EAC-ISBA Conference 2024 Poster No.:13

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

Ziyu Wang
Supervisor: Dr. Y. Gu, Department of Statistics & Actuarial Science

Introduction

Motivation:[2,3]

1st
Most Prevalent Cancer in Women (2.3M New Cases)

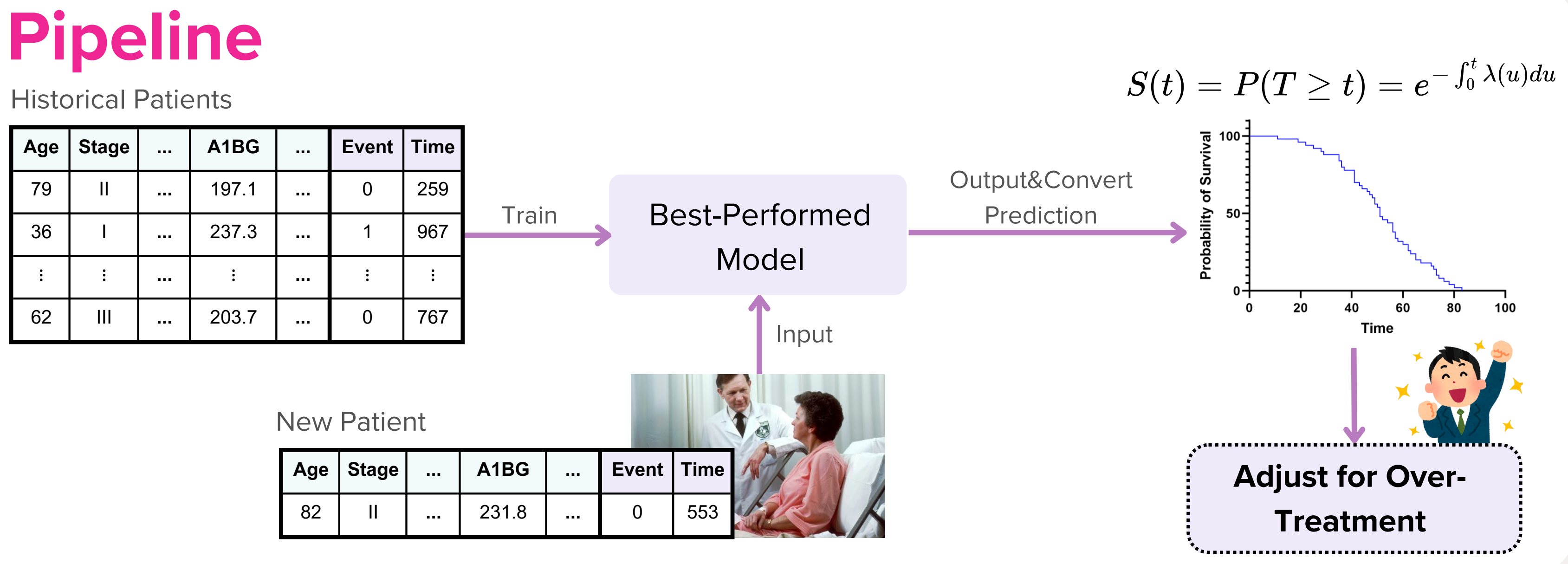
4th
Leading Cause of Cancer Mortality (666K Deaths)

0-54%
Rates of Overdiagnosis

Objectives:

- **To compare** ML models for predicting survival in breast cancer cases
- To explore **adjusting therapeutic interventions** based on survival predictions

Pipeline



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

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

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

Data & Methods

Breast invasive carcinoma data **sourced** from TCGA (The Cancer Genome Atlas)

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 & biospecimen data
- Data Quality Control
- Imputation (KNN)
- One-hot encoding
- Min-Max normalization

Dataframe Illustration

Clinical & Biospecimen (p = 96)

Gene Expression (p = 20217)

