Investigating Machine Learning Methods for Survival Prediction

--An Application to TCGA Breast Cancer Data

Presented by: Ziyu Wang (Group A-4)
Supervisor: Dr. Y. Gu

STAT3799 Directed Studies in Statistics (2023-24 2nd Semester)

Table of contents

01
Introduction

Breast Cancer Statistics Overview (2022)

1st

Most Prevalent Cancer in Women

2.3M

New Cases

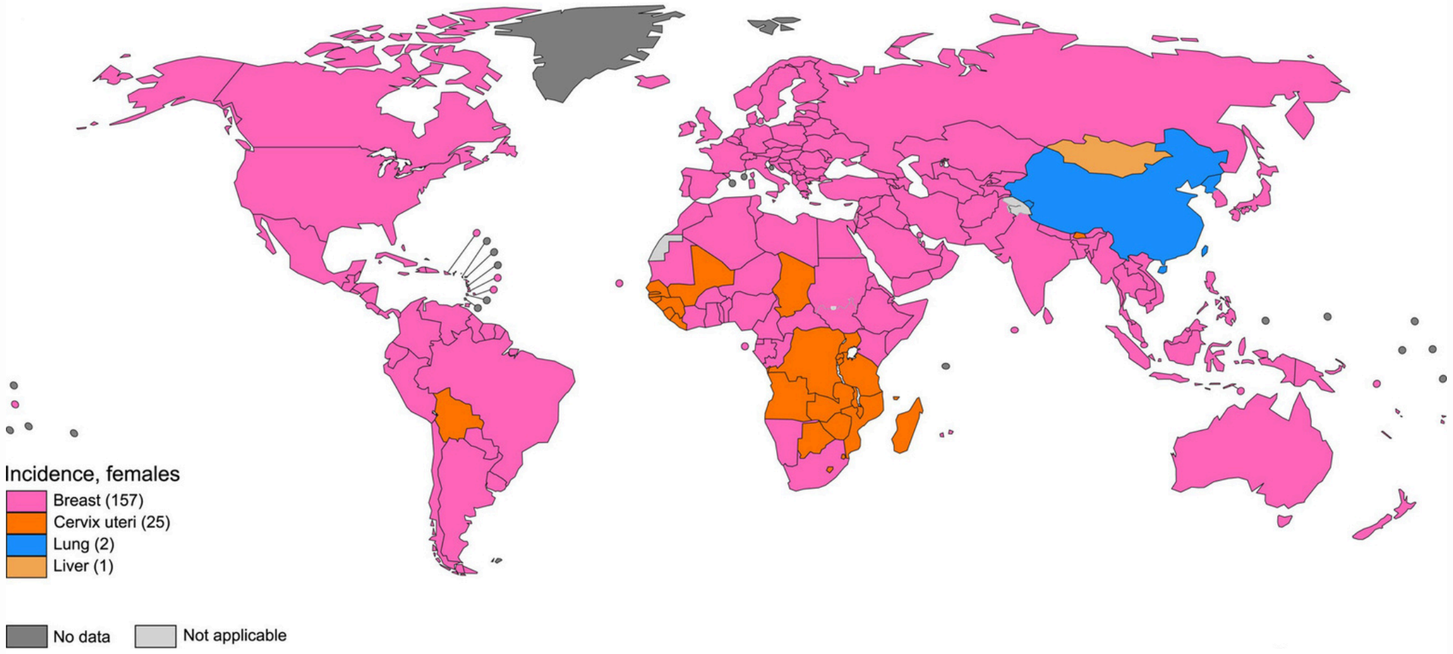
4th

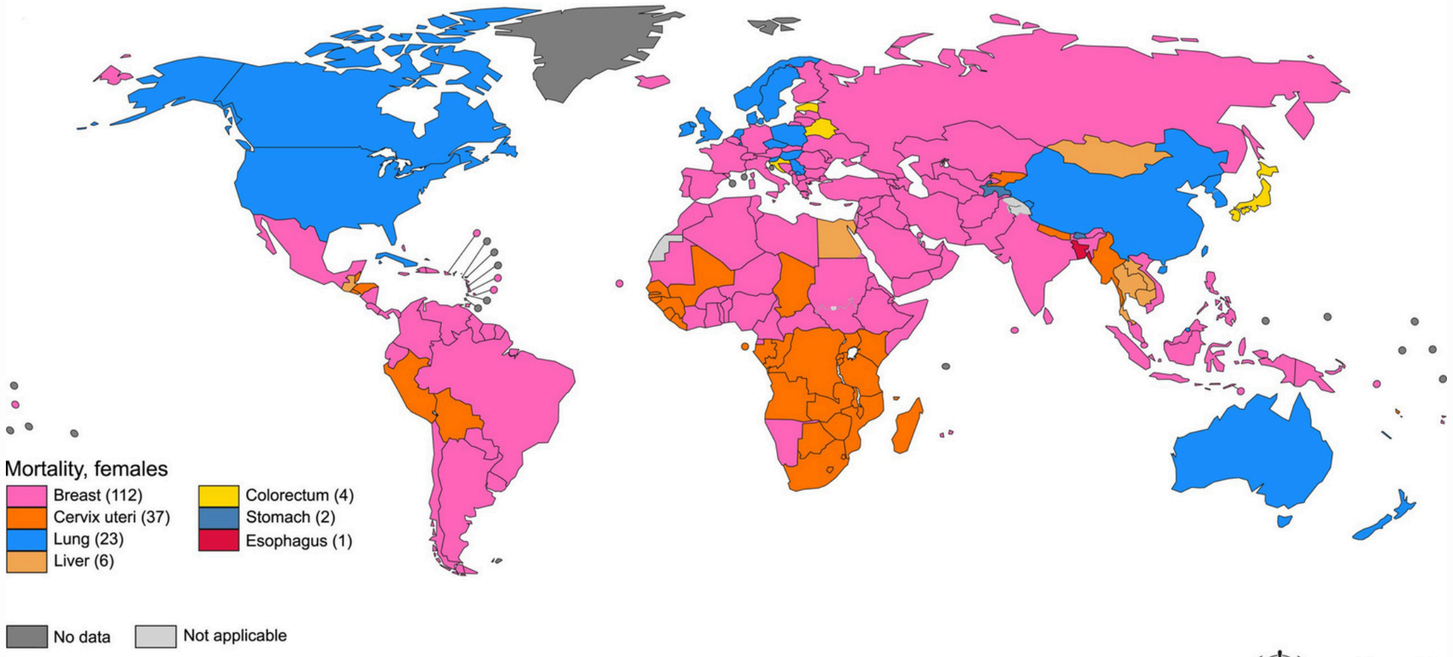
Leading Cause of Cancer Mortality

666K

Deaths

(Bray et al., 2024)





Data source: Globocan 2022
Map production: IARC
World Health Organization

World Health Organization

Ranking **1st** in:

157

countries **(Incidence)**

112

countries **(Mortality)**

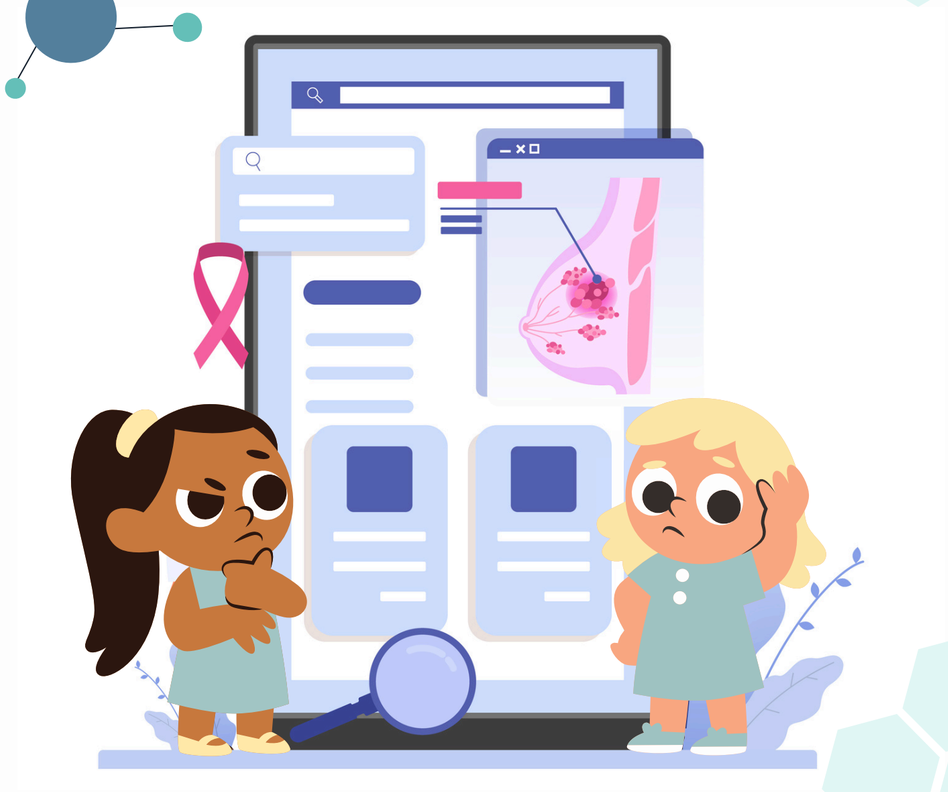
Out of 185 countries

Problem Statement

- **Current approach:**
