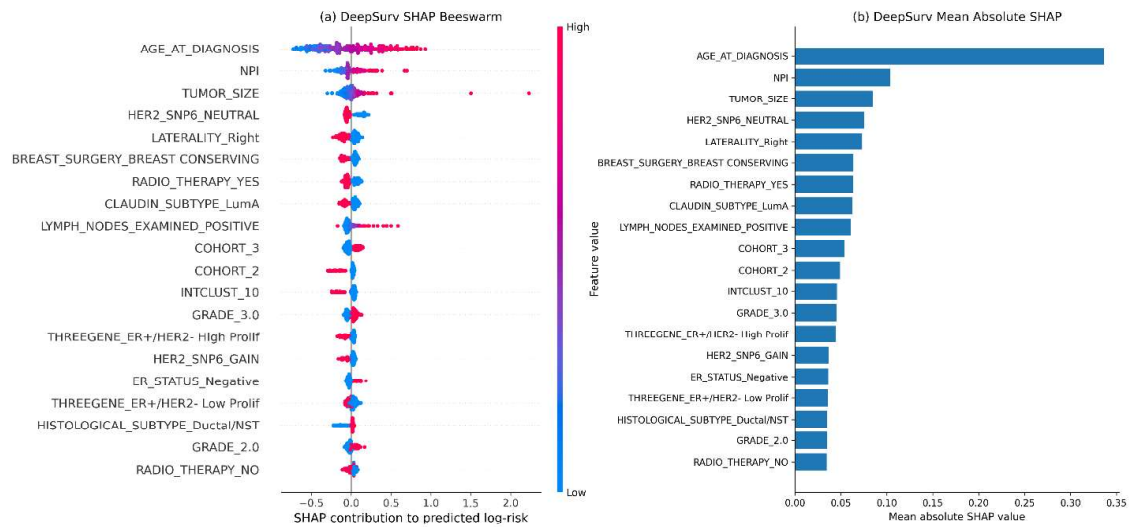


Figure 7. Decision curve analysis of the survival models at (a) 60 months and (b) 120 months.

4.5 Feature Importance

Figure 8 presents the explainability analysis for the Random Survival Forest model. Both the SHAP Beeswarm plot and permutation feature importance consistently identified Age at Diagnosis as the most influential predictor of survival, followed by the Nottingham Prognostic Index (NPI), positive lymph nodes examined, and tumor size. The SHAP plot further shows that higher values of these variables generally contributed to an increased predicted risk of mortality, whereas lower values were associated with reduced risk. Other clinical and molecular variables, including IntClust subtype, Claudin subtype, and chemotherapy status, had comparatively smaller contributions to the model predictions. The strong agreement between the SHAP and permutation importance analyses supports the robustness and interpretability of the Random Survival Forest model.

