香　港　大　學

THE UNIVERSITY OF HONG KONG

STAT3799 Directed studies in statistics

Individual Project Report

Investigating Machine Learning Methods for Survival Prediction with an Application to TCGA Breast Cancer Data

Supervisor: Dr. Y. Gu

Student name: Ziyu Wang (UID: 3035777547)

Session: 2023-24 2nd Semester

Word count: 8273

GitHub repository:

https://github.com/ZiyuWang1121/Investigating-machine-learning-methods-for-survival-prediction

Contents

Abstract

Background

Breast cancer remains one of the major global health challenges; It is the most commonly diagnosed cancer for women in the world. Despite significant progress in the implementation of mass screening programs, the problem of overdiagnosis and subsequent overtreatment remains prominent. This issue demands further exploration of more targeted treatments. Traditional statistical techniques used in survival analyses could provide some insight into the disease patterns; they are generally not adapted to the complex situations encountered in the real world. On the other hand, alternative methods, such as machine learning (ML), may lead to more accurate and efficient results.

Methods

This study evaluated and compared the performance of three ML models. They are: Deep Survival Analysis (DeepSurv), Survival Support Vector Regression (Survival SVR), and Random Survival Forest (RSF). Using the Cox proportional hazard (CPH) model as a benchmark, we compared the goodness of fit and calibration performance of the models. The evaluation is conducted using the concordance index (C index) to evaluate model goodness-of-fit and the integral Brier score (IBS).

We applied these ML models for survival prediction to a rich dataset consisting of clinical, biospecimen, and gene expression data from 1079 breast cancer cases within The Cancer Genome Atlas (TCGA). Our objective was to assess the effectiveness of these ML algorithms in predicting survival among breast cancer patients, aiming to enhance the identification and intervention strategies in managing breast cancer in practical healthcare scenarios.

Results

The findings confirm the superiority of deep learning performed over other algorithms considered. DeepSurv showed promising ability to discriminate (C-index: 0.831; 95% C.I.: 0.826, 0.837) and to calibrate (IBS: 0.122; 95% C.I.: 0.118, 0.126). In contrast, Survival SVR achieved moderate discrimination with a C-index of 0.611 (95% C.I.: 0.611, 0.611), while RSF was not as effective on these metrics, with a C-index of only

0.450 (95% C.I.: 0.447, 0.454) and an IBS of 0.213 (95% C.I.: 0.212, 0.214). The CPH model, employed as the baseline, displayed the poorest performance (C-index: 0.261, 95% C.I.: 0.261, 0.261; IBS: 0.336, 95% C.I.: 0.336, 0.336). Even though there were challenges with high right-censoring rates, all ML models outperformed the baseline, showing their potential to assist medical decisions. Moreover, DeepSurv introduces hope for improving prognostic evaluations among breast cancer patients. It can stratify individuals into risk groups based on their survival probabilities over specified time intervals provided by the model output.

Conclusions

Our study of machine learning models for predicting survival in breast cancer patients underscores the outstanding performance of DeepSurv. This model's success indicates the potential of deep learning in this area. By harnessing such model, clinicians can improve their decision-making processes, ultimately leading to better patient outcomes in breast cancer care.

Keywords: Breast Cancer; Survival Analysis; Machine Learning; TCGA; Prediction Model; Comparative Study; Deep Survival Analysis; Support Vector Regression; Random Survival Forest.

Introduction

Breast cancer is the most common cancer among women. It makes up about 11.7% of all cancer cases, which was around 2.3 million new cases. Breast cancer also comes fourth for cancer mortality at 666,000 deaths worldwide, representing 6.9% of all cancer-related deaths. Compared with other cancers, this major health challenge dominates the impact of female incidence rate and mortality in 157 countries and 112 countries, respectively (Bray et al., 2024).

Although screening methods, such as mammography, have helped to improve early detection, the occurrence of overdiagnosis and resulting overtreatment have complicated public health strategies (Sung et al., 2021). Here, the term overdiagnosis refers to the situation in which patients with the diagnosed tumor present no symptoms or life-threatening effects throughout their lifetime (Kale & Korenstein, 2018). This conception got into the focus of breast cancer screening methods and has become even more problematic, including the financial pressure on the patient and the medical institutions; it also triggers ethical disputes concerning the benefits and hazards of early detection.

According to prior findings, overdiagnosis rates range from 0% to 54% (Dunn et al., 2022). As a consequence, many patients may receive unnecessary surgery, radiation therapy, and various adjuvant therapies, which can harm the patient's quality of life. Therefore, the treatment and management of breast cancer have to be improved for the well-being of patients. Although statistical models such as the Cox Proportional Hazards (CPH) models are widely adopted, they might invoke some assumptions like proportional hazards and linearity of the log-hazard function, which holds only under limited situations (Katzman et al., 2016). Therefore, research on finding new methods to improve the accuracy and precision of survival prediction is on the rise.

In this study, we aimed to compare the performance of machine learning (ML) models in predicting the survival of breast cancer patients and explore their potential of adjusting treatment interventions. For this reason, we used breast invasive carcinoma data in The Cancer Genome Atlas (TCGA), which provides a wealth of clinical, biospecimen and gene expression information. Such a rich data source provides a basis

for our analysis.

Literature Review

ML has turned out to be one of the most vital area for survival prediction as it can handle complicated patterns. This can be obsearved in high-dimensional biomedical data in initiatives like TCGA. In this section, we aimed to to review relevant literature on TCGA gene expression data and its application in survival prediction methods of breast cancer so that consolidated previous research findings on this particular topic will be available towards a deeper understanding of the subject matter.

Deep learning is a powerful tool for survival analysis. It can unite multiple omics data types together to increase its prediction performance. Deep learning models were also examined by Chaudhary et al. (2018) on liver cancer patients and by Ching et al. (2018) on various types of cancer cases. Also, Katzman et al. (2016) proposed DeepSurv. It is is a personalized treatment recommendation system that utilizes gene and protein profiles from METABRIC and had achieved satisfying survival prediction performance.

To make the prognosis of breast cancer and uncover the molecular mechanisms of its pathogenesis, one should consider combining genetic information with gene expression data. Genetic algorithms and bioinformatics analyses allowed Zhang et al. and Divya & Suresh to identify key prognostic genes during their study; they provide valuable insights for personalized treatment strategies. Briefly speaking, molecular mechanisms, such as the action of somatic mutations and the DNA repair system, can be considered to provide a more accurate prognosis for female breast cancer.

The complexity of datasets like those from TCGA renders the use of dimension reduction techniques to improve model performance. Liu et al. (2016) and Zhao et al. (2015) proposed methods based on Principal Component Analysis (PCA), such as Block-Constraint Robust Principal Component Analysis (BCRPCA). This is useful for understanding an overview of genomic data and identifying important biomarkers for different cancers. Despite the appearance of new technologies, as discussed by Fanaee-T&Thoresen (2019), PCA is still a favorite method because of its robustness and efficiency.

A consistent theme observed in the progress in the field of ML for predicting survival in patients with breast cancer, particularly with the use of TCGA data, has been the use of complex imputation techniques to handle missing values. In 2022, Murugesan and Balamurugan came up with innovative approaches, such as Weighted Fuzzy Scores (WFS) and Bayesian Independent Principal Component Analysis (BIPCA), to address the inherent challenges associated with missing values. In addition, Zhu et al. (2021) suggested ensembling various single imputation models to enhance the prediction result and its reliability. Viñas et al. (2021) have investigated advanced deep-learning techniques such as PMI and GAIN-GTEx. These methods offer great advantages in terms of accuracy and computational efficiency compared to conventional methods. The use of these imputation methods has allowed researchers to overcome the limitations of traditional methods and improve the prognostic outcome of breast cancer treatment. On the other hand, Bhandari et al. (2022) highlighted the subtle tradeoff between imputation strategies, highlighting the importance of considering specific research objectives and data set characteristics.

Efficient normalization methods are essential for accurate and reliable gene expression analysis, especially in cancer research. Although quantile normalization was once a standard, it showed susceptibility to class-effect proportion and batch effects, which led to the development of "class-specific" normalization strategies to address these issues (Y. Zhao et al., 2020). New methods, such as Median Ratio Normalization (MRN) proposed by Maza et al. (2013), aim to reduce the bias associated with transcriptome size and show better consistency and robustness than existing methods. Henderi's research (2021) emphasized the importance of normalization in ML and pointed out that when using the k-NN algorithm to predict breast cancer, Min-max normalization is better than the Z score. In addition, Wang et al. (2024) explored the use of generative deep learning models (such as GANs and DMs) to generate synthetic gene expression data, demonstrating the potential of using Min-Max-GAN for preprocessing to improve the performance and reliability of downstream analysis. These findings emphasize the crucial role of standardized methods in ensuring the accuracy and interpretability of transcriptome data.

The accurate extraction and analysis of features from gene expression data are crucial for the prognosis of breast cancer. Strelcenia & Prakoonwit (2023) demonstrated the

effectiveness of feature engineering the diagnostics of breast cancer, using decision trees as an illustrative classifier. Tan et al. (2014) elucidated the capabilities of denoising autoencoders (DAs) to reveal complex patterns in genomic data, encapsulating both clinical and molecular insights. Q. Liu and Hu (2019) considered unsupervised deep learning, with results indicating that integrating gene expression with copy number alteration data through deep features could strengthen associations with clinical outcomes. Sun et al. (2017) introduced a novel deep learning approach, D-SVM, which combines deep neural networks with SVM to predict breast cancer prognosis, achieving notable success. He et al. (2019) focused on the synergistic effects of clinical variables and gene expression to stratify breast cancer, thereby enhancing the identification of subtypes associated with survival and recurrence rates. W. Liu et al. (2019) applied Independent Component Analysis (ICA) to proteogenomics data, yielding pathway-level insights. Lastly, Alzubaidi et al. (2022) tackled the challenge of identifying robust biomarkers through deep learning-based feature extraction, highlighting promising new biomarkers associated with hormone receptor status. These studies underscore the importance of advanced feature extraction and engineering techniques in improving breast cancer prognosis and guiding therapeutic decisions.

This literature review section describes various ML methods applied to survival prediction in breast cancer research. It also emphasizes the importance of integrating genetic data, using deep learning technology, methods such as dimensionality reduction, robust imputation, and effective normalization are crucial for maintaining data integrity and improving prediction accuracy.

