# Uncertainty Estimation in Cancer Survival Prediction

Hrushikesh Loya[1], Pranav Poduval[1], Deepak Anand[1], Neeraj Kumar[2], Amit Sethi[1,3]

Survival models predict survival and used in oncology for treatment planning and personalized therapy

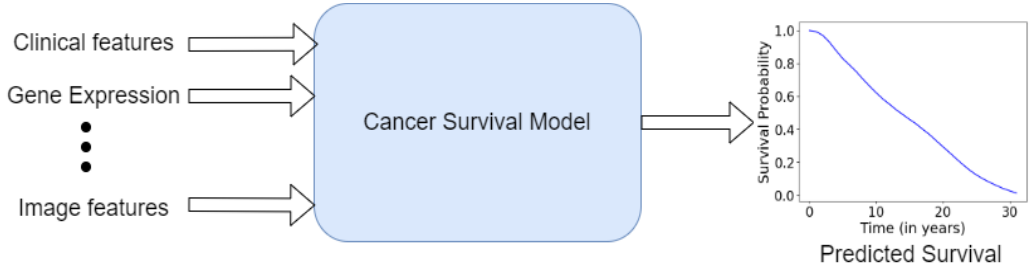


Fig 1: Cancer survival models predict survival probabilities

Existing models do not capture patient-specific uncertainty in prediction

- They only predict patient-specific survival
- They can estimate only overall model uncertainty
- These are barriers to interpretability and trust
- Model should recognize out-of-distribution samples as highly uncertain predictions

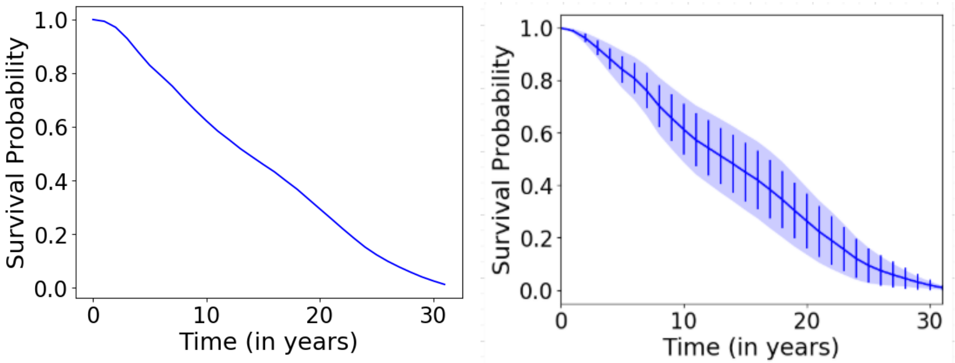


Fig 2: Survival probability without (L) and with (R) uncertainties

[1] Department of Electrical Engineering, Indian Institute of Technology Bombay
[2] Department of Biomedical Engineering, Case Western Reserve University
[3] Department of Pathology, University of Illinois at Chicago

Contact: hrushikesh.loya@iitb.ac.in

# Background – Setting

## Each Patient *i* has:

- Set of covariates $x_i$ (here, expression of PAM50[1] genes and clinical features)
- Time of adverse event since diagnosis (here, death) $T_i$
- Event indicator $E_i$ (0 means right censoring, i.e. loss to follow-up)

Table 1: Sample data input to the model

| Patient ID | Age | Subtype A | Gene B | $T_i$ (months) | $E_i$ |
|---|---|---|---|---|---|
| A-01 | 23 | 1 | 0 | 45 | 1 |
| A-02 | 52 | 0 | 1 | 23 | 0 |

Covariates

Time

Event

[1] Parker, Joel S., et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes." *Journal of clinical oncology* 27.8 (2009): 1160.