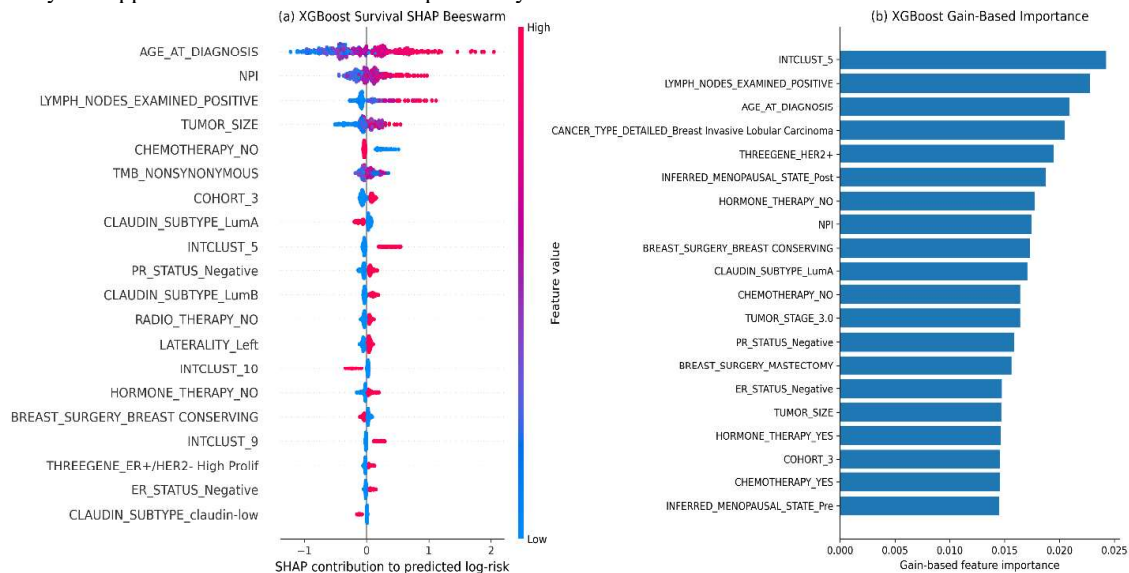


Figure 8. Random Survival Forest (RSF) explainability.

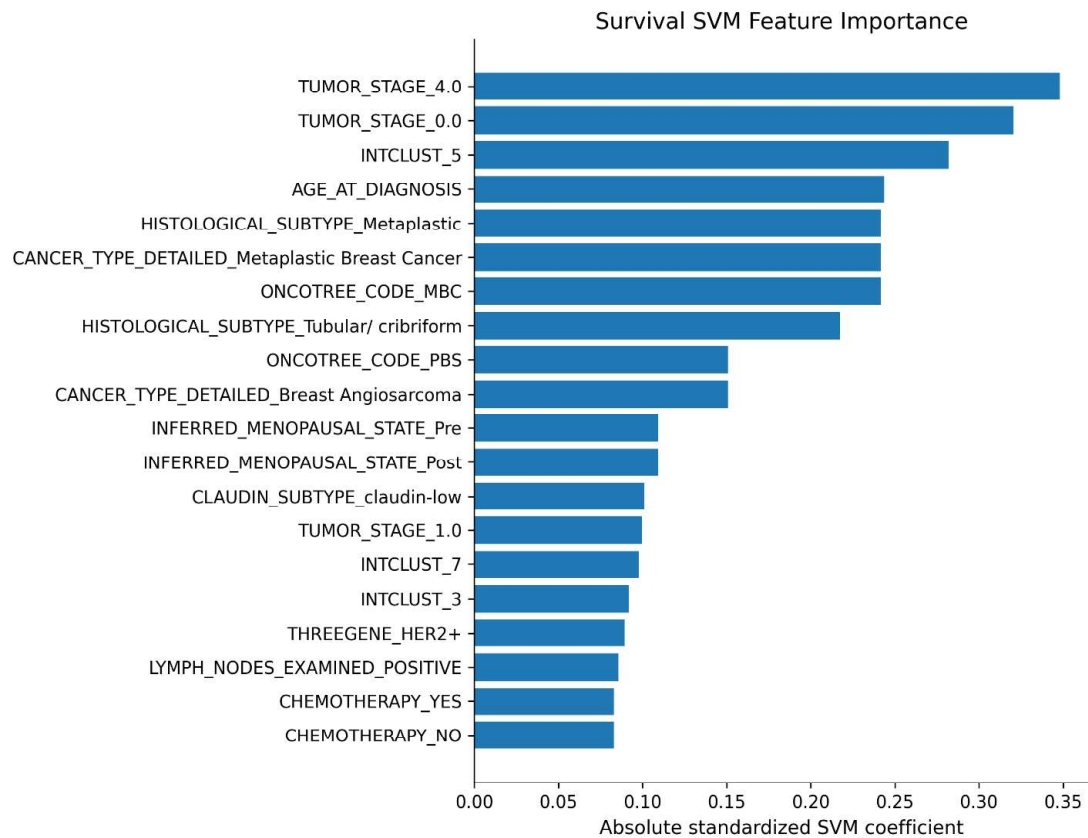


Figure 9. DeepSurv model explainability.