Considering the potential of deep learning, the prevalent use of SVM and RSF in TCGA, and the importance of integrating gene expression data obtained through RNA-sequencing, we opt to compare these models with PCA applied. This comparative analysis aims to discern the strengths and limitations of each approach, thereby contributing to the refinement of survival prediction methodologies in breast cancer research. Before proceeding with our methodological details, however, it is essential to review the foundational concepts of survival analysis.

Background to Survival Analysis

This section covered the basic principles and key concepts underpinning survival analysis. Understanding these fundamentals is essential as they provide the groundwork for the following methodology section.

Survival analysis, as a critical statistical branch, specializes in predicting the timing of events in distinctive fields - from the expected lifespan of biological organisms and the failure rate of mechanical structures to economic phenomena such as credit default. This analytical technique is particularly important in medical studies because it gives insights into the duration of major health events (inclusive of cancer recurrence or loss of life) before they occur. It is critical in the field of cancer research; it aimed toward estimating the timing of key events inclusive of cancer recurrence or mortality and presenting a deeper understanding of the effect of treatment on survival.

In survival analysis, we often face the issue of censoring. This happens when the observation of subjects does not persist till the occurrence of the event. Such a phenomenon is often the result of the untimely withdrawal or conclusion of the research concern or the studies itself before observing the events. One typical type of this problem in reality is right-censoring, where some subjects were not observed to the event occurrence. Due to the uncertainty of sthe exact time of the event, this brings complexity.

Studying failure time distribution is key to survival analysis; it covers some essential elements. This include the failure time distribution, described by a continuous non-negative random variable T, as well as various quantities of interest, consisting of the distribution function $F(t)$, the density function $f(t)$, the survival function $S(t)$, the hazard function $\lambda(t)$, and the cumulative hazard function $\Lambda(t)$.

It is critical to understand that the functions are interrelated. For example, the survival function $S(t)$ can be shown as an integral of the density function $f(t)$. We then considered the hazard function $\lambda(t)$. As the derivative of negative logarithm of the survival function, it gives an implication of instantaneous risk. Another relationship is that the cumulative hazard function $\Lambda(t)$ is the integration of the hazard function over

time; this provides further insight into risk accumulation.

While applying ML methods in survival analysis, models like DeepSurv and RSF could exploit the relationships between the quantities of interests to process their outputs. For example, DeepSurv and RSF outputs the log-risk function and the cumulative hazard function, respectively. By employing the relationships discussed earlier, we could convert these model outputs into the desired statistical estimates like the survival probabilities. Specifically, the formula of the survival probability can be:

$$S(t) = 1 - F(t) = P(T \geq t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u)du}$$

After the comprehension of the essential quantities of interest in survival analysis, we could then explore the methods we employed in this study. This could be important for effectively predicting survival outcomes.

Methods

1. Data Exploration

In this section, we examined the dataset utilized in our study and furnished a detailed introduction to the methods of data acquisition, filtering, and preprocessing. These steps set the stage for subsequent investigations and analysis by ensuring the reliability of the data.

1.1. Data Source

The dataset used in this investigation is obtained from The Cancer Genome Atlas (TCGA). It is a distinguished repository housing genomic and clinical data pertinent to numerous cancer types. It aggregates data from over 11,000 cases spanning 33 tumor types, thereby facilitating comprehension of cancer's molecular basis. The organizational sample collection and compilation time of these data were from 2013 to 2016, providing a solid data foundation for our research (Hutter & Zenklusen, 2018).

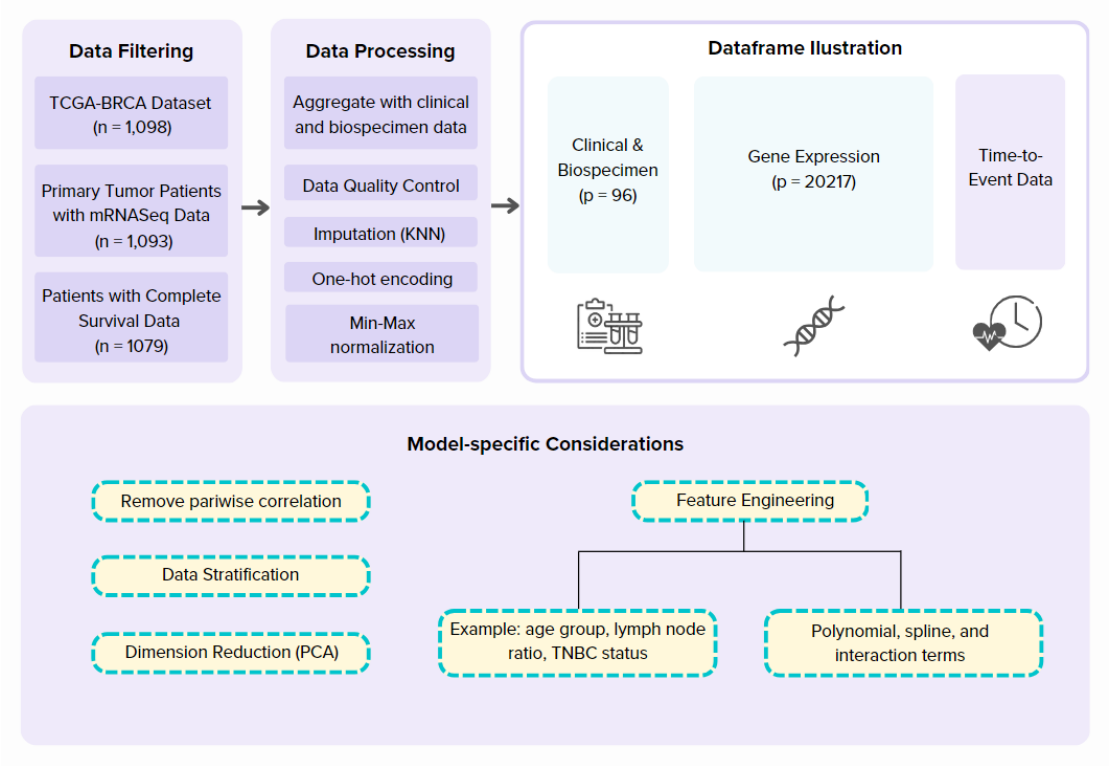
1.2. Data Filtering and Preprocessing

At the initial stage of data preprocessing, we first retrieved the original data from the breast cancer dataset of TCGA and focused on selecting the primary tumor samples that contained complete survival data and essential clinical and biospecimen information.

This selection process ensured that the data used in our analysis is highly relevant and comprehensive in inclusion. After filtering, we constructed a dataset containing 1079 individuals and then combined their gene expression data with clinical and biospecimen information based on patients' unique patient identifiers.

Figure 1

An illustration of data preprocessing



The preprocessing stage involved rigorous steps to ensure the quality and integrity of the data. We removed duplicate entries and features with constant values, as these did not contribute variability essential for analysis. Addressing missing data was crucial; we excluded features with over 20% missing data. We then filled the remaining gaps using the k-nearest Neighbors (KNN) imputation method with five nearest neighbors, preserving the dataset's structural integrity. For categorical variables, we used one-hot encoding scheme for encoding. We then process the data using Min-Max normalization to standardize data across different scales in order to prepare the dataset for subsequent analysis processes.

We also implemented preprocessing steps such as removing pairwise correlation, data

stratification, dimension reduction, and feature engineering. Specifically, principal component analysis (PCA) with 99% of the variance explained to reduce dimension.

Meanwhile, feature engineering plays a crucial role in mining deep-level information from data. For instance, we divided patients into different age groups and included lymph node ratios (calculated as the proportion of positive lymph nodes to total examined lymph nodes). At the same time, we also introduced a binary variable to indicate whether it is triple-negative breast cancer (TNBC). TNBC is a subtype of breast cancer with negative expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Given TNBC's aggressive nature, this feature could be meaningful for the analysis. In addition, we use polynomials, splines, and interaction terms to capture complex relationships in the data. By studying the interaction between nonlinear patterns and features in the data, we can further enhance the predictive ability of the model.

These additional preprocessing steps were considered only when they were beneficial to the model's performance. For example, DeepSurv can handle high-dimensional data well without the need for dimension reduction. In this case, employing PCA would reduce its predictive power due to the loss of information.

In our journey of exploring data, it is not an exaggeration to say that there is a complex connection between data preprocessing and subsequent training stages. The quality of data preprocessing has a significant impact on the performance of the model. Therefore, when training and validating models, it has become a necessary practice to constantly review and finely adjust data based on its performance. This iterative process is crucial for improving the overall performance of the model. Those well-performing datasets will be selected for further analysis. At the same time, the ones falling short of expectations undergo review to pinpoint areas for improvement or are discarded if their shortcomings cannot be remedied.

By applying the preprocessing steps as previously mentioned, we set up a solid foundation for further exploration of the dataset. These efforts had been pivotal in extracting pertinent information even as preserving the predictive performance necessary for subsequent analyses.

1.3. Exploratory Data Analysis (EDA)

In this subsection, we attempt to evaluate the distribution and interrelationships of clinical and biological sample variables (such as age, tumor stage, and number of lymph nodes examined), along with molecular data from gene expressions. This type of analysis is essential to understand the basic characteristics of the dataset.

We utilized various visualization techniques to better grasp our data. For example, histograms are helpful for examining the distribution of data like age and tumor stages. They show where most values lie and how they spread out. We also used box plots to display the quartiles and the extremes. In addition, heatmaps are very useful for observing the relationships between features and can clearly display the strength of these relationships.

This visual exploration deepen our understanding and directed the further stages of our analysis. They revealed issues such as skewness or outliers in the data; they help us identify areas that require further investigation or special attention. By counducting EDA, we can better decide on how to effectively preprocess data and choose the analysis techniques that are most suitable to meet our research objectives.

2. Statistical Analysis

In this section, we described the ML models and methods we performed. The models, DeepSurv, Survival SVR, and RSF were compared against the conventional CPH model. Each method has its own unique approach to dealing with complexity and patterns in the data.