Time-to-Event Data

Feature Engineering

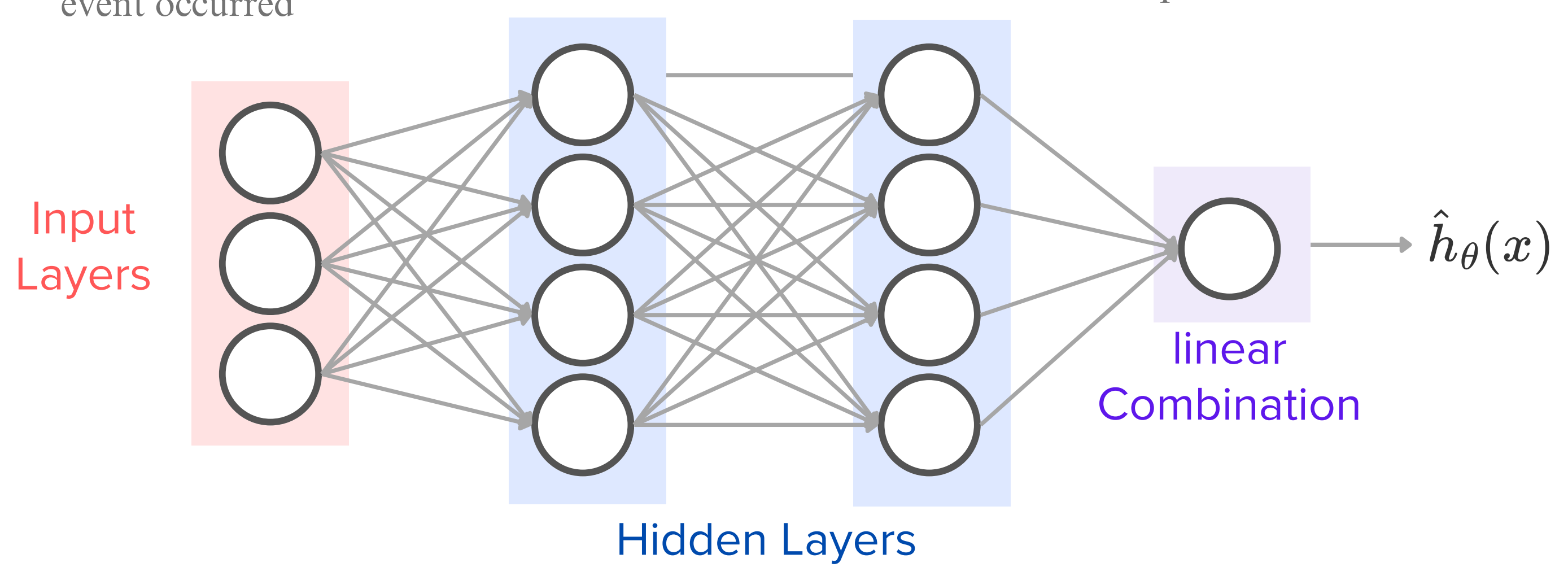
E.g., Age Group, Lymph Node Ratio, TNBC Status; Polynomial, spline, and interaction terms

Remove pariwise correlation **Data Stratification** **Dimension Reduction (PCA)**

1. Deep Survival Analysis (DeepSurv)

- Log Partial Likelihood[5]:

$$l(\theta) = -\frac{1}{\underbrace{N_{E=1}}_{\text{\# of patients with event occurred}}} \sum_{i:E_i=1} \Big(\overbrace{\hat{h}_\theta(x_i)}^{\text{log-risk function}} - log \sum_{j \in \Re(T_i)} \overbrace{e^{\hat{h}_\theta(x_j)}}^{\text{risk score r(x)}}\Big) + \underbrace{\lambda}_{l_2 \text{ regularization parameter}} \|\theta\|_2^2$$