Figure 9 presents the SHAP-based interpretation of the DeepSurv model. a) SHAP beeswarm plot illustrating the contribution of individual features to the predicted log-risk, where red and blue represent high and low feature values, respectively. (b) Mean absolute SHAP values ranking the most influential predictors in the DeepSurv model. Both the Beeswarm plot and the mean absolute SHAP values consistently identified Age at Diagnosis as the most influential predictor of breast cancer survival, followed by the Nottingham Prognostic Index (NPI) and tumor size. The Beeswarm plot indicates that higher values of age, NPI, tumor size, and the number of positive lymph nodes generally increased the predicted log-risk of mortality, while lower values were associated with reduced risk. Clinical and molecular variables, including HER2 SNP6 status, laterality, breast-conserving surgery, radiotherapy, Claudin subtype, and IntClust subtype, also contributed to prediction but with smaller effects. Overall, the agreement between the SHAP beeswarm plot and the feature importance ranking demonstrates that the DeepSurv model primarily relies on established clinical prognostic factors while incorporating additional molecular characteristics to improve individualized survival prediction.

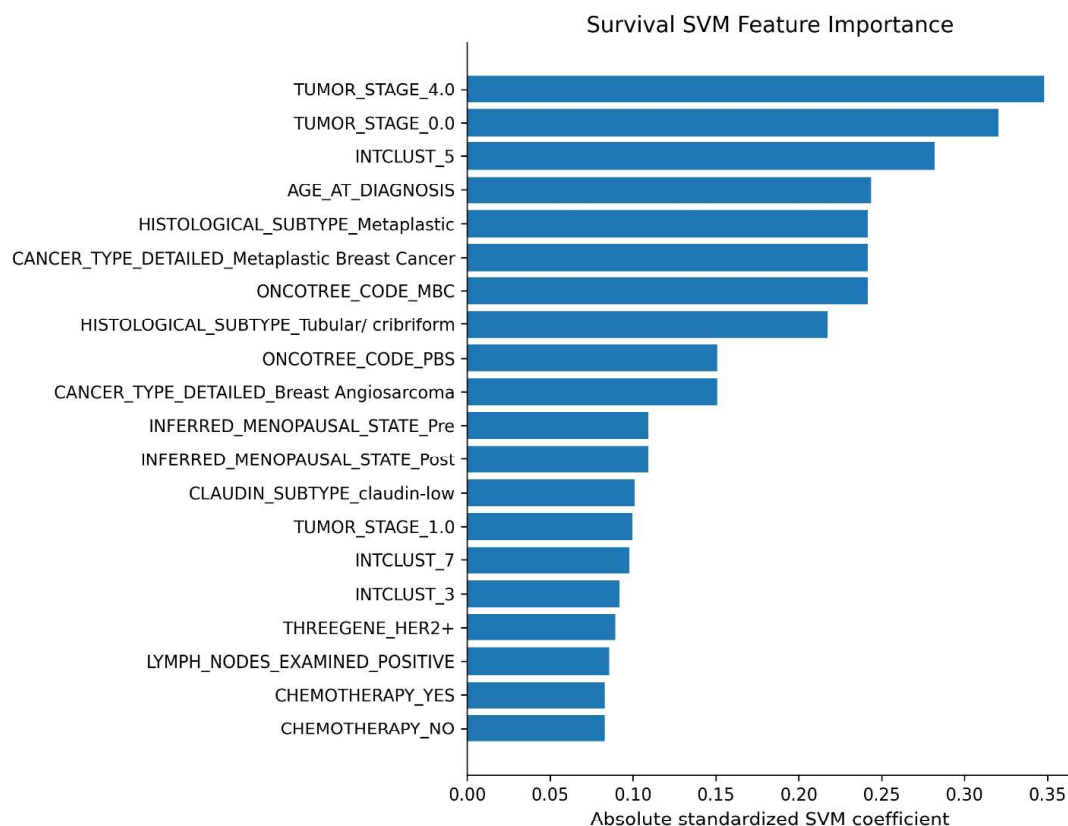


Figure 10. XGBoost Survival model explainability.