2.1. Cox Proportional Hazards (CPH) model

The CPH model estimates the log-risk function, $h(x)$, by $\widehat{h_\beta}(x)$, where β is the coefficient vector. Here is its hazard function:

$$\lambda(X) = \lambda_0(t)e^{h(x)},$$

Where $\lambda_0(t)$ denotes the baseline hazard function. The risk score $r(x) = e^{h(x)}$ captures the covariate effect on the hazard rate. Estimating the log-risk function, h(x), involves maximizing the log partial likelihood, which accounts for the observed events

and the set of patients still at risk at each time point. The log partial likelihood is shown as:

$$l(\beta) = \sum_{i:E_i=1} \left(\widehat{h_\beta}(x_i) - log \sum_{j \in \Re(T_i)} e^{\widehat{h_\beta}(x_j)} \right)$$

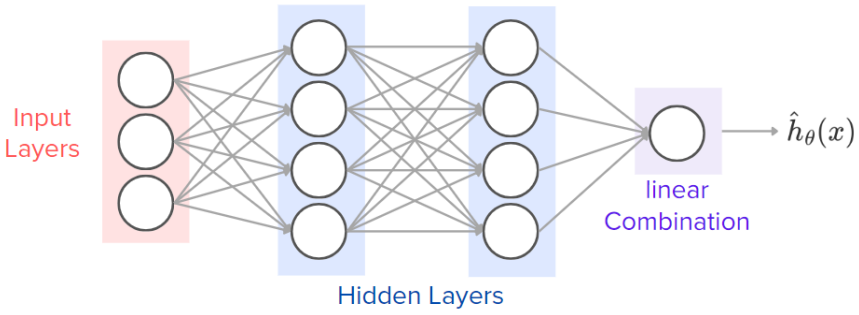
Where T_i, E_i, and x_i are denoted as the event time, event indicator, and covariates for the i^{th} subject, and $R(t) = \{i: T_i \geq t\}$ is the set of patients still at risk at time t.

2.2. Deep Survival Analysis (DeepSurv)

DeepSurv was proposed by Katzman et al. in 2016, presents an innovative approach to survival analysis through a deep learning perspective. It makes use of a feed-forward neural network architecture that excels at managing nonlinear log-risk functions and is crucial for describing complex relationships in survival data. This setting is beneficial for the model to grapple the interaction and non-linear effects that traditional models may overlook.

Figure 2

Model Architecture of DeepSurv



As you could see in Figure 2, the model structure includes fully-connected layers of neurons. At first, the input layer would take in the covariates and then pass them through one or more hidden layers. In these layers, each neuron employs a nonlinear activation function—usually the Rectified Linear Unit (ReLU)—to the weighted sum of its inputs, capturing the interactions of complex variables.

For the refinement of model parameters, DeepSurv leverages the gradient descent method to fine-tune the network's weights, aiming to minimize the loss function. This function is generally the negative log-likelihood from the Cox proportional hazards

model, modified with regularization strategies to curb overfitting:

$$l(\theta) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\widehat{h_\theta}(x_i) - log \sum_{j \in \Re(T_i)} e^{\widehat{h_\theta}(x_j)} \right) + \lambda ||\theta||_2^2$$

Where:

- $N_{E=1}$: number of patients with the event occurred
- λ: l_2 regularization parameter
- h(x): log-risk function
- T_i, E_i, x_i: event time, event indicator, and covariates for the ith subject
- $R(t) = \{i: T_i \geq t\}$: the set of patients still at risk at the time t

The fusion of a potent neural network architecture with a comprehensive optimization framework allows DeepSurv to deliver refined and precise predictions in survival analysis.

2.3. Survival Support Vector Regression (Survival SVR)

Survival SVR, proposed by Pölsterl et al. in 2015, extends the Support Vector Regression (SVR) concept to survival analysis. It predicts survival time by identifying the optimal hyperplane that maximally separates survival times.

The regression objective is based on an ordinary least squares problem with an l2 penalty, incorporating the consideration of censoring:

$$f(w, b) = \frac{1}{2}\mathbf{w^T w} + \frac{\gamma}{2} \sum_{i=0}^{n} \left(\zeta_{\mathbf{w},b}(y_i, x_i, \delta_i) \right)^2$$

where $\zeta_{w,b}(y_i, x_i, \delta_i)$ is defined as:

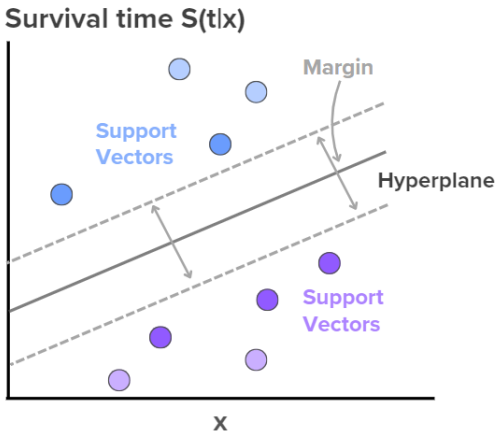
$$\zeta_{w,b}(y_i, x_i, \delta_i) = \begin{cases} max(0, y_i - w^T x_i - b) & \text{if } \delta_i = 0, \\ y_i - w^T x_i - b & \text{if } \delta_i = 1. \end{cases}$$

For each data point, $\zeta_{w,b}(y_i, x_i, \delta_i)$, represents the error between the predicted and observed survival time. This error is calculated differently depending on whether the observation is censored ($\delta_i = 1$) or uncensored ($\delta_i = 0$). In essence, censored subjects are only penalized for predictions less than their observed censoring time, while

uncensored observations are penalized for all incorrect predictions. Figure 3 below demonstrates the main idea of Survival SVR.

Figure 3

An illustration of Survival SVR



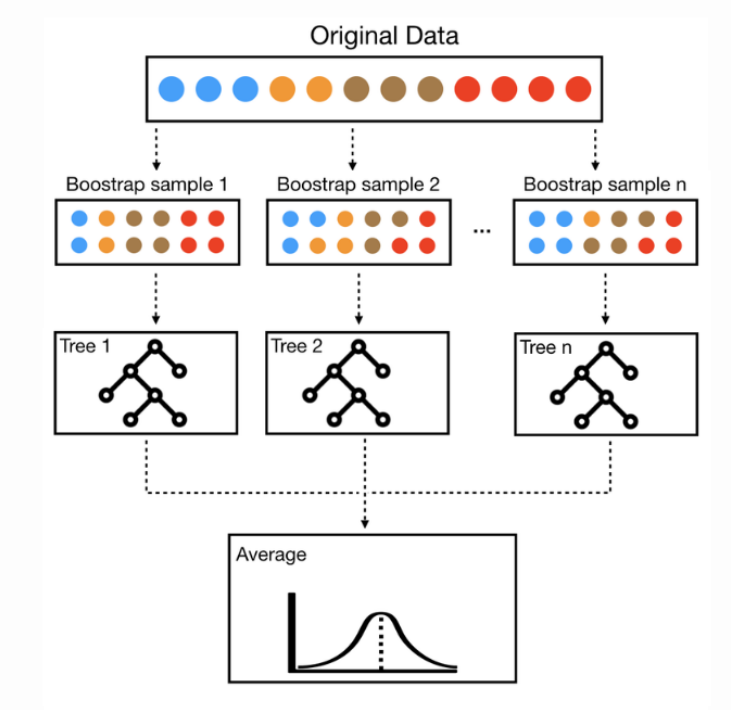
Coupled with truncated Newton optimization, the optimization process seeks to efficiently minimize this objective function to solve the convex quadratic problem. This approach ensures the model learns the optimal parameters to make accurate predictions while accommodating censored data.

2.4. Random Survival Forest (RSF)

The Random Survival Forest (RSF), developed by Ishwaran et al. in 2008, is an ensemble learning approach to estimate the cumulative hazard function (CHF). It mitigates overfitting by randomly sampling data instances and features at each tree node. In RSF implementation, B bootstrap samples are drawn, excluding approximately 37% as out-of-bag (OOB) data, and then survival trees are constructed for each sample. At each node, a subset of p variables is randomly selected, and splitting occurs based on the variable with the highest log-rank test statistic, ensuring maximal survival difference. The trees are grown to full size under the constraint that the terminal node must have at least d0 numbers of unique deaths. The ensemble CHF is derived by averaging the CHFs computed for each tree, with terminal node predictions estimated using the Nelson-Aalen estimator. Then, prediction errors of the ensemble CHF are evaluated using the OOB data, resulting in accurate and robust CHF estimation. The following Figure 4 roughly illustrates the procedures involved in RSF.

Figure 4

An illustration of random forests (Boehmke, 2018)



Using OOB data is crucial because calculating prediction errors requires predicting results. Since the decision trees in the model are trained on a bootstrapped sample of the original data, the OOB data represents the data points that were not included in the training of a particular tree. This approach provides an almost unbiased estimate of model prediction error, allowing for an accurate assessment of the model's prediction result on unseen data.

The ML models in this section are advantaged over traditional models such as CPH. Specifically, they can potentially enhance the accuracy and reliability of breast cancer management and survival prediction in other fields by means of capturing complex interactions in data.

3. Model Training & Evaluation

In this subsection, we outlined the training methodologies like hyperparameter tuning, and primary performance metrics used to determine and compare the efficacy of ML models for survival prediction

After preprocessing, the data undergoes a train-test split in an 80:20 ratio, allocating the majority for training to mitigate overfitting while ensuring a dependable evaluation of unseen data. The models are trained using this cohort, with hyperparameters fine-tuned, primarily through repeated cross-validation techniques—five iterations of 5-fold cross-validation for most models, with DeepSurv utilizing a random search to accommodate the lack of computational power.

The following table (Table 1) summarizes the optimization hyperparameters and related software packages for the models used in this study. Although the CPH model typically operates effectively under default settings, other models, such as DeepSurv, Survival SVR, and RSF require more custom configurations to achieve optimal performance.

Table 1

Optimized hyperparameters and software used

Models	Optimized hyperparameters	Software (Package)
CPH (as baseline)	None	R (survival)
DeepSurv	Activation = "relu", num_nodes = [32, 64], learning_rate = 0.1, epochs = 512, dropout = 0.2, batch_size = 256	Python (Pycox)
Survival SVR	alpha = 0.0009765625	Python (scikit-surv)
RSF	mtry = 771, nodesize = 50	R (randomForestSRC)

For example, the DeepSurv model employs an activation function referred to as Rectified Linear Unit (ReLU), which is $f(z) = max(0, z)$, across multiple hidden layers with varying node counts. We used a dropout rate of 0.2 in our training to prevent the model from overfitting and ran the training for 512 epochs, handling 256 batches of data in each one. For the Survival SVR model, we set the regularization alpha value to a specific 0.0009765625. We also adjusted the 'mtry' and 'nodesize' settings to get the best splits in the trees and the right sizes for the leaf nodes. These details are key because they show how important it is to adjust the settings just right to catch the complex patterns of survival in different types of model setups.