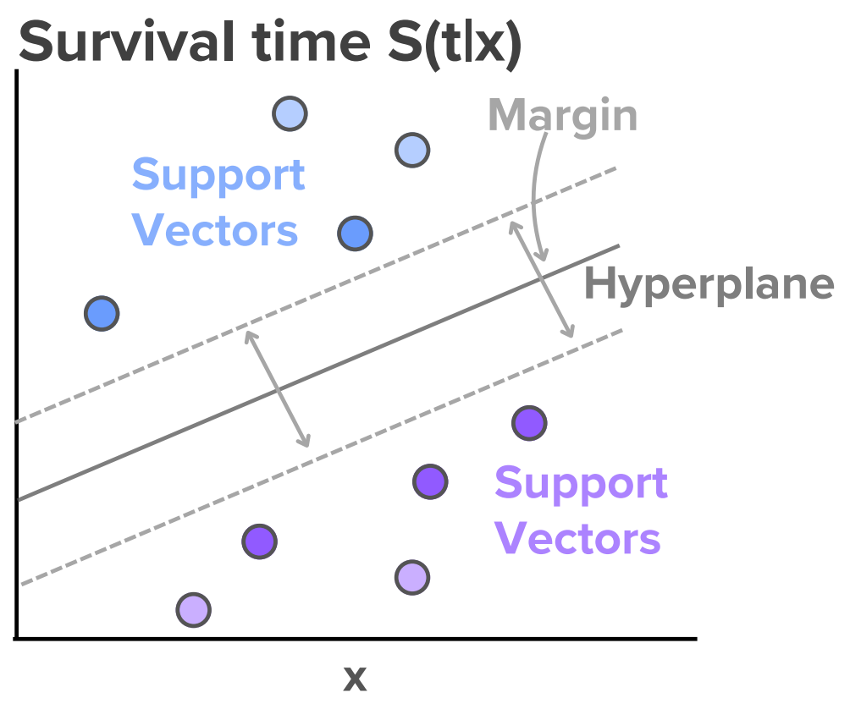


2. Survival Support Vector Regression (Survival SVR)

- Objective function[6]:

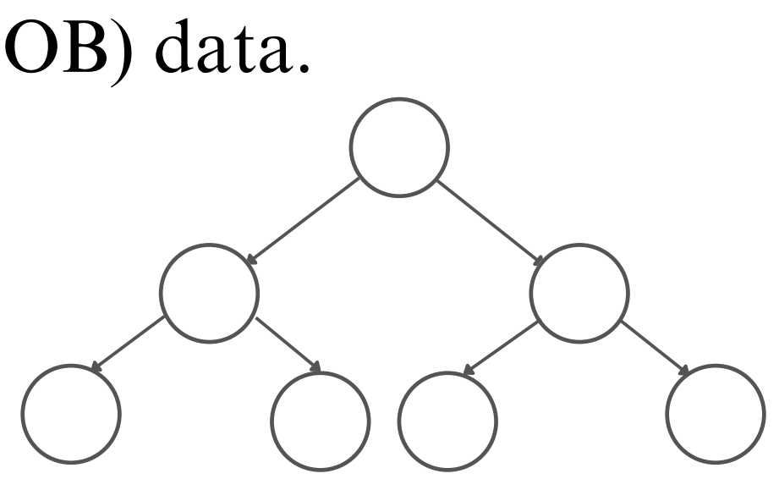
$$f(w,b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{\overset{\text{regulation parameter}}{\gamma}}{2}\sum_{i=0}^{n}(\zeta_{\mathbf{w},b}(y_i, x_i, \delta_i))^2$$

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



3. Random Survival Forest (RSF) [4]

- **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 d_0 unique deaths
- **Predict** terminal node CHF values.
- **Average** cumulative hazard function (CHF) from all trees.
- **Evaluate** prediction error using OOB data.



Evaluation & Results

Train-Test Split (at 4 to 1 rate)

Hyperparameter Tuning (5 repeats of 5-fold cross-validation)

Repeated Evaluation (20 times)