# Background – Multitask Logistic Regression (MTLR)[2]

- Divides the time in *m+1* bins & fits a logistic regression for survival probability in each bin

$$P_{\theta_i}(T \geq t_i \mid x) = (1 + exp(\vec{\theta}_i.\vec{x} + b_i))^{-1}; 0 \leq i \leq m$$

- The parameters $\vec{\theta_i}$ and $b_i$ depend on the time interval $i$

- MTLR encoded survival time of patient as binary sequence $y = (y_1, y_2, ..., y_m)$, where $y_i$ is survival status at $t_i$
- The **joint** likelihood of observing a sequence is then given by:

$$P_\theta(Y = (y_1, y_2, y_3 ... y_m) \mid \vec{x}) = \exp\left[\sum_{k=j}^{m} y_i(\vec{\theta}_i.\vec{x} + b_i)\right] / \left[\sum_{k=0}^{m} \exp(f_\theta(\vec{x}, k))\right]; \quad f_\theta(\vec{x}, k) = \sum_{j=k+1}^{m} (\vec{\theta}_i.\vec{x} + b_i)$$

[2] Yu, Chun-Nam, et al. "Learning patient-specific cancer survival distributions as a sequence of dependent regressors." *Advances in Neural Information Processing Systems*. 2011.

# Methods – Variational Inference

- Assumes an approximate posterior $q_\psi(\theta)$, and fits it to be close to (in KL divergence) the actual posterior $p(z|x)$
- Equivalent to minimizing the variational free energy[3,4], given by:

$$\hat{\mathcal{L}}(\psi) \quad = -\mathbb{E}_{q_\psi(\theta)}[\log p(D^i|\theta^i)] + KL(q_\psi(\theta)||P(\theta))$$

  - $-\mathbb{E}_{q_\psi(\theta)}[\log p(D^i|\theta^i)]$ is expectation of negative log-likelihood, which is approximated using unbiased MC samples
  - $KL(q_\psi(\theta)||P(\theta))$ is the KL divergence between assumed posterior and prior

[3] Karl Friston, J ́er ́emie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. Variationalfree energy and the laplace approximation.Neuroimage, 34(1):220–234, 2007.

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty inneural networks.arXiv preprint arXiv:1505.05424, 2015.

# Methods – Variational Inference

- **Data uncertainty:** We used standard trick of predicting not only mean but also the variance[5]

$y_{out} = \hat{y} + \hat{\sigma}.\epsilon; \epsilon \sim N(0, 1)$ where, $\hat{y}$ and $\hat{\sigma}$ are approximated using MC samples

- **Model uncertainty:** We compute the variance in survival probability curves for multiple forward passes (after sampling from the prior) through the network

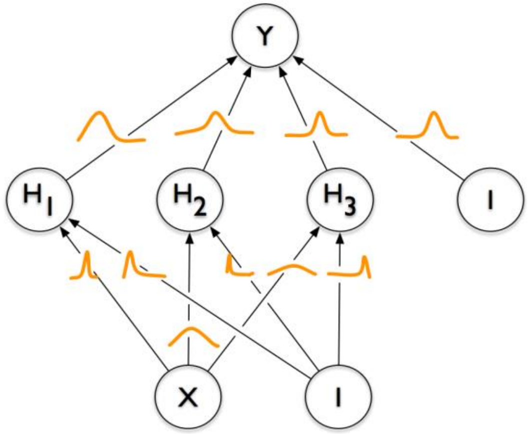


Fig 3: "Weight uncertainty in neural networks" Blundell et.al.

[5] Kendall, Alex, and Yarin Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?." *Advances in neural information processing systems*. 2017.

# Methods – Spike and slab prior and posterior

- Spike and slab prior → sparse solutions
- Closed form solutions for ELBO[6]

$$p(\theta) = \prod_{i=1}^{N} (\alpha \mathcal{N}(\theta_i; 0, 1) + (1 - \alpha)\delta(\theta_i))$$

$$q_\psi(\theta) = \prod_{i=1}^{N} (\gamma_i \mathcal{N}(\theta_i; \mu_i, \sigma_i^2) + (1 - \gamma_i)\delta(\theta_i))$$