To minimize randomness and variability in model performance, we performed 20

repeated evaluations. The Concordance Index (C-index) and the Integrated Brier Score (IBS) were used to prediction accuracy and calibration of the model. We started with introducing the C-index.

Here, the C-index is a measure of the models' ability to accurately rank individuals based on their predicted survival times based on their risk scores. When there is no censoring, the C-index reduces to the Area Under the Curve (AUC) of a Receiver Operating Characteristic (ROC) curve. The formula for calculating the C-index is:

$$\text{C-index} = \frac{\sum_{i,j} I\left(T_j < T_i\right) \cdot I\left(r_j > r_i\right) \cdot \delta_j}{\sum_{i,j} I\left(T_j < T_i\right) \cdot \delta_j}$$

Where T_i and T_j are the survival times of individuals i and j, respectively. Also, r_i and r_j are their corresponding predicted risk scores. Lastly, δ_j is an indicator function denoting whether individual j was censored or not.

In this equation, the numerator counts the numbers of concordant pairs, instances where an individual j is uncensored and experienced the event before subject i. And if individual j also has a higher predicted risk score than i, such pair is considered concordant. Meanwhile, the denominator part calculates the number of all the permissible pairs, including all combinations of censored and uncensored individuals. We then obtain the C-index by evaluating the proportion of the numbers of concordant pairs over all permissible pairs. The resulted value ranges from 0 to 1. A C-index of 1 means the model ranks the survival times perfectly.

The Integrated Brier Score (IBS) also accounts for the issure of right-censoring, while calculating the mean squared difference between predicted probabilities and actual outcomes. It evaluates the overall accuracy and calibration of a model. Its calculation is as follows:

$$IBS(\tau) = \frac{1}{\tau}\int_0^{\tau} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(0 - \hat{S}(t|x_i)\right)^2 \cdot I(Y_i \le t, \delta_i = 1)}{\hat{G}(Y_i)} + \frac{\left(1 - \hat{S}(t|x_i)\right)^2 \cdot I(Y_i > t)}{\hat{G}(t)}\right) dt$$

Where $\hat{S}(x_i)$ is the predicted survival probability for uncensored and censored subjects, respectively. And $I(Y_i \le t, \delta_i = 1)$ indicates the occurrence of the event. The denominators are the Kaplan-Meier estimated survival probabilities of the censoring distribution. It is meant to adjust the IBS using the inverse probability of censoring

weights method. These squared differences are then integrated over time to calculate IBS. A lower IBS value signifies better model performance.

By reviewing the key elements in our methodology for training and evaluating survival analysis models, we aimed to ensure the reliability and effectiveness of the models in real-world applications.

Results

1. Patient characteristics

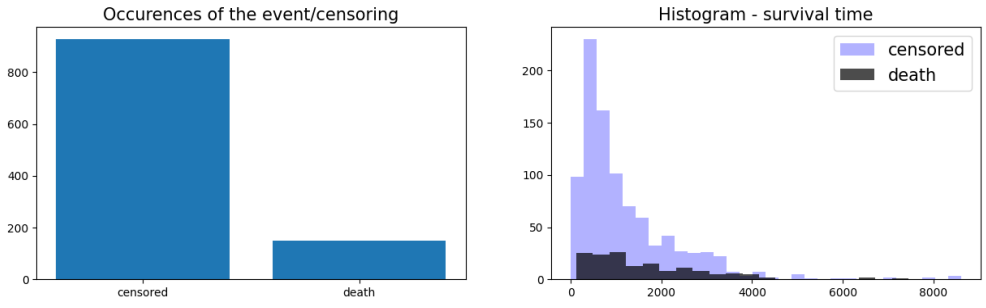
In our study, we carefully reviewed the demographic information, clinical symptoms, treatments, biomarkers, lymph node, and tumor features of the participants. The detailed insights provide a solid foundation for understanding the structure of the study cohort.

Participant Distribution

When looking at participant distribution, we noted that 86.0% of the participants exhibited a high censoring rate. This means that the vast majority did not experience the event of breast cancer mortality during the study period. This high proportion of censoring is important for our analysis as it is likely to profoundly affect the performance of our models.

Figure 5

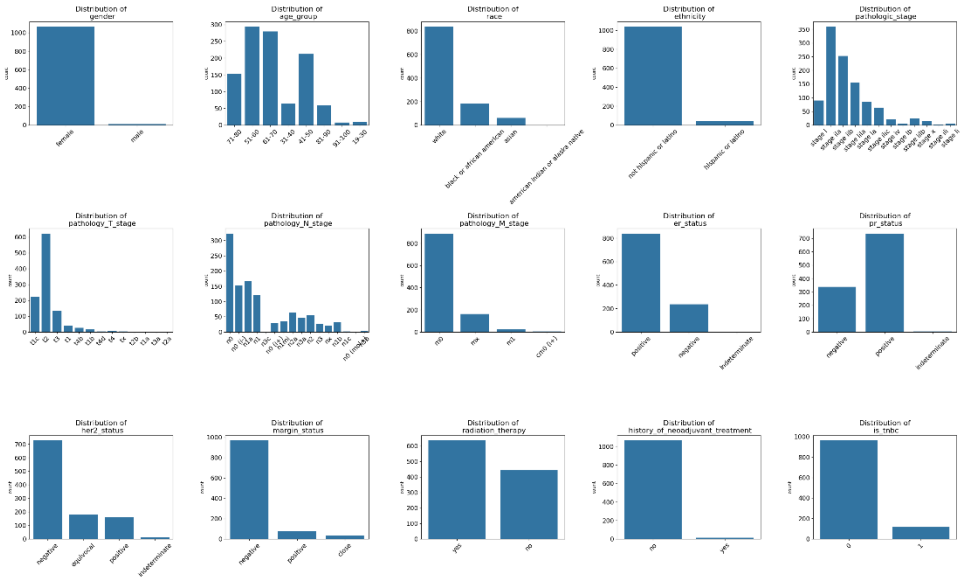
The occurrences of the event/censoring and the histogram of the survival times for the censoring and event



The visual representation in Figure 6 elucidates the distribution of categorical variables, which is summarized as follows:

Figure 6

Main bar plots of clinical and biospecimen features



Demographics

- **Gender and Race:** Predominantly female (98.9%), reflecting the higher incidence of breast cancer in women. The diversity in race, with a majority of White followed by Black or African American and Asian participants, underscores the importance of inclusivity in research.
- **Ethnicity and Age Distribution:** Most participants are non-Hispanic or Latino. The age distribution is primarily between 51-70 years, aligning with the typical age range for breast cancer diagnosis and underscoring the necessity for age-specific clinical and research approaches.

Clinical Profiles

- **Pathological Staging:** Most patients are in the early stages (IIA and IIB), with a notable proportion at stage IIIA, emphasizing the need for stage-specific interventions. The prevalence of T2 tumors highlights the importance of tumor size in prognosis and treatment planning.
- **Lymph Node Staging and Metastasis**: Predominantly N0 status, followed by N1a and N0 (i-), is pivotal for understanding disease progression and for treatment and prognostic assessments. Most patients are metastasis-free (M0), with some cases having unknown status (MX) and a smaller portion confirmed as metastatic (M1).

Treatment Characteristics

- **Surgical Procedures and Radiation Therapy:** A variety of surgical techniques

reflects the personalized nature of breast cancer treatment. Additionally, the widespread use of radiation therapy emphasizes its role in improving patient's prognostic outcomes.

- **Neoadjuvant Treatment:** The low proportion of patients receiving neoadjuvant treatment highlights suggests variations in treatment strategies and underscores the importance of evaluating its impact on survival.

TNBC Status

- The presence of triple-negative breast cancer in a subset of cases necessitates tailored therapeutic strategies due to the diversity of molecular subtypes.

Biomarkers

- **Hormone Receptor and HER2 Status:** The predominance of hormone receptor-positive and HER2-negative cases illustrates the molecular complexity of breast cancer.

Meanwhile, we could also obtain valuable insights into the heterogeneity of breast cancer by looking into the patient characteristics related to lymph nodes and tumors in the following Figure 7 and Table 2.

Figure 7

Histogram and Boxplots of clinical/biospecimen features

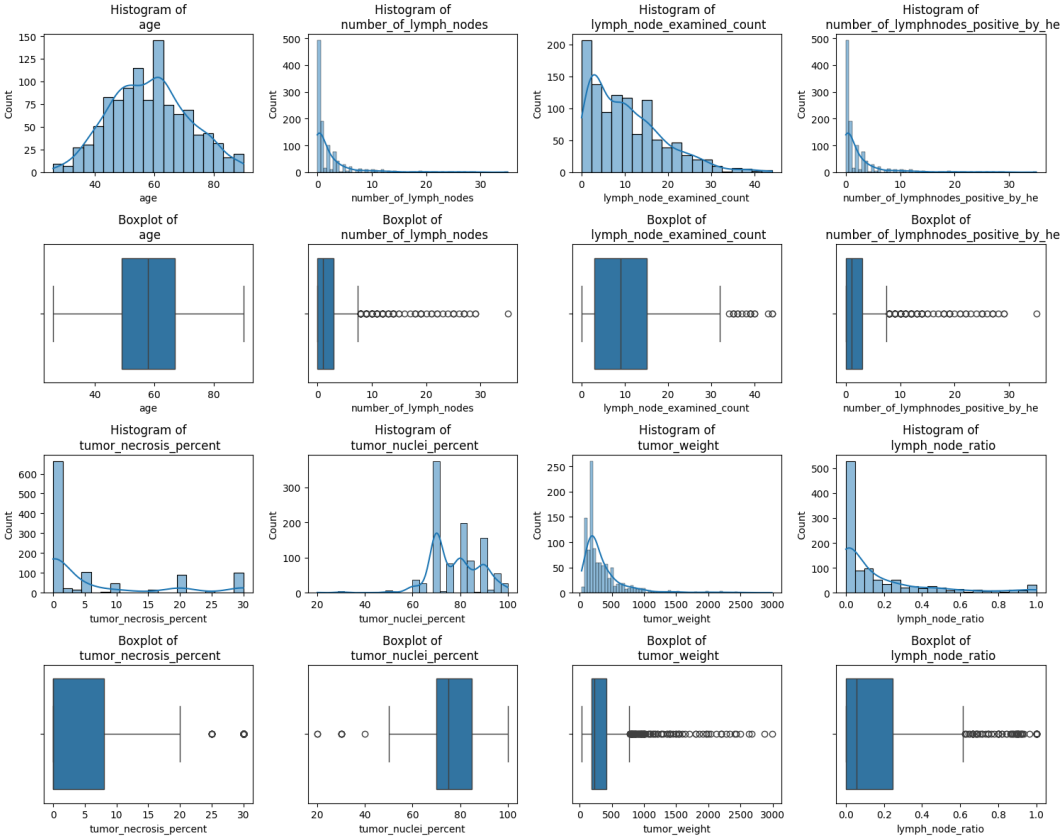


Table 2

Summary of lymph node and tumor characteristics

Features	Censored	Uncensored
Number of Lymph Nodes	2.03 ± 3.94	3.87 ± 6.08
Lymph Node Examined Count	10.09 ± 8.28	12.96 ± 8.46
Positive Lymph Nodes Count	2.03 ± 3.94	3.87 ± 6.08
Tumor Necrosis Percentage	6.49 ± 10.18	3.31 ± 7.18
Tumor Nuclei Percentage	77.90 ± 10.79	76.70 ± 10.44
Tumor Weight	364.42 ± 371.31	396.47 ± 400.69