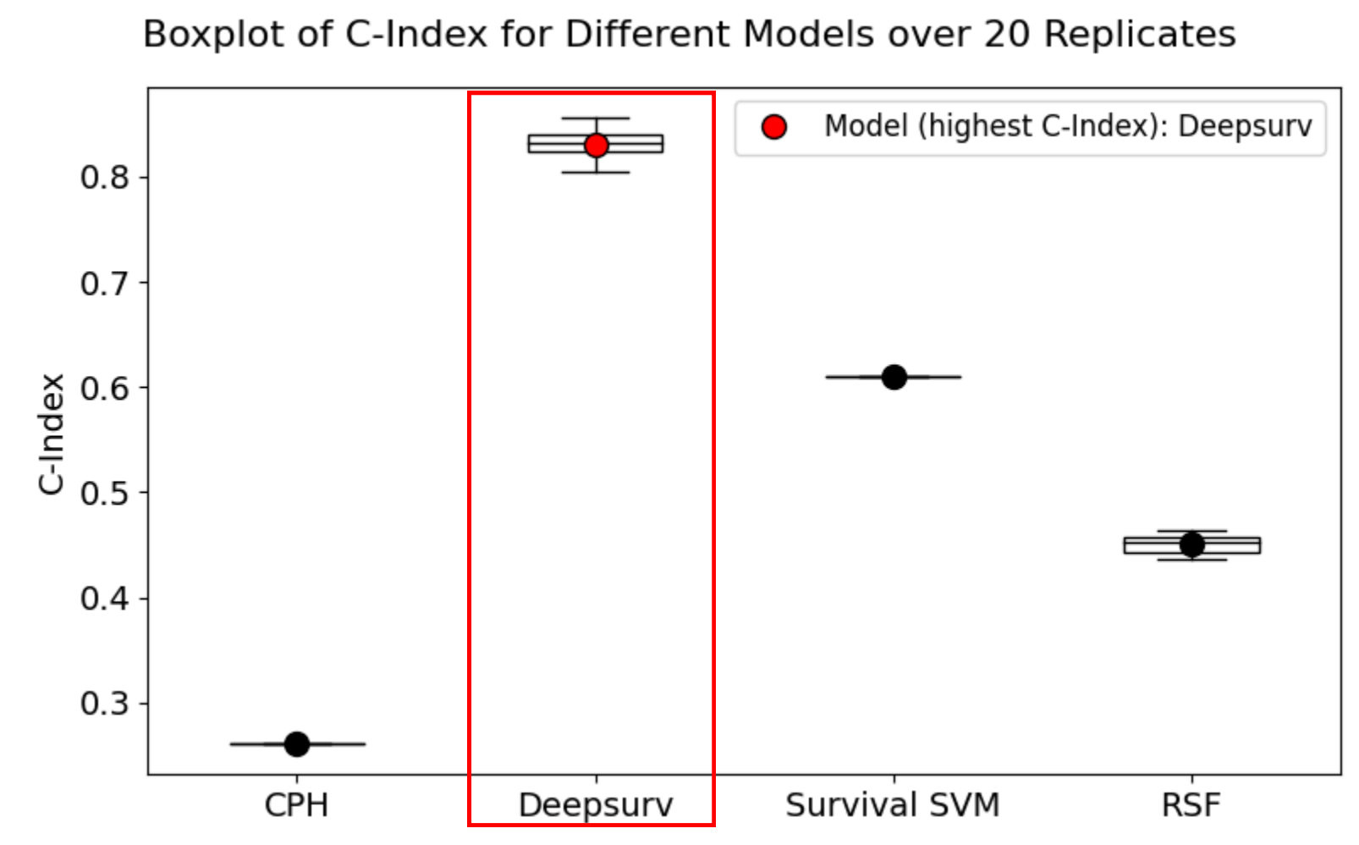
- **Evaluation Metrics:**
  - **Concordance index (C-index):**
    - Measures the ability to rank individuals by survival times correctly.
    - **Higher** values (closer to 1) indicate **better** predictive performance[7].