  - Population-wide breast cancer screening programs

- **Limitations:**
  - Overdiagnosis & overtreatment
  - Rates of overdiagnosis: **0-54%**

-> Need for more targeted approaches



Objectives

To Compare

machine learning models for predicting survival in breast cancer cases with survival outcomes

To Explore

adjusting therapeutic interventions based on survival predictions

Study Pipeline

Historical Patients

Age	Stage	...	A1BG	...	Event	Time
79	II	...	197.1	...	0	259
36	I	...	237.3	...	1	967
⋮	⋮	...	⋮	...	⋮	⋮
62	III	...	203.7	...	0	767

Train

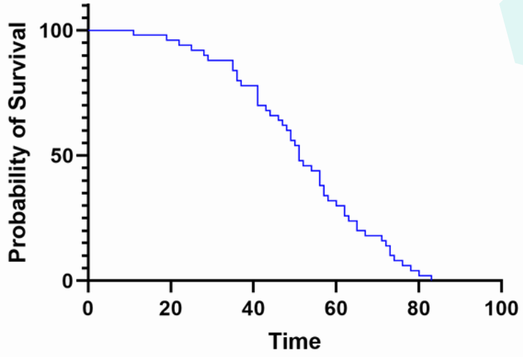
Best-Performed Model

Input

New Patients

Age	Stage	...	A1BG	...	Event	Time
82	II	...	231.8	...	0	553

Output Prediction



Adjust for Over-Treatment

02
Data Exploration

Data Exploration

Data Filtering

TCGA-BRCA Dataset
(n = 1,098)