Lymph Node Characteristics

In the censored group, about 2.03 lymph nodes on average are observed, which can vary as much as 3.94. On the other hand, the uncensored group exhibits more lymph nodes, averaging 3.87, with variations going up to 6.08. The variation is twice of that in the censored group. As for the total lymph nodes examined, the censored group

averages around 10.09 nodes, with variations up to 8.28. In contrast, the uncensored group examines more average positive nodes with slightly larger variations. This might mean clinicians generally observe more lymph nodes in the uncensored group. Moreover, the number of positive lymph nodes—indicating disease presence—is also higher in the uncensored group, with an average of 3.87 and variations up to 6.08, compared to just 2.03 (with variations up to 3.94) in the censored group. These indicate possibly more frequent disease occurrences in the uncensored group. The differences in the lymph node characteristics could reflect varying levels of progression in breast cancer.

Tumor Characteristics

The table also provides details on various tumor characteristics in both groups. The tumor necrosis percentage indicates the dead tissue proportion within a tumor. We found that it is higher in the censored group at an average of 6.49, with variations up to 10.18. In contrast, the uncensored group has a lower average of 3.31, with variations up to 7.18. It seems tumors in the censored group may be dying off more. Another measurement, the tumor nuclei percentage, shows the proportion of nuclei in the tumor. The censored group showed a slightly higher average of 77.90 compared to 76.70 in the uncensored group. Regarding the weight of these tumors, the censored group's tumors weigh about 364.42 grams on average, with a range of up to 371.31 grams. The uncensored group's tumors, however, are heavier, averaging 396.47 grams with a maximum variation of 400.69 grams. This might suggest that tumors are generally larger in the uncensored group.

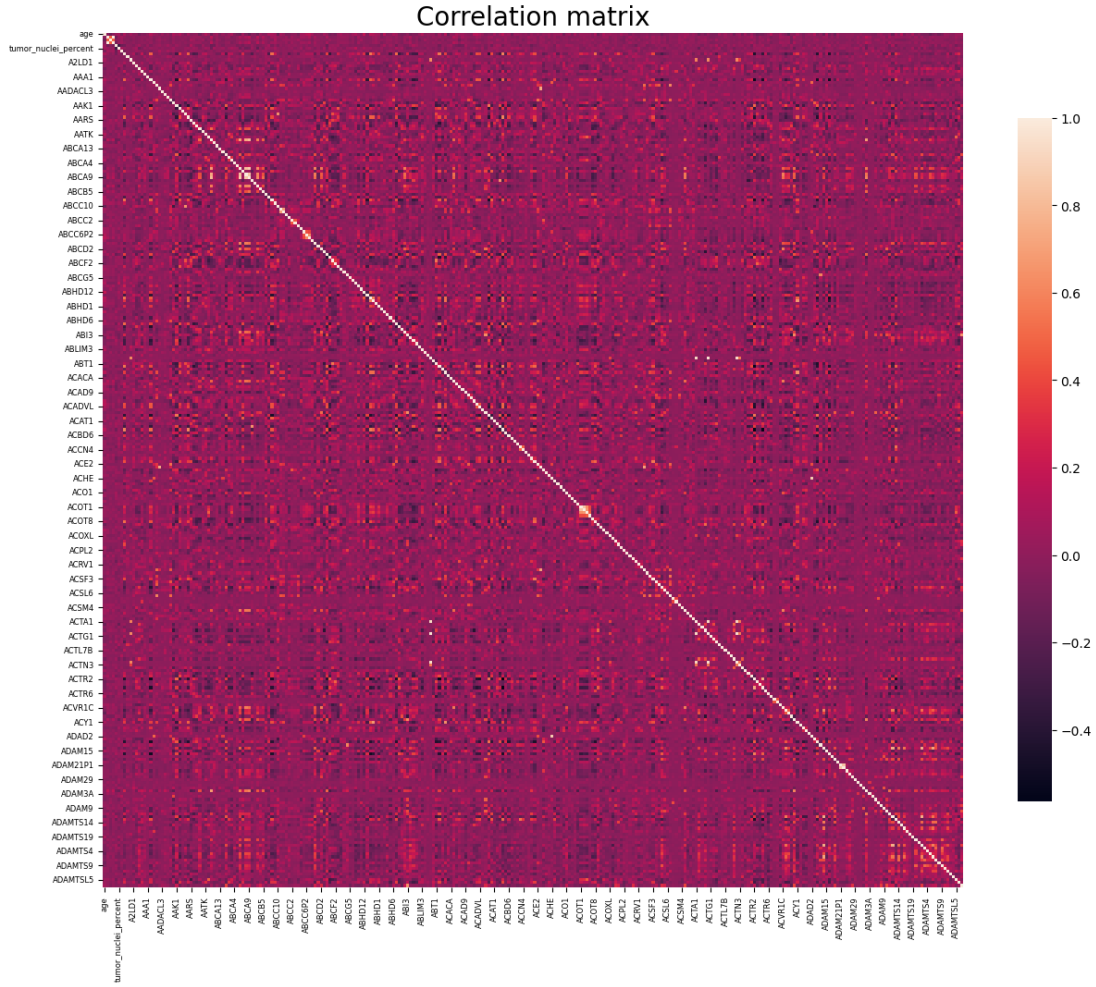
These insights help us understand the physical and cellular differences in lymph nodes and tumors between the censored and uncensored groups, which might reflect different stages of disease progression and possibly the aggressiveness of the conditions.

Correlation Matrix

The correlation matrix delineates the interrelationships among various engineered features and survival outcomes in patients with breast cancer, functioning as an instrumental resource for pinpointing features with significant correlations to patient survival. This tool aids in identifying potential predictive markers and understanding the mechanisms that impact outcomes.

Figure 8

The correlation matrix of the first 300 features as an example



As illustrated in Figure 8, which showcases the correlation matrix of the initial 300 features, the correlations with the time-to-event data are generally modest (<0.4). Nonetheless, specific engineered features, including interaction and polynomial terms within the matrix, show noteworthy correlations, approximately around the 0.2 mark. Although these are not high correlation values, they reveal significant relationships that merit further consideration.

Moreover, particular pathological staging variables such as ‘pathologic_stage=stage iv’ (0.242), ‘pathology_N_stage=n1b’ (0.339), and ‘pathology_M_stage=m1’ (0.263) exhibit relatively stronger correlations with survival outcomes compared to other features. These results underline the pivotal influence of disease progression and

metastasis in determining patient prognosis. The highlighted correlations of these staging variables reinforce their criticality as prognostic indicators for survival in breast cancer patients.

The comprehensive analysis of patient characteristics provides valuable insights into the diversity of breast cancer manifestations, which is helpful for clinical decision-making and prognosis evaluation.

2. Model performance

The performance of different machine learning models in predicting the survival outcome of breast cancer patients with right censored data characteristics was comprehensively evaluated. The DeepSurv model performed the best and significantly outperformed baseline Cox proportional hazard (CPH) models, among others. The performance of the model was compared using metrics such as C-index and Integrated Brier Score (IBS). The results were presented in the box plot and table as follows.