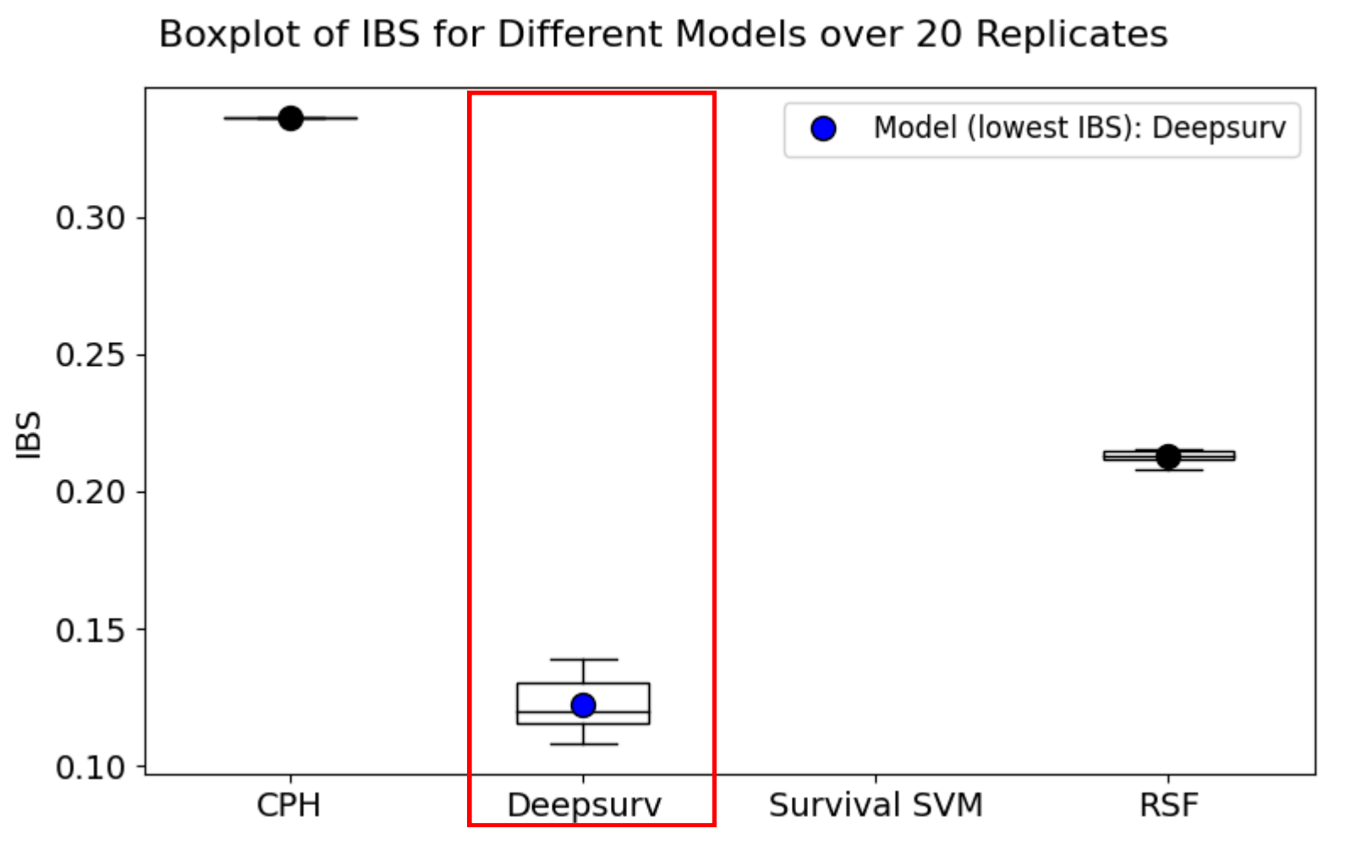
$$\text{C-index} = \frac{\sum_{i,j} I(T_j < T_i) \cdot I(r_j > r_i) \cdot \delta_j}{\sum_{i,j} I(T_j < T_i) \cdot \delta_j}$$

  - **Integrated Brier Score (IBS):**
    - Reflects overall model accuracy and calibration.
    - **Lower** values (closer to 0) signify **better** performance[7].