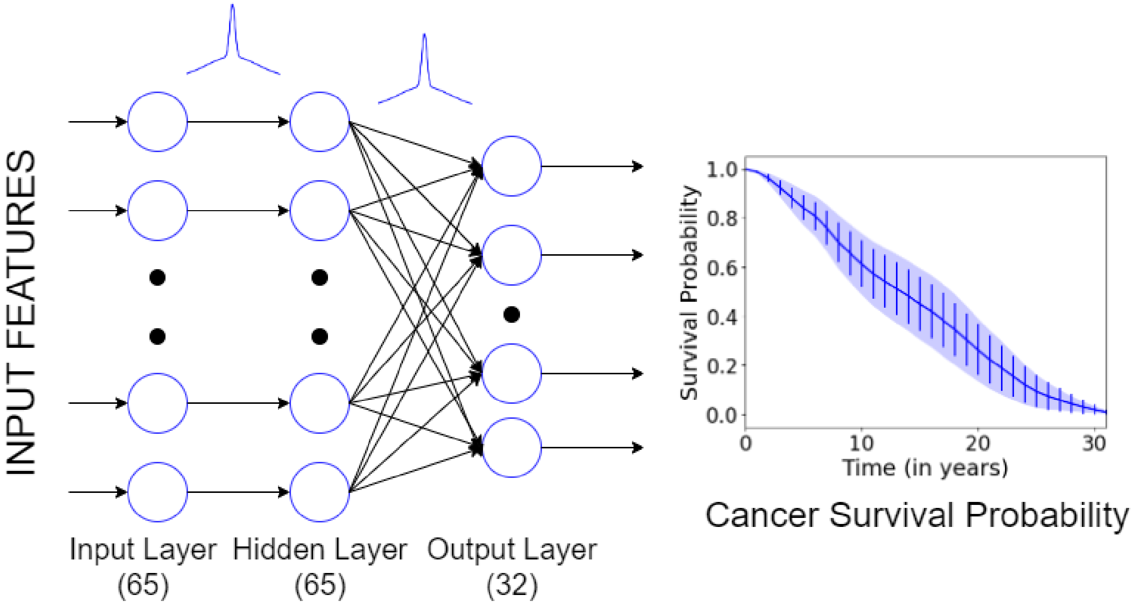


Fig 4: Proposed neural network architecture gives survival probability (solid curve), along with data uncertainty (vertical bars), and model uncertainty (shaded region).

$$\hat{\mathcal{L}}(\psi) = -\log p(D^i|\theta^i) + \frac{1}{M}\sum_{i=1}^{N}\left(\frac{\gamma_i}{2}(\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2)) + (1 - \gamma_i)\log(\frac{1-\alpha}{1-\gamma_i}) + \gamma_i \log(\frac{\alpha}{\gamma_i})\right)$$

[6] Tonolini, Francesco, Bjorn Sand Jensen, and Roderick Murray-Smith. "Variational Sparse Coding." (2019).

# Results – Survival Predictions

- **C-index**: Generalization of the area under the ROC curve (AUC)

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{1}_{T_j<T_i} \mathbb{1}_{\eta_j>\eta_i} \delta_j}{\sum_{i,j} \mathbb{1}_{T_j<T_i} \delta_j}$$

where, $\eta_i$ is risk score of patient i

- **IBS**: Integration of Brier score (weighted MSE) for censored variables defined below

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{(0-\hat{S}(t,\vec{x}_i))^2 \mathbb{1}_{T_i\leq t,\delta_i=1}}{\hat{G}(T_i^-)} - \frac{(1-\hat{S}(t,\vec{x}_i))^2 \mathbb{1}_{T_i>t}}{\hat{G}(t)} \right)$$

(where, $\hat{G}(T) = P(C \geq t)$ is prob. of censoring, calculated using KM curve)

Table 2: Comparison of C-index and IBS across survival models using TCGA-BRCA for training and METABRIC for testing

| Method | C-index | IBS |
|---|---|---|
| CoxPH[7] | 0.65 ± 0.1 | 0.2 ± 0.07 |
| MTLR | 0.68 ± 0.06 | 0.21 ± 0.06 |
| Neural MTLR[8] | 0.68 ± 0.02 | 0.16 ± 0.04 |
| Our method | **0.71 ± 0.05** | **0.12 ± 0.02** |

[7] D. R. Cox. Regression models and life-tables.Journal of the Royal Statistical Society. Series B(Methodological), 34(2):187–220, 1972. ISSN 00359246.

[8] Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework, 2018.