Primary Tumor Patients with mRNASeq Data
(n = 1,093)

Patients with Complete Survival Data
(n = 1079)

→

Data Processing

Aggregate with clinical and biospecimen data

Data Quality Control

Imputation (KNN)

One-hot encoding

Min-Max normalization

→

Dataframe Ilustration

Clinical & Biospecimen
(p = 96)

Gene Expression
(p = 20217)

Time-to-Event Data

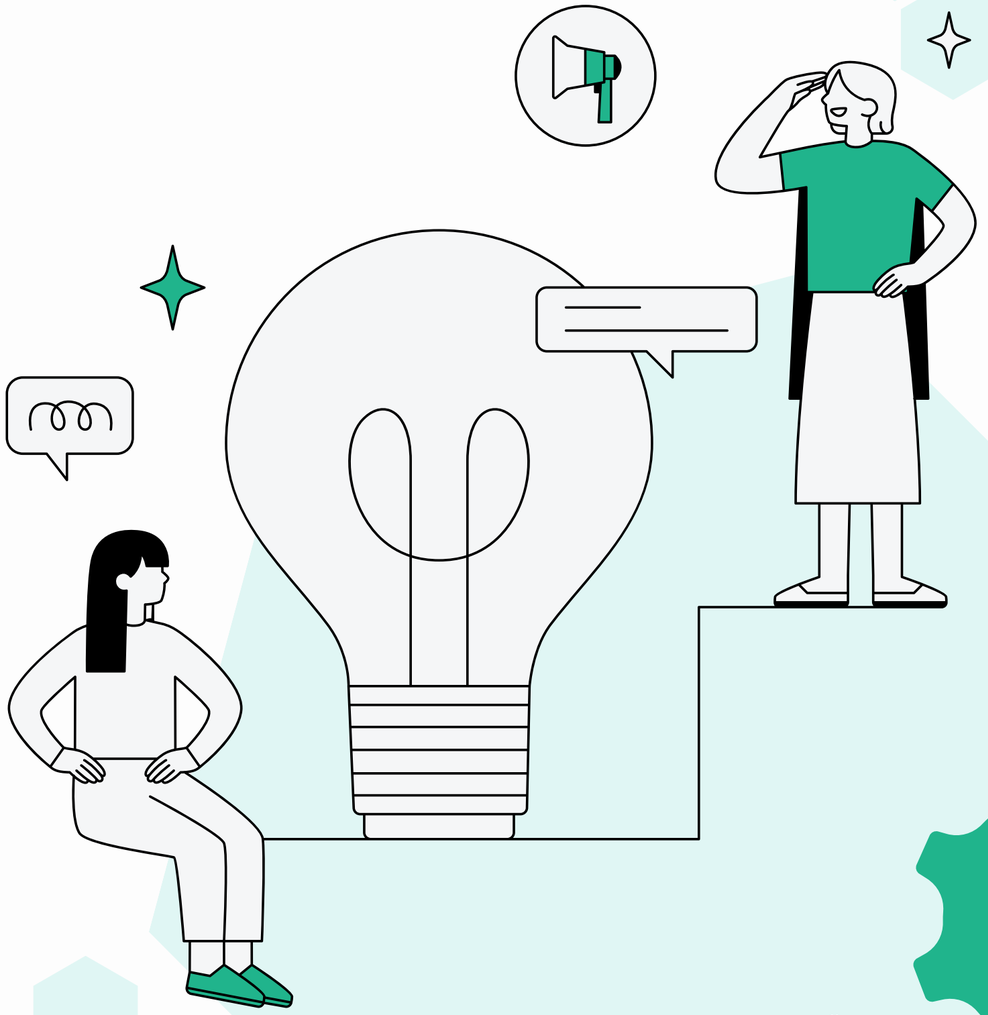
Data Source: Breast invasive carcinoma data from TCGA (The Cancer Genome Atlas)

Data Exploration

Model-specific Considerations

- Remove pariwise correlation
- Data Stratification
- Dimension Reduction (PCA)
- Feature Engineering
  - Example: age group, lymph node ratio, TNBC status
  - Polynomial, spline, and interaction terms

03
Methods



Methods

1. Deep Survival Analysis (DeepSurv)

- Log Partial Likelihood[3]:

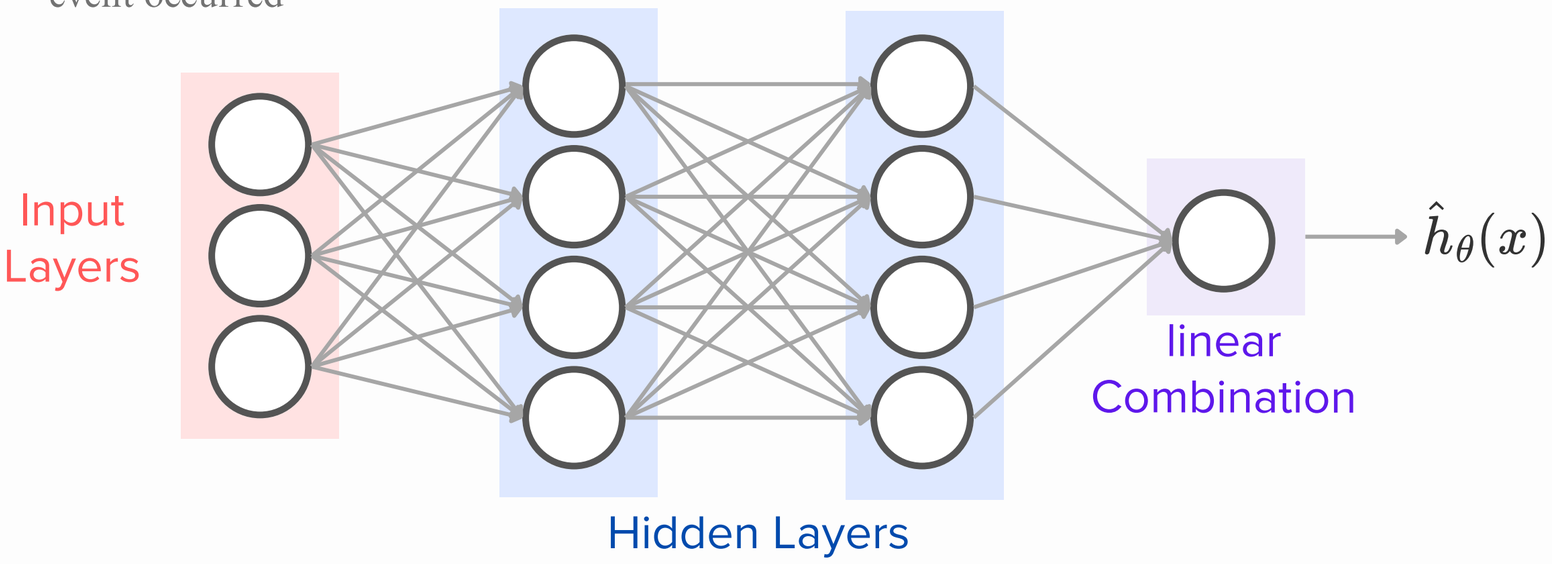
log-risk function

risk score r(x)

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

# of patients with event occurred

l_2 regularization parameter



Methods

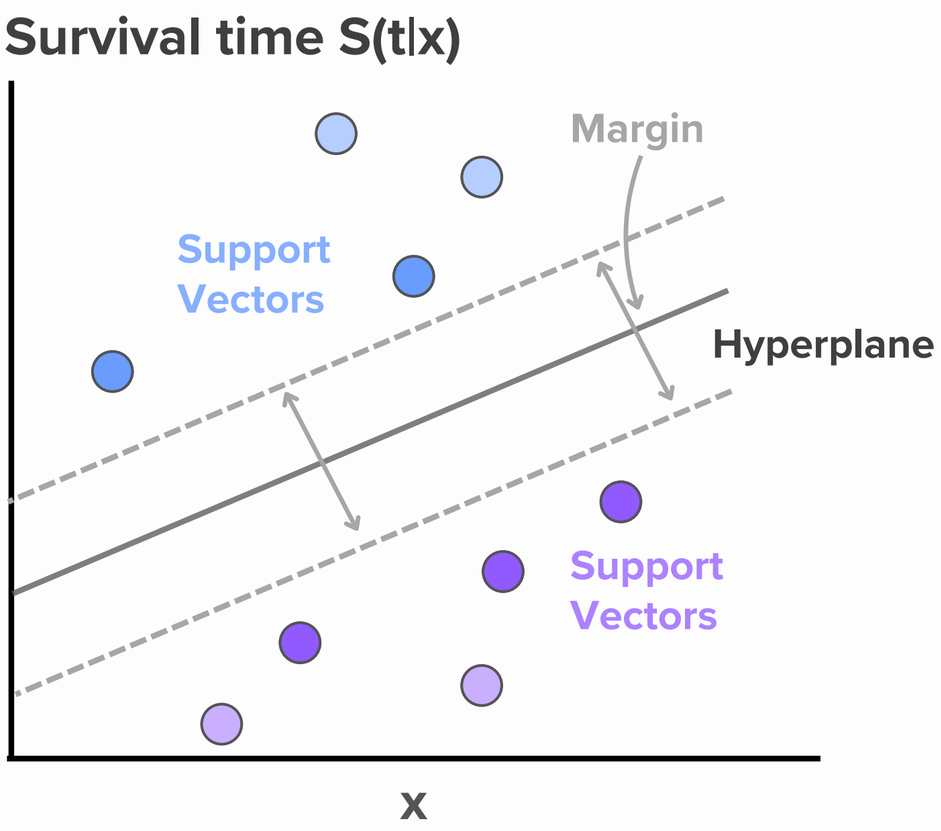
2. Survival Support Vector Regression (Survival SVR)

- Objective function[4]:

regulation parameter

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

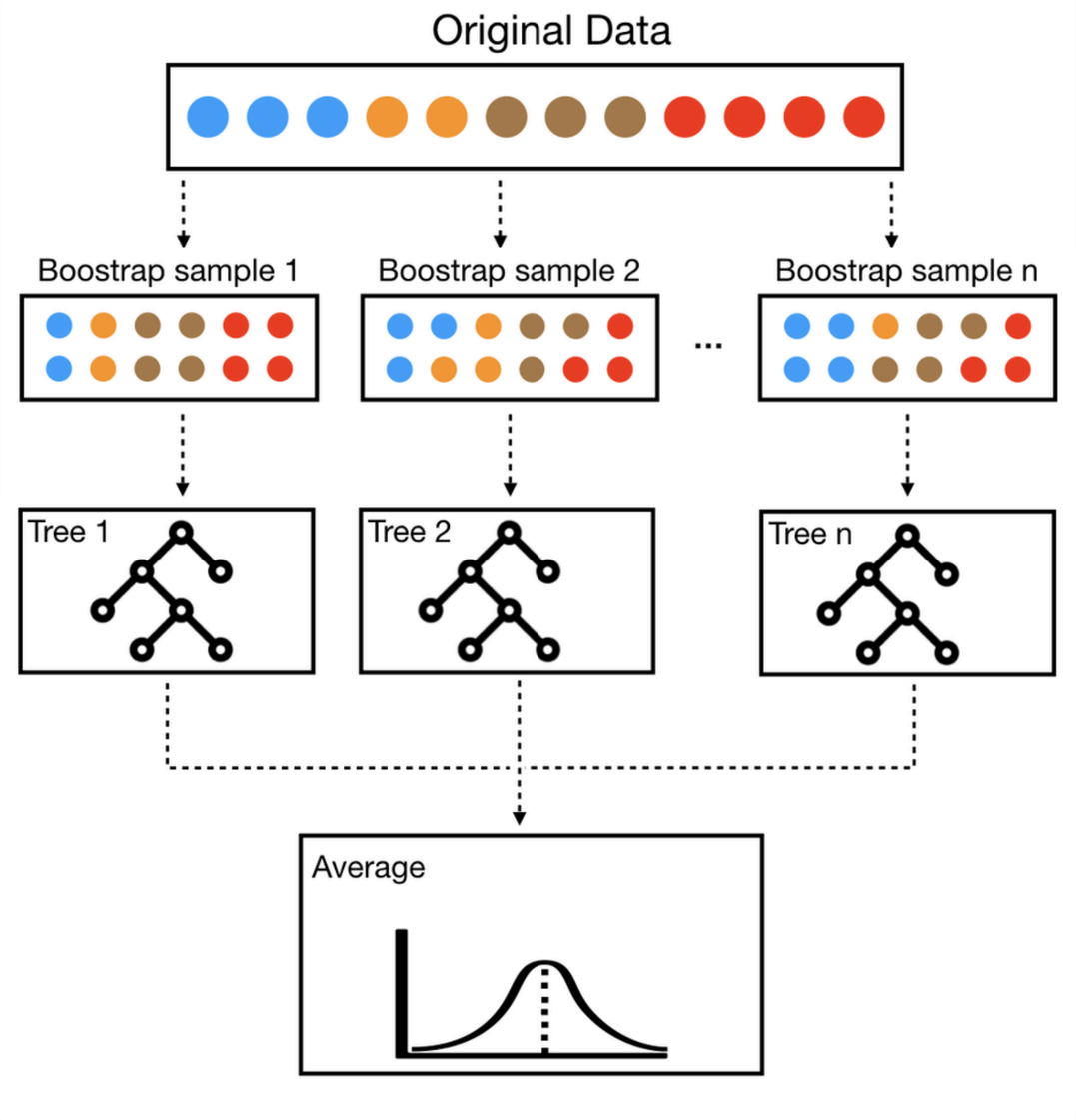
$$\zeta_{w,b}(y_i, x_i, \delta_i) = \begin{cases} max\left(0, y_i - (w^T x_i + b)\right) & if \delta_i = 0, \\ y_i - (w^T x_i + b) & if \delta_i = 1, \end{cases}$$



Methods

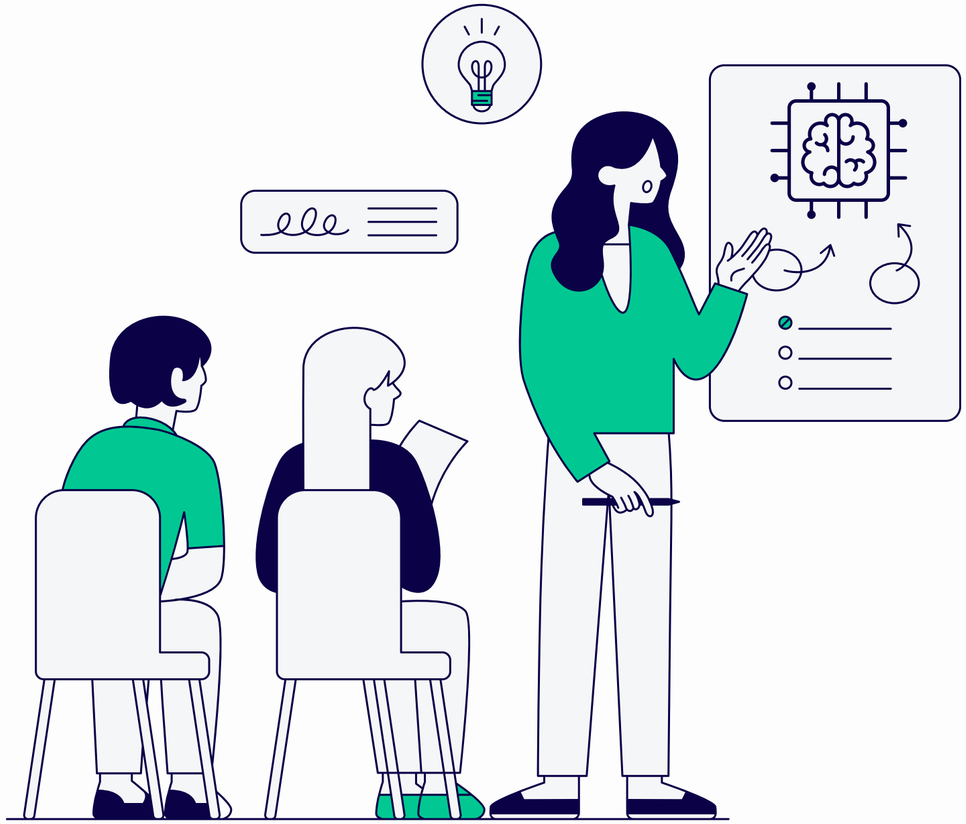
3. Random Survival Forest (RSF) [2]

- **Bootstrap B** samples, each excluding 37% as out-of-bag (OOB) data.
- **Grow survival trees** to the full size**:**
  - Randomly select p variables at each node.
  - Perform log-rank splits.
  - Constraint:
    - Terminal nodes must have at least d0 unique deaths
- **Predict** terminal node CHF values.
- **Average** cumulative hazard function (CHF) from all trees.
- **Evaluate** prediction error using OOB data.