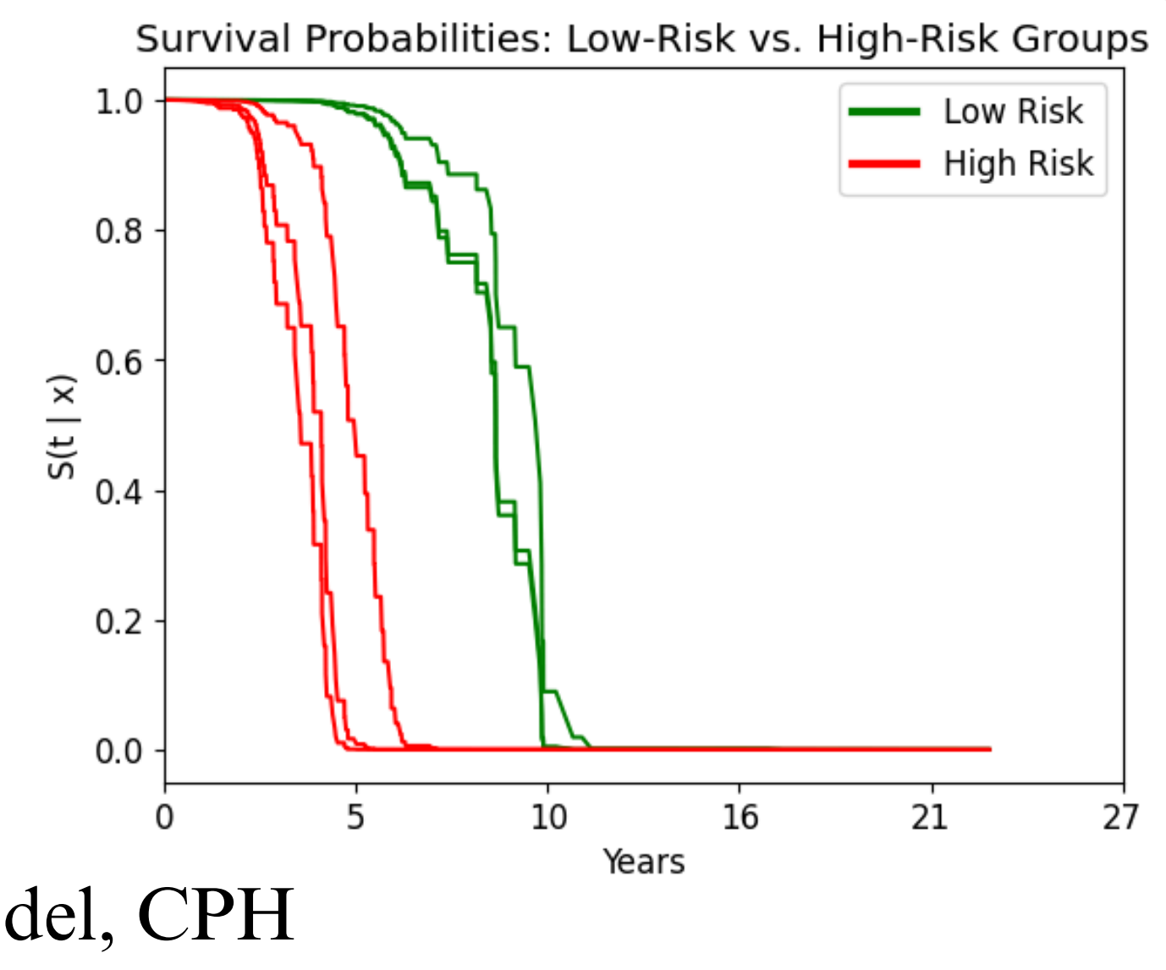
$$IBS(\tau) = \frac{1}{\tau}\int_0^\tau \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(0 - \hat{S}(t|x_i)\right)^2 \cdot I(Y_i \leq t, \delta_i = 1)}{\hat{G}(Y_i)} + \frac{\left(1 - \hat{S}(t|x_i)\right)^2 \cdot I(Y_i > t)}{\hat{G}(t)}\right)dt$$

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)


Boxplot of C-Index for Different Models over 20 Replicates


Boxplot of IBS for Different Models over 20 Replicates

Conclusion & Discussion


Survival Probabilities: Low-Risk vs. High-Risk Groups

- **Clinical Insights:**
  - **Low-risk Group:** employ watchful waiting or less aggressive treatments.
  - **High-risk Group:** optimize treatment plans and ensure timely interventions.
- **Overall Implications:**
  - All ML models **outperformed** the baseline model, CPH
  - **DeepSurv** demonstrated superior predictive performance
- **Future Directions:**
  - Alternative dimension reduction methods [1]
  - Unique challenges of subtypes like TNBC
  - Additional Multi-omics data for refined predictive models

References and 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. 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
3. 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
4. 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
5. 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
6. 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.
7. 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

I want to express my sincere gratitude to Dr. Y. Gu, my supervisor, for her invaluable guidance and unwavering support throughout this endeavor. I also appreciate 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.