Figure 9

Boxplots of C-Index and IBS for different models over 20 replicates

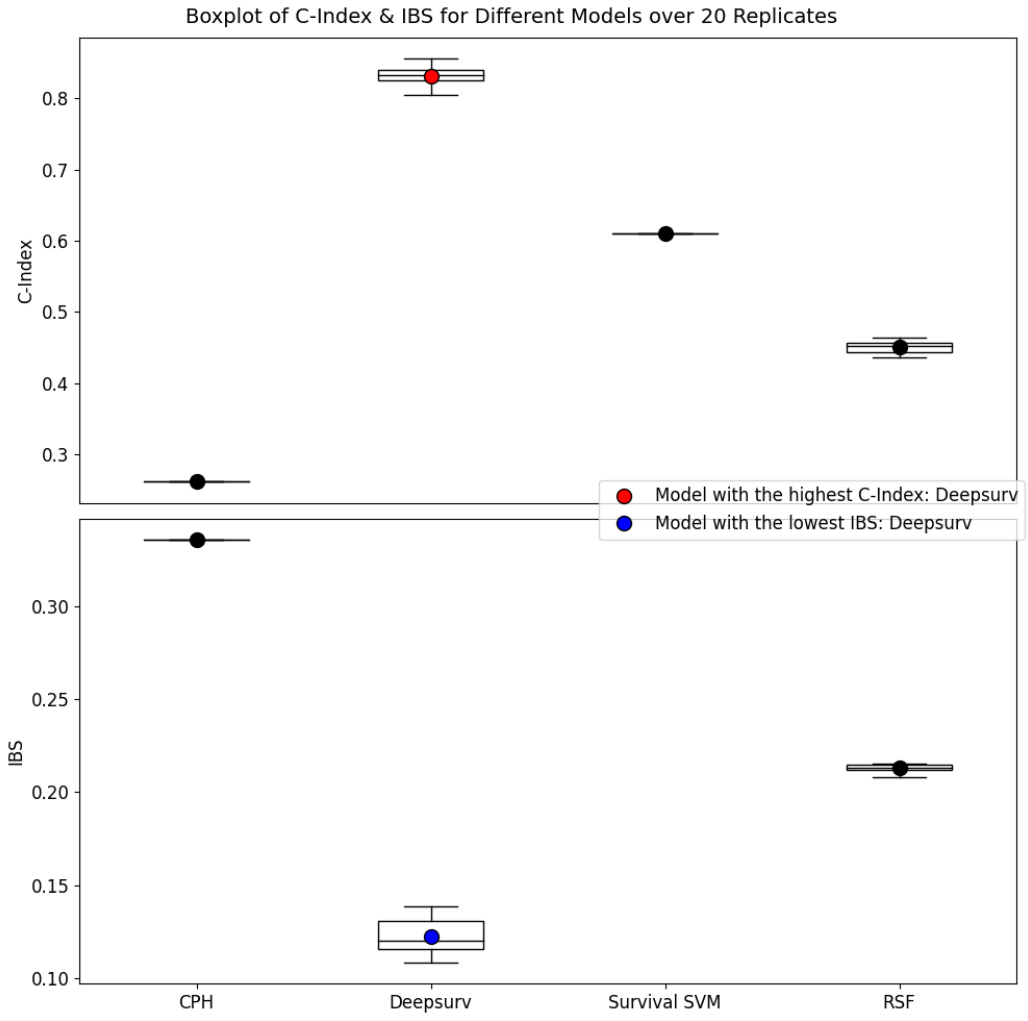


Table 3

Model evaluation results

Models\Metrics	C-index (95% C.I.)	IBS (95% C.I.)
CPH (as baseline)	0.261 (0.261, 0.261)	0.336 (0.336, 0.336)
DeepSurv	0.831 (0.826, 0.837)	0.122 (0.118, 0.126)
Survival SVR	0.611 (0.611, 0.611)	NA
RSF	0.450 (0.447, 0.454)	0.213 (0.212, 0.214)

In detail, the CPH model, utilized as the baseline, demonstrated limited predictive capability, with a C-index of 0.261 (95% C.I.: 0.261, 0.261) and an IBS of 0.336 (95% C.I.: 0.336, 0.336). DeepSurv, however, showcased remarkable predictive accuracy, achieving a C-index of 0.831 (95% C.I.: 0.826, 0.837) and an IBS of 0.122 (95% C.I.: 0.118, 0.126). This performance surpassed that of other ML models evaluated, including Survival SVR and RSF. While Survival SVR showed moderate predictive ability with a C-index of 0.611, its IBS value was unavailable. RSF performed less optimally with a C-index of 0.450 (95% C.I.: 0.447, 0.454) and an IBS of 0.213 (95% C.I.: 0.212, 0.214).

Survival SVR functions without assuming any Cox-type models; however, it then lacks the capability to predict the baseline hazard and hazard function. This limitation is crucial because it impedes the transformation of predictions into survival probabilities. As a result, Survival SVR is unable to compute the IBS, which depends on these probabilities. Despite this inherent limitation within the model's structure, the C-index provides a reliable alternative. This measure does not require estimating the censoring distribution, making it a viable option. Therefore, despite Survival SVR's shortcomings, we can confidently assert that DeepSurv exhibits superior performance compared to other models.

Given the high censoring rate of 86% observed in this study, it is understandable why specific models such as Survival SVR and RSF exhibited suboptimal performance. For Survival SVR, the penalization of most observations occurs when predictions fall below the observed censoring time. In the case of RSF, the constraint imposed on tree growth—that the terminal node must have a certain number of unique deaths—limits

the size of the trees that can be grown, thereby diminishing their predictive power.

An interesting observation from prior interpretations of the correlation matrix highlights the relatively high correlations (greater than 0.2) between numerous engineered features, such as polynomial and interaction terms, and the time-to-event data, compared to other features. These engineered features are exclusively employed in training DeepSurv, indicating their significant role as interpretation tools. Their incorporation suggests that complex interactions and nonlinear effects are crucial in influencing patient survival outcomes. Although the correlations of these features are not extraordinarily high, their inclusion in predictive models like DeepSurv is justified due to their capability to capture nuanced relationships within the data. However, it is critical to note that the utility of these features might be confined to specific modeling frameworks. While beneficial in DeepSurv, their application in other models leads to reduced predictive accuracy.

Table 4

Experimented combinations that generate the best performance for each model

Models	Addressing High Pairwise Correlation	Data Stratification	Dimension Reduction (PCA)	Feature Engineering
CPH (as baseline)	×	√	√	×
DeepSurv	×	√	×	√
Survival SVR	√	×	√	×
RSF	×	√	√	×

Table 4 presents the preprocessing steps experimented on the models to predict outcomes better. For the CPH model, we applied data stratification and PCA to enhance interpretability and efficiency; we did not leverage feature engineering because of its underlying assumption of linear relationships. On the other hand, DeepSurv benefits from feature engineering and data stratification in finding complex patterns and doesn't require the use of PCA, thanks to its automatic feature extraction capability. The Survival SVR model exploits PCA to reduce dimensionality and address high pairwise

correlation, but it doesn't rely much on data stratification. Lastly, RSF uses both data stratification and PCA to improve its generalization ability. It also has less emphasis on feature engineering due to its capability to manage nonlinear relationships. These tailored preprocessing methods are designed to optimize each model's prediction performance.

Discussion

In this section, we discussed the practical implications of DeepSurv particularly in the TCGA breast cancer dataset. DeepSurv, identified as the most effective model, adeptly stratifies patients into high-risk and low-risk groups according to their median survival probabilities.

1. Practical Implications

By looking at the visualizations provided by DeepSurv, taking the survival probabilities as a reference, healthcare providers can improve their patient care strategies. For patients in the low-risk group, clinicians could consider reducing unnecessary medical interventions and treatments. This could mean adopting watchful waiting or less aggressive treatments for certain conditions, which might potentially prevent overdiagnosis or overtreatment. For high-risk patients, however, doctors might want to concentrate on optimized treatment plans and timely medical interventions instead. This situation demands closer monitoring and more aggressive treatment approaches.

Figure 10

Survival probabilities for the low-risk and high-risk groups

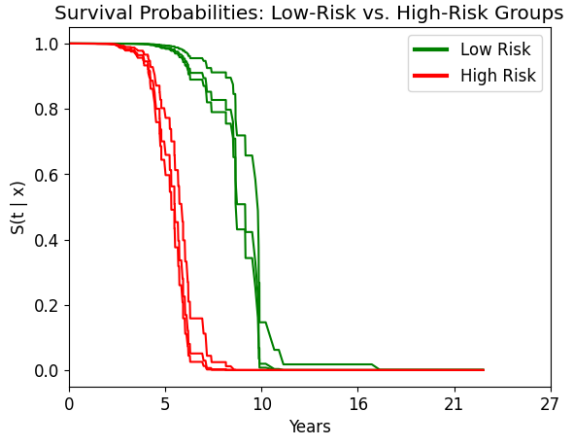
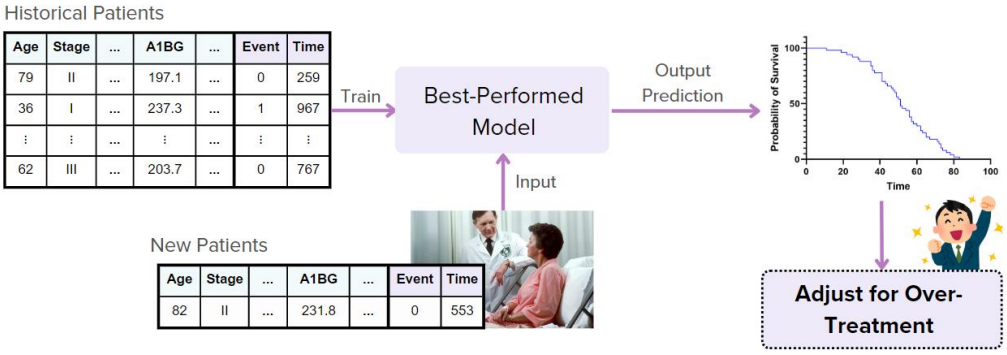


Figure 11 below visually displaces the pipeline of our study, where the ML predictions are integrated into clinical practice. After training the best-performing model (i.e., DeepSurv) with the historical patients' data, whenever there is a new patient comes in, we could output his/her survival distribution. This prediction could then serve as a reference for adjusting treatment strategies accordingly. Such a procedure emphasizes that ML is really significant in making healthcare delivery more precise and effective.

Figure 11

Pipeline of the study



The applications of DeepSurv extend well beyond breast cancer to various cancer types. For instance, Yang et al. (2022) found its utility in non-small cell lung cancer (NSCLC), where DeepSurv outperformed conventional models by incorporating radiomic features into a nomogram. It achieved an impressive C-index of 0.821 in the training and 0.768 in the validation cohorts. Similarly, Wang et al. (2023) saw that DeepSurv was more accurate than the traditional CPH model for treating early hepatocellular carcinoma (HCC); it yielded better discrimination with a C-index of 0.7028. This enhancement in predictive accuracy provides valuable insights for tailoring treatment decisions to individual risk profiles, as Li et al. (2023) further discussed. Not limited to the cancer types studied so far, in the case of gastric adenocarcinoma, Choi et al. (2021) determined that DeepSurv was more accurate over time horizons with a C-index of 0.772 than both the CPH and RSF models. This perspective was corroborated by Zeng et al. (2023). Also, Yang et al. (2022) highlighted that DeepSurv has consistent superiority in prognostic predictions for patients in coronary care units (CCU). It achieved a C-index of 0.833, which further affirmed this model's capacity. These

studies across NSCLC, early HCC, gastric adenocarcinoma, and CCU settings have shown DeepSurv's remarkable predictive accuracy and transformative potential in crafting personalized treatment strategies.

2. Interpretation Challenge

In our study, DeepSurv emerged as the top-performing model, but it's hard to understand or draw interpretations from the results due to the inherent black-box nature of deep learning algorithms. This interpretation challenge hinders its use in real-world clinics. And the correlation matrix we discussed earlier doesn't provide much assistance. Meanwhile, RSF was having trouble computing variable importance with over 20,000 features. To support the practical application of models like DeepSurv, we then explore alternative feature selection methods or interpretable models.

This dilemma highlights the importance of advancements like Autosurv, introduced by Jiang et al. (2024). It is a deep-learning framework that provides interpretability. In this case, the authors conducted a cancer survival analysis by incorporating clinical and multi-omics data, which include gene and miRNA expression data. These allow Autosurv to outperform existing machine learning and deep learning approaches.