04

Model Training & Evaluation



Model Training & Evaluation

- **Perform Train-Test Split** (80% for training and 20% for testing).
- **Conduct Hyperparameter Tuning:**
  - Utilize 5 repeats of 5-fold cross-validation, primarily through grid search.
  - Employ random search specifically for DeepSurv.
- **Repeat model evaluation 20 times** to account for the uncertainty of results.

Evaluation Metrics

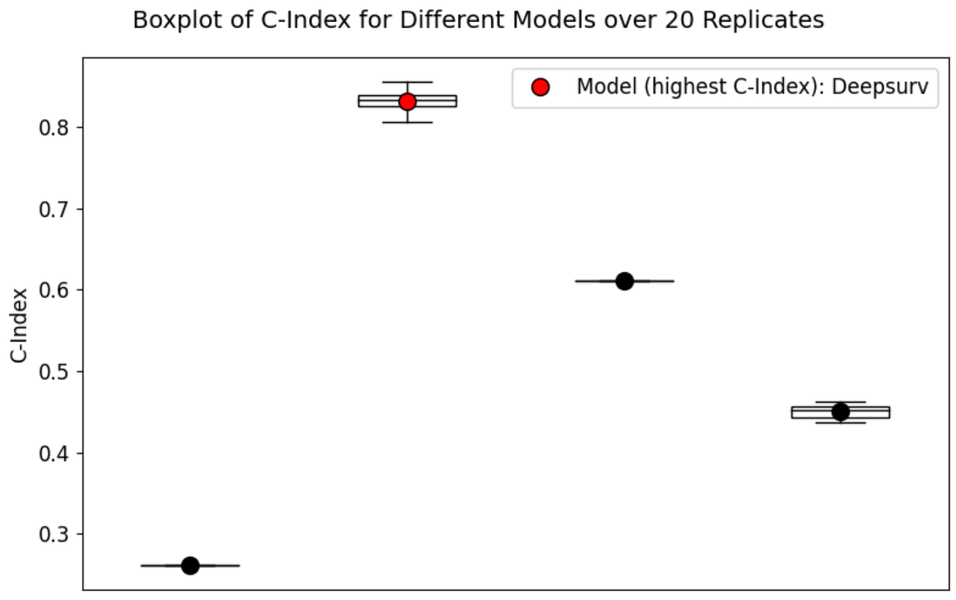
- **Concordance index (C-index):**
  - Measures the ability to rank individuals by survival times correctly.
  - **Higher** values (closer to 1) indicate **better** predictive performance[5].

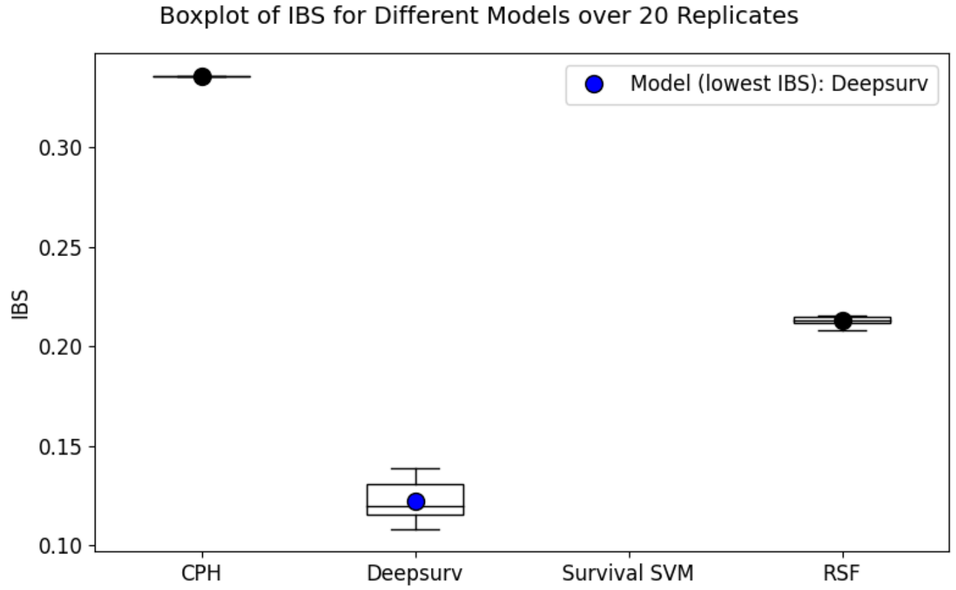
$$\text{C-index} = \frac{\sum_{i,j} I(T_j < T_i) \cdot I(r_j > r_i) \cdot \delta_j}{\sum_{i,j} I(T_j < T_i) \cdot \delta_j}$$

- **Integrated Brier Score (IBS):**
  - Reflects overall model accuracy and calibration.
  - **Lower** values (closer to 0) signify **better** performance[5].