# Results – Feature Ranking

- Identifying key features from high-dimensional gene expression data
- Most important genetic features:
  - **BCL2 -** Antiapoptotic protein, good prognostic marker for Luminal A breast cancers
  - **CDC20 -** Oncoprotein that promotes the development and progression of breast cancer
  - **RASGRF1 -** Role in Tumor cell proliferation and and inflammation

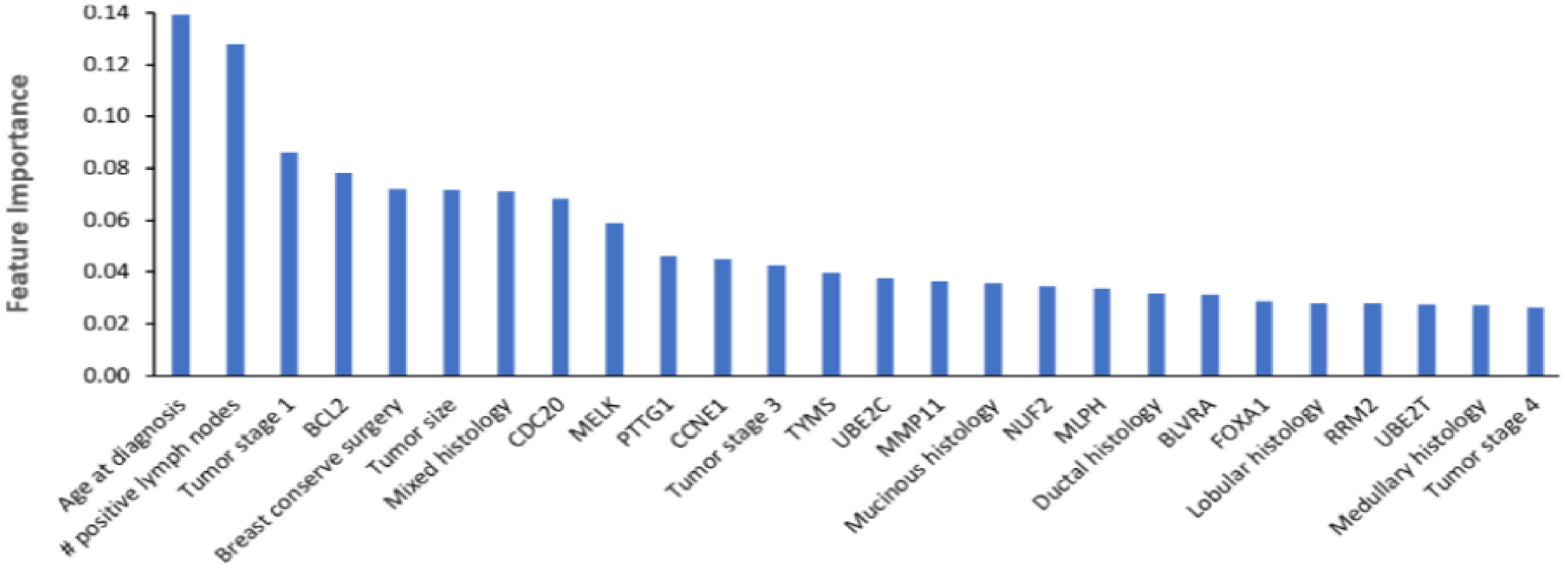
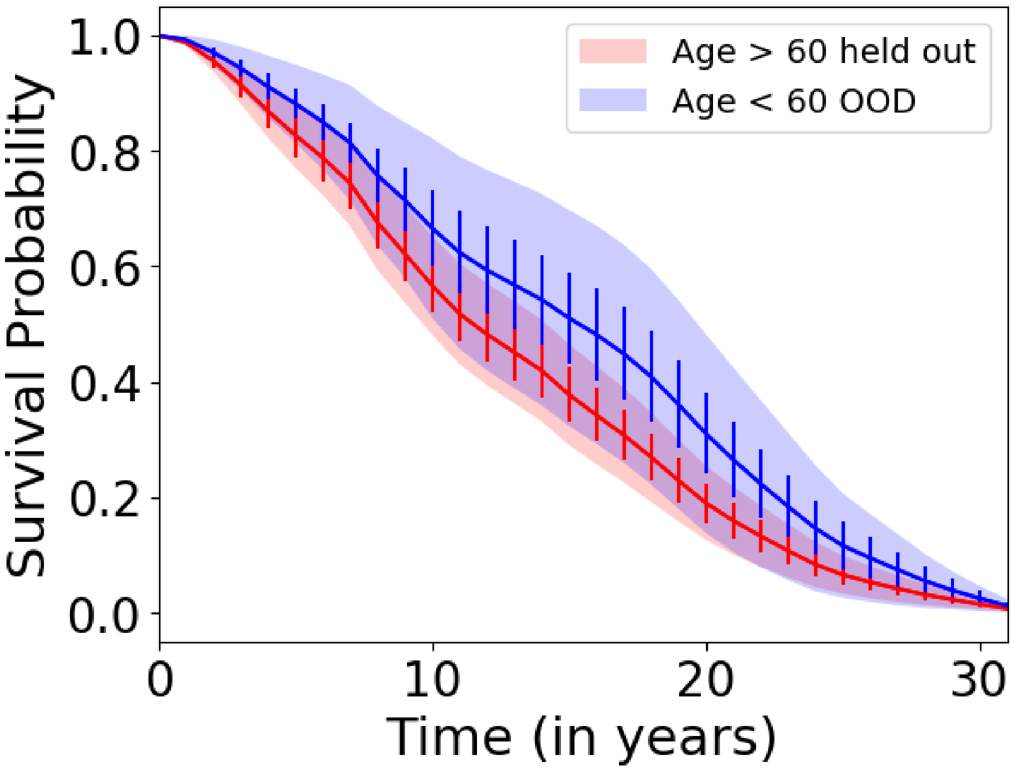