Baidoo (2023) explored influential factors in breast cancer patient survival using the RSF ensemble technique and DeepSurv. He used a technique called the Shapley Additive explanation (SHAP). It effectively facilitates the interpretation of variables. Similarly, Moncada-Torres et al. (2021) evaluated ML techniques compared to the standard Cox Proportional Hazards analysis. The study was devoted to survival prediction in breast cancer patients. They found that ML methods can be just as good or even better. Extreme Gradient Boosting (XGB) was especially comparable because it can model complex nonlinearities. They also utilized SHAP to explain how different features affect their predictions, which shows the promising landscape of adopting explainable ML techniques in clinical settings.

Additionally, Hao et al. (2019) proposed Cox-PASNet. It is a biologically interpretable pathway-based sparse deep neural network. This model was tailored for survival analysis with the use of both genomic and clinical data. They built information from databases about biological pathways into the model's architecture, helping the model

find complex patterns that affect patient survival. The model identifies nonlinear and hierarchical associations underlying cancer patient survival quite well. Its final result surpasses benchmarking methods and maintains room for interpretation.

All these studies collectively show how ML methods for predicting survival are evolving in survival prediction. It is important to consider techniques like SHAP or alternative interpretable models in such a context. The purpose is to address interpretability challenges and translate innovations into actionable insights for clinical practice.

2. Future Directions

After the discussion, we started to explore future research directions We first considered alternative ways of dimension reduction and feature selection. The reason is to ignite the predictive abilities of extensive features in the TCGA, which in our current study was not elaborated enough.

Many studies have shown that altering dimension reduction and feature selection methods can make predictions more accurate. For example, Xie et al. (2016) found that combining Random Projection (RP) with other methods like Principal Component Analysis (PCA) and Feature Selection (FS) to improve classification accuracy on genomics data significantly. In 2021, Jiang and Jin proposed a new method for searching mutated and cancer-related genes. Here, the staged feature selection algorithms play an important role in optimal model construction. Bartenhagen et al. (2010) compared PCA with nonlinear methods like Locally Linear Embedding and Isomap for microarray data visualization, highlighting superior performance in capturing underlying data structure. In 2017 and 2019, Alkuhlani and Fan demonstrated that selecting features through multistage feature selection integrating filter and wrapper methods is key to finding reliable genetic evidence in cancer data. More recently, Taghizadeh et al. (2022) showed the effect of detecting breast cancer by selecting features and classifying them using a hybrid ML approach. Rajpoot et al. (2024) also showed that advanced ML optimization techniques have the potential to help predict breast cancer. Here, many researchers choose to integrate methods such as Locally Linear Embedding, Isomap, and hybrid feature selection algorithms into their study. This might enhance model performance and contribute to effective breast cancer

diagnosis and treatment.

In our work, we achieved promising results using Deepsurv to analyze mainly the gene expression data. However, we believe that using more types of omics data, such as copy number variation and miRNA expression, can help us better understand and predict cancer. For example, pioneering studies by Orsini et al. (2023) and Rossi et al. (2022) have begun using these data types to identify new disease biomarkers and therapeutic targets that could transform personalized medicine. Research by Yuan et al. (2019) showed that ML is very useful for understanding complex gene regulatory networks and improving predictions of cancer. Similarly, Y. Fan et al. (2022) demonstrated that combining genomic and transcriptomic patterns can help us understand important genes and the underlying disease mechanisms. Guo et al. (2022) utilized advanced computational models like graph convolutional networks. They could effectively integrate heterogeneous omics data for improved subtype classification and prognostic accuracy. Together, the evolving integration of multi-omics data through ML might also help to enhance the accuracy of survival predictions in our future study.

In addition, it is important to prioritize the research on triple-negative breast cancer (TNBC). This type is more complex and has a higher recurrence rate than other breast cancer subtypes, which has raised the attention of scholars and clinicians. For example, Li et al. (2023) developed a new ML-based immune infiltrating cell (IIC) associated signature (MLIIC). They amalgamated data from multiple transcriptome sources, including purified immune cells and TNBC tissues, their methodology enhanced the predictive accuracy of TNBC prognosis. Similarly, J. Li et al. (2023) contributed to this evolving landscape by establishing a ROS signature (ROSig). It targets the reactive oxygen species in TNBC progression and treatment response. These innovative approaches, combining ML algorithms with prognostic indicators, enhanced risk prediction accuracy and facilitated a deeper understanding of tumor heterogeneity and therapeutic implications.

The discussion highlights the practical significance of DeepSurv's effectiveness in predicting patient survival. We also recognized the challenges related to its interpretability and seek methods to resolve them. More research on different dimension reduction and feature selection methods using different multi-omics data

sources is recommended. We also mentioned the need for more detailed studies of TNBC.

Conclusion

This study evaluated the performance of ML models in survival prediction through the application of TCGA breast cancer data. This study demonstrates the advanced performance of advanced ML technologies. The models inclued DeepSurv, Survival SVR, and RSF, in evaluating traditional statistical models such as CPH. Among all the models we compared, DeepSurv performs the best. It demonstrates the potential of deep learning in capturing complex relationships and improving treatment strategies.

Although the prediction performance is satisfying, challenges still exist, mainly in handling data with high censoring rates. Such a challenge is evident in survival SVR and RSF. In addition, the discussion emphasized the practical significance of DeepSurv. The challenges related to interpretability are acknowledged, which could be addressed using SHAP or other interpretable models.

This study also outlines future research directions, including exploring alternative techniques for dimension reduction and feature selection, integrating multi-omics resources, and focusing on the research of triple-negative cancer. These emphasized the multifaced nature of survival prediction proclems and complexities involved in guiding treatment approaches.

References

Alkuhlani, A., Nassef, M., & Farag, I. (2017). Multistage feature selection approach for high-dimensional cancer data. Soft Computing, 21(22), 6895–6906. https://doi.org/10.1007/s00500-016-2439-9

Alzubaidi, A., Tepper, J., Inden, B., & Lotfi, A. (2022). MRNA biomarkers for invasive breast cancer based on a deep feature selection approach. Journal of Biomedical Research & Environmental Sciences, 3(10), 1163–1176. https://doi.org/10.37871/jbres1572

Baidoo, T. G. (2023). Advanced prognostic modeling for breast cancer patients: Leveraging data-driven approaches for survival analysis. University of Texas Rio Grande Valley.

Bartenhagen, C., Klein, H.-U., Ruckert, C., Jiang, X., & Dugas, M. (2010). Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinformatics, 11(1). https://doi.org/10.1186/1471-2105-11-567

Bhandari, N., Walambe, R., Kotecha, K., & Khare, S. P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. Frontiers in Molecular Biosciences, 9. https://doi.org/10.3389/fmolb.2022.907150

Boehmke, B. (2018, December 5). Decision trees, bagging, & random forests. Github.Io. https://bradleyboehmke.github.io/random-forest-training/slides-source.html

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. https://doi.org/10.3322/caac.21834

Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning–based

multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 24(6), 1248–1259. https://doi.org/10.1158/1078-0432.ccr-17-0853

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Computational Biology, 14(4), e1006076. https://doi.org/10.1371/journal.pcbi.1006076

Divya, P., & Suresh, S. (2024). Bioinformatics analysis in the identification of prognostic signatures for ER-negative breast cancer data. Journal of the Indian Society for Probability and Statistics. https://doi.org/10.1007/s41096-024-00187-8

Dunn, B. K., Woloshin, S., Xie, H., & Kramer, B. S. (2022a). Cancer overdiagnosis: A challenge in the era of screening. Journal of the National Cancer Center, 2(4), 235–242. https://doi.org/10.1016/j.jncc.2022.08.005

Dunn, B. K., Woloshin, S., Xie, H., & Kramer, B. S. (2022b). Cancer overdiagnosis: A challenge in the era of screening. Journal of the National Cancer Center, 2(4), 235–242. https://doi.org/10.1016/j.jncc.2022.08.005

Evangeline I., K., Kirubha, S. P. A., & Precious, J. G. (2023). Survival analysis of breast cancer patients using machine learning models. Multimedia Tools and Applications, 82(20), 30909–30928. https://doi.org/10.1007/s11042-023-14989-8

Fan, S., Tang, J., Tian, Q., & Wu, C. (2019). A robust fuzzy rule based integrative feature selection strategy for gene expression data in TCGA. BMC Medical Genomics, 12(S1). https://doi.org/10.1186/s12920-018-0451-x

Fan, Y., Kao, C., Yang, F., Wang, F., Yin, G., Wang, Y., He, Y., Ji, J., & Liu, L. (2022). Integrated multi-omics analysis model to identify biomarkers associated with prognosis of breast cancer. Frontiers in Oncology, 12. https://doi.org/10.3389/fonc.2022.899900

Fanaee-T, H., & Thoresen, M. (2019). Performance evaluation of methods for integrative dimension reduction. Information Sciences, 493, 105–119.

https://doi.org/10.1016/j.ins.2019.04.041

Fotso, S. (2019). PySurvival: Open source package for survival analysis modeling. https://www. pysurvival. io.