Figure 10 presents the explainability results of the XGBoost Survival model using SHAP values and gain-based feature importance. The SHAP Beeswarm plot identified Age at Diagnosis, NPI, number of positive lymph nodes, and tumor size as the strongest contributors to the predicted log-risk, with higher values generally associated with an increased risk of mortality. The gain-based importance ranking similarly highlighted INTCLUST_5, number of positive lymph nodes, Age at Diagnosis, NPI, and several clinicopathological variables, including tumor stage, hormone therapy, and breast surgery, as important predictors of survival. While the gain-based importance measures each variable's contribution to model construction, the SHAP analysis provides patient-level explanations by quantifying both the magnitude and direction of each feature's effect on the predicted risk. Overall, the agreement between the two explainability approaches demonstrates that the XGBoost Survival model relies on well-established clinical prognostic factors while also incorporating molecular subtype information to improve breast cancer survival prediction.

Figure 11 presents the feature importance of the Survival SVM model based on the absolute standardized coefficients. Variables are ranked according to the absolute standardized model coefficients, where larger coefficients indicate greater influence on the predicted survival risk. Tumor stage (Stage IV) was identified as the most influential predictor, followed by tumor stage (Stage 0), INTCLUST 5, and Age at Diagnosis, indicating their strong contribution to survival prediction. Several histopathological and molecular characteristics, including histological subtype, cancer type, OncoTree classification, and INTCLUST subtypes, also ranked among the most important predictors. Clinical variables such as positive lymph nodes, chemotherapy status, and menopausal status contributed to the model but had comparatively smaller effects. Overall, the Survival SVM model incorporated both conventional clinicopathological factors and molecular characteristics, demonstrating that breast cancer prognosis is influenced by a combination of clinical and biological features.