Fig2: Importance scores for a truncated list of the features

# Results – Out of Distribution Scenario 1 : Age difference

- Trained on older and tested on younger patients
  - Old: Age > 60
  - Young: Age < 60

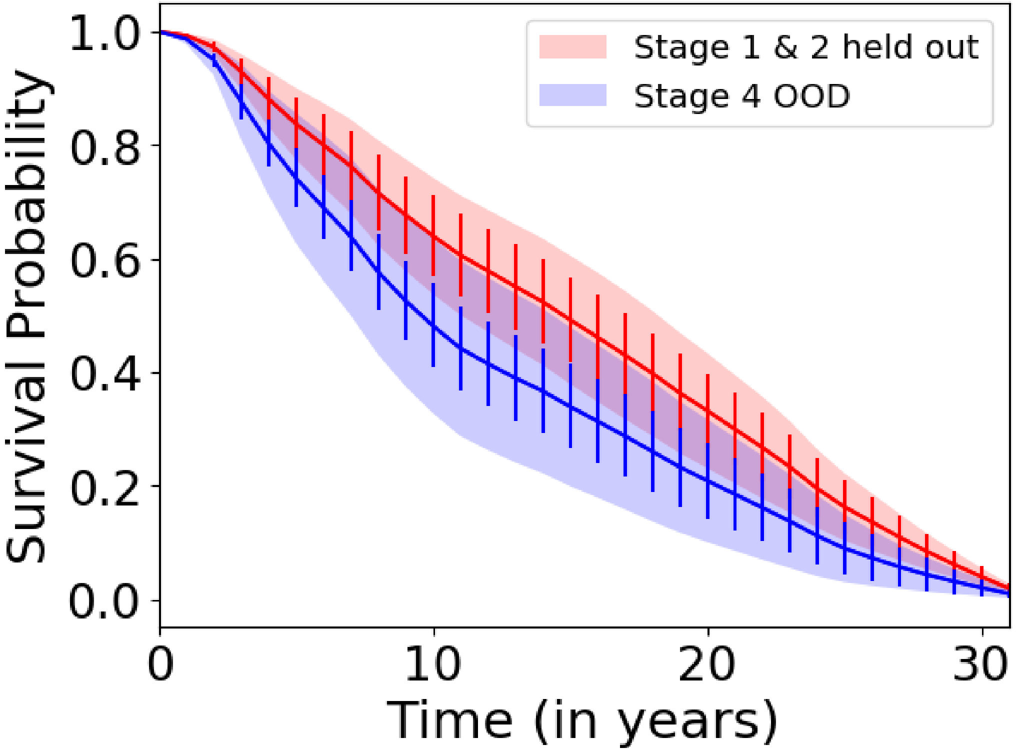
- Model more uncertain on young patients
  - **110%** higher mean uncertainty on young patients (OOD) compared to held-out old patients



# Results – Out of Distribution Scenario 2 : Stage difference

- Trained on lower stage and tested on higher stage patients
  - Lower: Stage 1 or 2
  - Higher: Stage 4

- Model more uncertain on higher stage patients
  - **43%** higher mean uncertainty on higher stage patients (OOD) compared to held-out lower age patients


Stage 1 & 2 held out
Stage 4 OOD
Survival Probability
Time (in years)