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

05
Results & Discussion





Results & Discussion

- **DeepSurv** demonstrated **superior** predictive performance
  - Highest C-index and Lowest IBS
- All ML models **outperformed** the baseline model, CPH

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 SVM	0.611 (0.611, 0.611)	NA
RSF	0.450 (0.447, 0.454)	0.213 (0.212, 0.214)

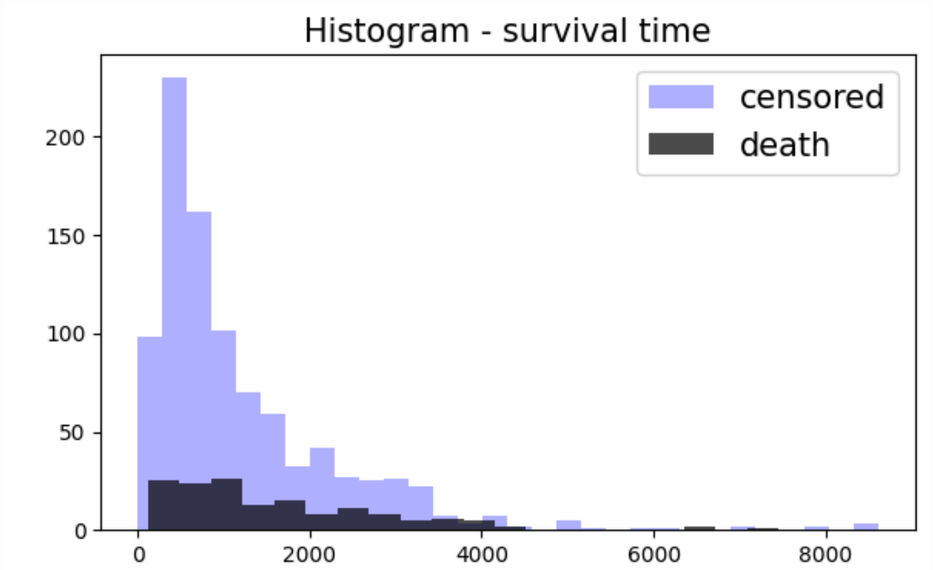
Results & Discussion

- **Limitations of Survival SVR**
  - Doesn't assume a Cox-type model
  - -> Unable to:
    - Predict baseline hazard or hazard function
    - Access survival probability
  - -> Incapable of obtaining IBS measure

- However, C-index provides a **reliable alternative**
  - No need for estimating censoring distribution

Results & Discussion

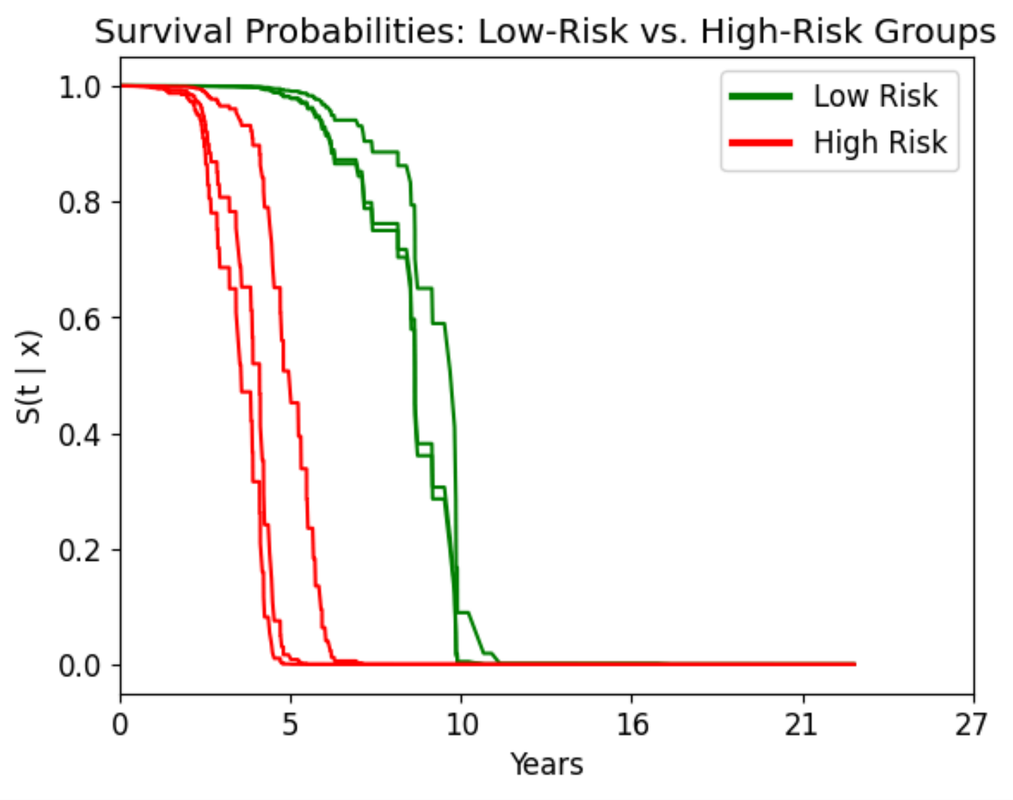
- **High censoring rate (86%)**
  - Common challenge for survival analysis/breast cancer studies
  - **Impact:** Survival SVR and RSF performed less optimally
    - **Survival SVR:** Most observations are only penalized for predictions less than the observed censoring time, impacting performance.
    - **RSF:** Trees can't grow to full potential due to constraints on terminal node deaths, reducing predictive power.