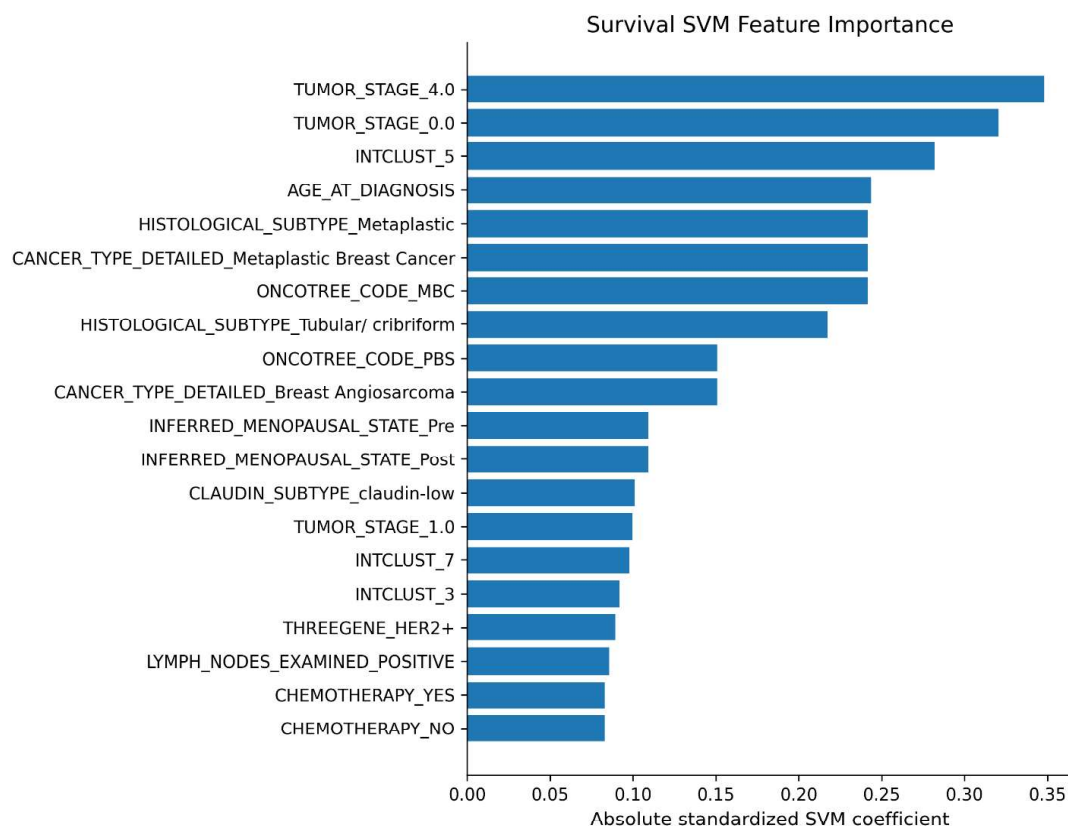


Figure 11. Survival Support Vector Machine (Survival SVM) feature importance.

4.6 Discussion

This study compared five survival modeling approaches Cox Proportional Hazards (CPH), Random Survival Forest (RSF), Survival Support Vector Machine (Survival SVM), DeepSurv, and XGBoost Survival for predicting breast cancer survival using multiple discrimination, calibration, prediction error, and clinical utility metrics. Overall, the results demonstrate that all models achieved acceptable predictive performance, with Harrell's C-index values ranging from 0.681 to 0.701, indicating good discrimination between patients with different survival outcomes. Among all models, the Survival SVM achieved the highest Harrell's C-index (0.7011), Uno's C-index (0.6875), and mean time-dependent AUC (0.7541), suggesting great overall discrimination. However, its inability to estimate individual survival probabilities prevented the calculation of the Integrated Brier Score (IBS) and excluded it from calibration and Decision Curve Analysis (DCA), limiting its usefulness for clinical risk prediction despite its strong ranking performance. The Random Survival Forest demonstrated the best overall balance between discrimination, calibration, prediction accuracy, and clinical applicability. It produced the lowest IBS (0.1725), indicating the smallest prediction error, while maintaining competitive discrimination performance and consistently providing the greatest net clinical benefit across both the 60 and 120-month decision curve analyses. These findings suggest that RSF effectively captures complex nonlinear relationships and interactions among clinical and molecular variables while providing reliable survival probability estimates required for clinical decision-making. Although the Cox Proportional Hazards model produced slightly lower discrimination than the machine learning models, it remained highly competitive and exhibited excellent calibration at both 60 and 120 months. The Cox model successfully addressed violations of the proportional hazards assumption while preserving the interpretability of hazard ratios for the remaining predictors. The forest plot further demonstrated that higher tumor stage, larger Nottingham Prognostic Index (NPI), and positive lymph node involvement were associated with an increased mortality risk, findings that are consistent with established clinical knowledge (Baidoo & Rodrigo 2025; Cai *et al.* 2025; Jin *et al.* 2025). DeepSurv achieved discrimination comparable to the Cox model while modeling nonlinear relationships between predictors and survival outcomes. Its calibration curves closely followed the ideal reference line, and the SHAP analysis identified Age at Diagnosis, NPI, tumor size, and lymph node involvement as the primary contributors to survival prediction. Although DeepSurv did not outperform RSF or Survival SVM, it demonstrated that deep learning can effectively model complex survival patterns while maintaining