Guo, H., Lv, X., Li, Y., & Li, M. (2022). Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification. In bioRxiv (p. 2022.09.05.506572). https://doi.org/10.1101/2022.09.05.506572

Hao, J., Kim, Y., Mallavarapu, T., Oh, J. H., & Kang, M. (2019). Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. BMC Medical Genomics, 12(S10). https://doi.org/10.1186/s12920-019-0624-2

He, Z., Zhang, J., Yuan, X., Xi, J., Liu, Z., & Zhang, Y. (2019). Stratification of breast cancer by integrating gene expression data and clinical variables. Molecules (Basel, Switzerland), 24(3), 631. https://doi.org/10.3390/molecules24030631

Henao, J. D., Lauber, M., Azevedo, M., Grekova, A., Theis, F., List, M., Ogris, C., & Schubert, B. (2023). Multi-omics regulatory network inference in the presence of missing data. Briefings in Bioinformatics, 24(5), bbad309. https://doi.org/10.1093/bib/bbad309

Henderi, H. (2021). Comparison of min-max normalization and Z-score normalization in the K-nearest neighbor (kNN) algorithm to test the accuracy of types of breast cancer. IJIIS: International Journal of Informatics and Information Systems, 4(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

Hutter, C., & Zenklusen, J. C. (2018). The cancer genome atlas: Creating lasting value beyond its data. Cell, 173(2), 283–285. https://doi.org/10.1016/j.cell.2018.03.042

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

Jiang, L., Xu, C., Bai, Y., Liu, A., Gong, Y., Wang, Y.-P., & Deng, H.-W. (2024). Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. Npj Precision Oncology, 8(1), 1–16. https://doi.org/10.1038/s41698-023-00494-6

Jiang, Q., & Jin, M. (2021). Feature selection for breast cancer classification by integrating somatic mutation and gene expression. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.629946

Kale, M. S., & Korenstein, D. (2018). Overdiagnosis in primary care: framing the problem and finding solutions. BMJ (Clinical Research Ed.), 362, k2820. https://doi.org/10.1136/bmj.k2820

Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T., & Kluger, Y. (2016). DeepSurv: Personalized treatment recommender system using A Cox proportional hazards deep neural network. In arXiv [stat.ML]. http://arxiv.org/abs/1606.00931

Kim, D. H., & Lee, K. E. (2022). Discovering breast cancer biomarkers candidates through mRNA expression analysis based on The Cancer Genome Atlas database. Journal of Personalized Medicine, 12(10), 1753. https://doi.org/10.3390/jpm12101753

Kim, W.-J., Choi, B. R., Noh, J. J., Lee, Y.-Y., Kim, T.-J., Lee, J.-W., Kim, B.-G., & Choi, C. H. (2024). Comparison of RNA-Seq and microarray in the prediction of protein expression and survival prediction. Frontiers in Genetics, 15. https://doi.org/10.3389/fgene.2024.1342021

Leung, K. L., Verma, D., Azam, Y. J., & Bakker, E. (2020). The use of multi-omics data and approaches in breast cancer immunotherapy: a review. Future Oncology (London, England), 16(27), 2101–2119. https://doi.org/10.2217/fon-2020-0143

Li, J., Liang, Y., Zhao, X., & Wu, C. (2023). Integrating machine learning algorithms to systematically assess reactive oxygen species levels to aid prognosis and novel treatments for triple -negative breast cancer patients. Frontiers in Immunology, 14.

https://doi.org/10.3389/fimmu.2023.1196054

Li, S., Zhang, N., Zhang, H., Zhou, R., Li, Z., Yang, X., Wu, W., Li, H., Luo, P., Wang, Z., Dai, Z., Liang, X., Wen, J., Zhang, X., Zhang, B., Cheng, Q., Zhang, Q., & Yang, Z. (2023). Artificial intelligence learning landscape of triple-negative breast cancer uncovers new opportunities for enhancing outcomes and immunotherapy responses. Journal of Big Data, 10(1). https://doi.org/10.1186/s40537-023-00809-1

Li, Xianguo, Bao, H., Shi, Y., Zhu, W., Peng, Z., Yan, L., Chen, J., & Shu, X. (2023). Machine learning methods for accurately predicting survival and guiding treatment in stage I and II hepatocellular carcinoma. Medicine, 102(45), e35892. https://doi.org/10.1097/md.0000000000035892

Liñares-Blanco, J., Pazos, A., & Fernandez-Lozano, C. (2021). Machine learning analysis of TCGA cancer data. PeerJ. Computer Science, 7(e584), e584. https://doi.org/10.7717/peerj-cs.584

Liu, J.-X., Gao, Y.-L., Zheng, C.-H., Xu, Y., & Yu, J. (2016). Block-constraint robust principal component analysis and its application to integrated analysis of TCGA data. IEEE Transactions on Nanobioscience, 15(6), 510–516. https://doi.org/10.1109/tnb.2016.2574923

Liu, Q., & Hu, P. (2019). Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer. Cancers, 11(4), 494. https://doi.org/10.3390/cancers11040494

Liu, W., Payne, S. H., Ma, S., & Fenyö, D. (2019). Extracting pathway-level signatures from proteogenomic data in breast cancer using independent component analysis. Molecular & Cellular Proteomics: MCP, 18(8), S169–S182. https://doi.org/10.1074/mcp.tir119.001442

Mahin, K. F., Robiuddin, M., Islam, M., Ashraf, S., Yeasmin, F., & Shatabda, S. (2022). PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. Genomics, 114(2), 110264.

https://doi.org/10.1016/j.ygeno.2022.01.001

Maza, E., Frasse, P., Senin, P., Bouzayen, M., & Zouine, M. (2013). Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. Communicative & Integrative Biology, 6(6), e25849. https://doi.org/10.4161/cib.25849

Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific Reports, 11(1), 1–13. https://doi.org/10.1038/s41598-021-86327-7

Murugesan, V., & Balamurugan, P. (2022). Weighted fuzzy score normalization and Bayesian independent principal component analysis imputation for breast cancer gene expression analysis. https://doi.org/10.22266/ijies2022.0630.08

Orsini, A., Diquigiovanni, C., & Bonora, E. (2023). Omics technologies improving breast cancer research and diagnostics. International Journal of Molecular Sciences, 24(16), 12690. https://doi.org/10.3390/ijms241612690

Pei, J., Wang, Y., & Li, Y. (2020). Identification of key genes controlling breast cancer stem cell characteristics via stemness indices analysis. Journal of Translational Medicine, 18(1). https://doi.org/10.1186/s12967-020-02260-9

Pölsterl, S., Navab, N., & Katouzian, A. (2015). Fast training of support vector machines for survival analysis. In Machine Learning and Knowledge Discovery in Databases (pp. 243–259). Springer International Publishing.

Rajpoot, C. S., Sharma, G., Gupta, P., Dadheech, P., Yahya, U., & Aneja, N. (2024). Feature selection-based machine learning comparative analysis for predicting breast cancer. Applied Artificial Intelligence: AAI, 38(1). https://doi.org/10.1080/08839514.2024.2340386

Rossi, C., Cicalini, I., Cufaro, M. C., Consalvo, A., Upadhyaya, P., Sala, G., Antonucci,

I., Del Boccio, P., Stuppia, L., & De Laurenzi, V. (2022). Breast cancer in the era of integrating “Omics” approaches. Oncogenesis, 11(1), 17. https://doi.org/10.1038/s41389-022-00393-8

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1), 128–138. https://doi.org/10.1097/ede.0b013e3181c30fb2

Strelcenia, E., & Prakoonwit, S. (2023). Effective feature engineering and classification of breast cancer diagnosis: A comparative study. BioMedInformatics, 3(3), 616–631. https://doi.org/10.3390/biomedinformatics3030042

Sun, D., Wang, M., Feng, H., & Li, A. (2017). Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: Supervised feature extraction and classification for breast cancer prognosis prediction. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1–5.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660

Tan, J., Ung, M., Cheng, C., & Greene, C. S. (2014). Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Biocomputing 2015.

Viñas, R., Azevedo, T., Gamazon, E. R., & Liò, P. (2021). Deep learning enables fast and accurate imputation of gene expression. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.624128

Wang, J., Qin, D., Ye, L., Wan, L., Wang, F., Yang, Y., Ma, Y., Yang, H., Yang, Z., Chen, M., Jiang, W., & Zhang, Q. (2022). CCL19 has potential to be a potential

prognostic biomarker and a modulator of tumor immune microenvironment (TIME) of breast cancer: a comprehensive analysis based on TCGA database. Aging, 14(9), 4158–4175. https://doi.org/10.18632/aging.204081

Wang, Y., Chen, Q., Shao, H., Zhang, R., & Shen, H. (2024). Generating bulk RNA-Seq gene expression data based on generative deep learning models and utilizing it for data augmentation. Computers in Biology and Medicine, 169(107828), 107828. https://doi.org/10.1016/j.compbiomed.2023.107828

Xie, H., Li, J., Zhang, Q., & Wang, Y. (2016). Comparison among dimensionality reduction techniques based on Random Projection for cancer classification. Computational Biology and Chemistry, 65, 165–172. https://doi.org/10.1016/j.compbiolchem.2016.09.010

Yang, B., Liu, C., Wu, R., Zhong, J., Li, A., Ma, L., Zhong, J., Yin, S., Zhou, C., Ge, Y., Tao, X., Zhang, L., & Lu, G. (2022). Development and validation of a DeepSurv nomogram to predict survival outcomes and guide personalized adjuvant chemotherapy in non-small cell lung cancer. Frontiers in Oncology, 12. https://doi.org/10.3389/fonc.2022.895014

Yang, R., Huang, T., Wang, Z., Huang, W., Feng, A., Li, L., & Lyu, J. (2021). Deep-learning-based survival prediction of patients in coronary care units. Computational and Mathematical Methods in Medicine, 2021, 1–10. https://doi.org/10.1155/2021/5745304

Yuan, L., Guo, L.-H., Yuan, C.-A., Zhang, Y., Han, K., Nandi, A. K., Honig, B., & Huang, D.-S. (2019). Integration of multi-omics data for gene regulatory network inference and application to breast cancer. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(3), 782–791. https://doi.org/10.1109/tcbb.2018.2866836

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

Zhang, D., Yang, S., Li, Y., Yao, J., Ruan, J., Zheng, Y., Deng, Y., Li, N., Wei, B., Wu, Y., Zhai, Z., Lyu, J., & Dai, Z. (2020). Prediction of overall survival among female patients with breast cancer using a prognostic signature based on 8 DNA repair–related genes. JAMA Network Open, 3(10), e2014622. https://doi.org/10.1001/jamanetworkopen.2020.14622

Zhang, Y., Yang, W., Li, D., Yang, J. Y., Guan, R., & Yang, M. Q. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. BMC Medical Genomics, 11(S5). https://doi.org/10.1186/s12920-018-0419-x

Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., & Ma, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Briefings in Bioinformatics, 16(2), 291–303. https://doi.org/10.1093/bib/bbu003

Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. Scientific Reports, 10(1), 1–11. https://doi.org/10.1038/s41598-020-72664-6

Zhong, S., Jia, Z., Zhang, H., Gong, Z., Feng, J., & Xu, H. (2021). Identification and validation of tumor microenvironment-related prognostic biomarkers in breast cancer. Translational Cancer Research, 10(10), 4355–4364. https://doi.org/10.21037/tcr-21-1248

Zhu, C., Zhang, S., Liu, D., Wang, Q., Yang, N., Zheng, Z., Wu, Q., & Zhou, Y. (2021). A novel gene prognostic signature based on differential DNA methylation in breast cancer. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.742578

Zhu, X., Wang, J., Sun, B., Ren, C., Yang, T., & Ding, J. (2021). An efficient ensemble method for missing value imputation in microarray gene expression data. BMC Bioinformatics, 22(1). https://doi.org/10.1186/s12859-021-04109-4