Results & Discussion

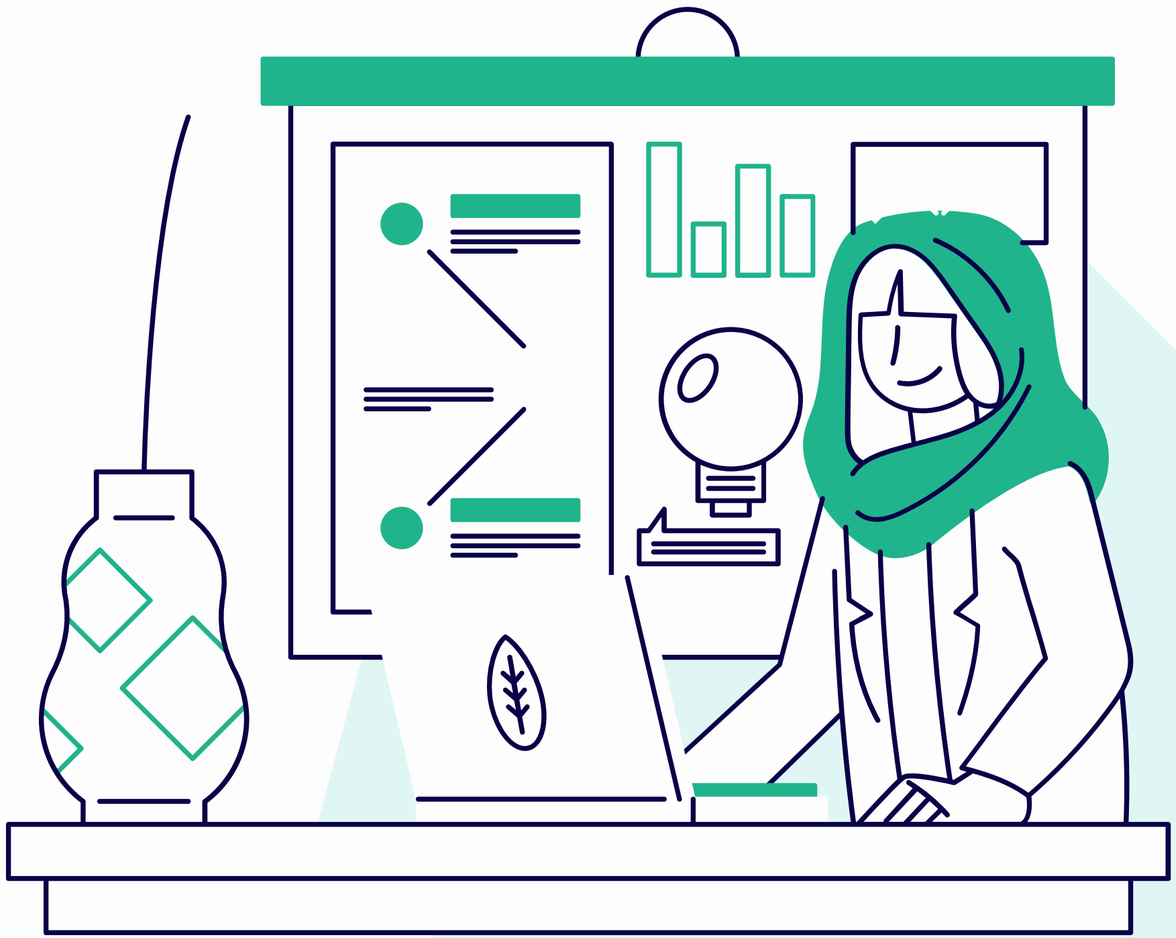
- **Comment on DeepSurv:**
  - **Deep Learning Architecture:**
    - Enables extraction of complex relationships
    - Benefits from a variety of engineered features
  - **Promising** performance even with high censoring rates

Results & Discussion



- **DeepSurv** as the optimal model
  - Survival prediction & Visualization of Risk Groups
- **Clinical Insights:**
  - Low-risk Group:
    - Employ watchful waiting or less aggressive treatments.
  - High-risk Group:
    - Optimize treatment plans and ensure timely interventions.

05
Conclusion



Conclusion

- **Insights from ML in SA for TCGA Breast Cancer Data:**
  - DeepSurv outperformed other models, highlighting deep learning's potential
  - Effective handling of complex survival data, aiding in therapeutic treatment

- **Future Directions:**
  - Alternative dimension reduction methods
  - Unique challenges of subtypes like TNBC
  - Multi-omics data for refined predictive models

- **Overall Implication:**
  - Data-driven ML approaches in SA enhance breast cancer management.

References & Acknowledgement

1. 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
2. Boehmke, B. (2018, December 5). Decision trees, bagging, & random forests. Github.Io. https://bradleyboehmke.github.io/random-forest-training/slides-source.html
3. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. https://doi.org/10.3322/caac.21834
4. Dunn, B. K., Woloshin, S., Xie, H., & Kramer, B. S. (2022). 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
5. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-aoas169
6. 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
7. 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. https://doi.org/10.48550/ARXIV.1606.00931
8. 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.
9. 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
10. 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

I want to express my sincere gratitude to Dr. Y. Gu, my supervisor, for her invaluable guidance and unwavering support throughout this endeavor. Additionally, I am deeply grateful to the researchers and contributors involved in the TCGA breast cancer dataset for their efforts in generating and sharing this valuable data, which formed the foundation of my research.

Thank you!

Do you have any questions?