interpretability through explainable artificial intelligence techniques. The XGBoost Survival model achieved the lowest overall discrimination and the highest prediction error among the evaluated probability-based models. Nevertheless, its SHAP explanations and gain-based feature importance consistently identified clinically relevant prognostic variables, including Age at Diagnosis, NPI, tumor size, lymph node involvement, and molecular subtypes, confirming that the model learned meaningful biological relationships despite its comparatively lower predictive performance. Across all explainability analyses, Age at Diagnosis, NPI, tumor size, and positive lymph nodes consistently emerged as the most influential predictors regardless of modeling approach, reinforcing well-established variables importance in breast cancer prognosis (Baidoo & Rodrigo 2025; Cai *et al.* 2025; Jin *et al.* 2025).

5. Conclusion

This study demonstrates that machine learning and deep learning survival models provide competitive performance for breast cancer survival prediction while offering complementary strengths over traditional statistical methods. Survival SVM achieved the highest discrimination performance, whereas Random Survival Forest provided the best overall combination of predictive accuracy, calibration, low prediction error, model explainability, and clinical utility. The Cox Proportional Hazards model remained a strong benchmark because of its interpretability and robust calibration, while DeepSurv offered an effective nonlinear alternative with meaningful SHAP-based explanations. Collectively, these findings suggest that Random Survival Forest represents the most suitable model for clinical survival prediction because it combines high predictive performance with accurate probability estimation and practical decision support. Although several machine learning and deep learning survival models were compared, future research should validate these findings using external multi-center datasets, incorporate additional genomic and imaging biomarkers, and evaluate emerging transformer-based and foundation survival models to further improve individualized breast cancer prognosis. Also, the Survival SVM model could not estimate individual survival probabilities, preventing the calculation of the Integrated Brier Score (IBS) and excluding it from calibration and decision curve analyses. Despite incorporating explainability techniques such as SHAP and feature importance, key predictors of breast cancer progression, such as genetic and lifestyle factors, were not included in the dataset, which may limit a comprehensive understanding of the disease and the ability to establish causal relationships between prognostic factors and survival outcomes. Therefore, in the future, larger datasets will be needed to support the use of deep learning methods for a more thorough understanding of breast cancer.

Declaration of Competing Interest.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The dataset used in this study is publicly available through Cbioportal and can be accessed at: https://www.cbioportal.org/study/clinicalData?id=brca_metabrc.

References

- Baidoo, T.G. & Rodrigo, H. (2025), "Data-driven Survival Modeling for Breast Cancer Prognostics: A Comparative Study with Machine Learning and Traditional Survival Modeling Methods", *PloS one* **20**(4), e0318167.
- Breiman, L. (2001), "Random Forests", *Machine learning* **45**(1), 5-32.
- Buyrukoğlu, G. (2024), "Survival Analysis in Breast Cancer: Evaluating Ensemble Learning Techniques for Prediction", *PeerJ Computer Science* **10**, e2147.
- Cai, W., Qi, Y., Zheng, L., Wu, H., Yang, C., Zhang, R., Wu, C. & Yu, H. (2025), "Comparison of Random Survival Forest Based-Overall Survival With Deep Learning and Cox Proportional Hazard Models in HER-2-Positive HR-Negative Breast Cancer", *Cancer Reports* **8**(7), e70262.
- Chen, T. & Guestrin, C. (2016), "Xgboost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data mining*, 785-794.
- Cox, D.R. (1984), "Regression Models and Life-tables", *Journal of the Royal Statistical Society: Series B* **34**, 588-591.
- Friedman, J.H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine", *Annals of statistics*, 1189-1232.
- Gneiting, T. & Raftery, A.E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation", *Journal of the American statistical association* **102**(477), 359-378.
- Goel, M.K., Khanna, P. & Kishore, J. (2010), "Understanding Survival Analysis: Kaplan-Meier Estimate", *International journal of Ayurveda research* **1**(4), 274.

- 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.
- Harrell Jr, F.E. (2015), "Multivariable Modeling Strategies", In *Regression modeling strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer, 63-102.
- Harrell Jr, F.E., Lee, K.L. & Mark, D.B. (1996), "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors", *Statistics in Medicine* **15**(4), 361-387.
- Heagerty, P.J., Lumley, T. & Pepe, M.S. (2000), "Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker", *Biometrics* **56**(2), 337-344.
- Heagerty, P.J. & Zheng, Y. (2005). "Survival Model Predictive Accuracy and ROC Curves", *Biometrics* **61**(1), 92-105.
- Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. (2008), "Random Survival Forests", *The Annals of Applied Statistics* **2**(3), 841-860.
- Islam, M.M., Haque, M.R, Iqbal, H., Hasan, M.M., Hasan, M. & Kabir, M.N. (2020), "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", *SN Computer Science* **1**(5), 290.
- Jin, Y., Zhao, M., Su, T., Fan, Y., Ouyang, Z. & Lv, F. (2025), "Comparing Random Survival Forests and Cox Regression for Nonresponders to Neoadjuvant Chemotherapy among Patients with Breast Cancer: Multicenter Retrospective Cohort Study", *Journal of Medical Internet Research* **27**, e69864.
- Kaindal, S. & Venkataramana, B. (2025), "A Comparative Analysis of Parametric Survival Models and Machine Learning Methods in Breast Cancer Prognosis", *Scientific Reports* **15**(1), 31288.
- Kaplan, E.L. & Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations", *Journal of the American Statistical Association* **53**(282), 457-481.
- 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.
- Kleinbaum, D.G. (1998), "Survival Analysis, A Self-learning Text", In: Wiley Online Library.
- Lundberg, S.M. & Lee, S.I. (2017), "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems* **30**.
- Pölsterl, S. (2020), "Scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn", *Journal of Machine Learning Research* **21**(212), 1-6.
- Pölsterl, S., Navab, N. & Katouzian, A. (2015), "Fast Training of Support Vector Machines for Survival Analysis", In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Charn: Springer International Publishing, 243-259.
- Rich, J.T., Neely, J.G., Paniello, R.C., Voelker, C.C., Nussenbaum, B. & Wang, E.W. (2010), "A Practical Guide to Understanding Kaplan-Meier Curves", *Otolaryngology—Head and Neck Surgery* **143**(3), 331-336.
- Siegel, R.L., Kratzer, T.B., Giaquinto, A.N., Sung, H. & Jemal, A. (2025), "Cancer Statistics, 2025", *CA Cancer J. Clin* **75**, 10-45. <https://doi.org/10.3322/caac.21871>
- Takele, K. & Chen, D.G. (2026), "A Comparative Analysis of Data-driven Models for Breast Cancer Survival Prediction", *Scientific Reports* **16**(1), 10114.
- Therneau, T.M. & Grambsch, P.M. (2000), "The Cox Model", In *Modeling Survival Data: Extending the Cox Model*, Springer, 39-77.
- Tiwari, M., Bharuka, R., Shah, P. & Lokare, R. (2020), "Breast Cancer Prediction using Deep Learning and Machine Learning Techniques", *Available at SSRN*, 3558786.
- Trayes, K. P. & Sarah, E. (2021), "Cokenakes" *Breast Cancer Treatment. American Family Physician* **104**(2), 171-178.
- Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B. & Wei, L.J. (2011), "On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data", *Statistics in Medicine* **30**(10), 1105-1117.
- Van Belle, V., Pelckmans, K., Van Huffel, S. & Suykens, J.A. (2011), "Support Vector Methods for Survival Analysis: A Comparison between Ranking and Regression Approaches", *Artificial Intelligence in Medicine* **53**(2), 107-118.
- Vickers, A.J. & Elkin, E.B. (2006), "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models", *Medical Decision Making* **26**(6), 565-574.