Dual Mixture-of-Experts Framework for Discrete-Time Survival Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Survival analysis is a task to model the time until an event of interest occurs, widely used in clinical and biomedical research. A key challenge is to model patient heterogeneity while also adapting risk predictions to both individual characteristics and temporal dynamics. We propose a dual mixture-of-experts (MoE) framework for discrete-time survival analysis. Our approach combines a feature-encoder MoE for subgroup-aware representation learning with a hazard MoE that leverages patient features and time embeddings to capture temporal dynamics. This dual-MoE design flexibly integrates with existing deep learning–based survival pipelines. On METABRIC and GBSG breast cancer datasets, our method consistently improves performance, boosting the time-dependent C-index up to 0.04 on the test sets, and yields further gains when incorporated into the Consurv framework.

1 Introduction

Survival analysis aims to predict the time until an event while properly accounting for censoring. A long-standing approach is the Cox Proportional Hazards (CPH) model [Cox, 1972], which assumes that hazard ratios between patients remain proportional over time. While effective in many settings, this assumption often fails in real-world clinical data, where risk dynamics are non-proportional over time. To address this limitation, recent deep learning models (e.g., DeepHit [Lee et al., 2018], ConSurv [Lee et al., 2024]) replace the CPH constraint with flexible neural architectures and are trained with negative log-likelihood objectives [Gensheimer and Narasimhan, 2019, Ren et al., 2019, Curth et al., 2021, Lee et al., 2024], enabling the modeling of non-proportional hazards.

Despite this progress, most deep survival models still rely on a single shared feature encoder. In practice, patients form heterogeneous subgroups with distinct risk profiles, and a single encoder tends to favor dominant patterns while underrespenting minority subgroups [Zhou et al., 2021, Guo et al., 2018, Jin et al., 2023]. Hazard estimation is likewise commonly implemented with a single network, yet survival risk is both time-varying and patient-specific: two patients at the same time point may exhibit markedly different risk trajectories depending on their clinical characteristics. A single network implicitly ties all patients and all time bins to one shared functional form, leaving further room for improvement in capturing patient heterogeneity and temporal dynamics.

In this paper, we propose a dual mixture-of-experts (MoE) [Shazeer et al., 2017] framework that integrates a mixture of feature encoders and a mixture of hazard networks to address these limitations (Fig. 1). The feature-encoder MoE models patient heterogeneity through soft routing based on each patient's encoded features, while the hazard-network MoE outputs a full hazard vector over the prediction horizon with a soft router conditioned on both patient representation and time embedding. This joint design enables experts to specialize along temporal horizons while adapting to patient subgroups, resulting in finer-grained, context-aware hazard modeling. Experiments on the METABRIC and GBSG datasets show consistent improvements in both overall and time-dependent

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

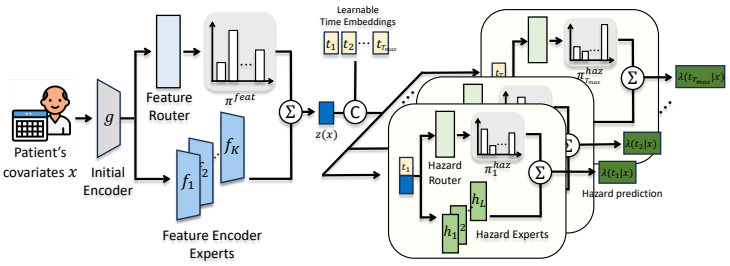

Figure 1. Overall architecture of the proposed framework. Unlike prior survival models that use a single encoder and a single hazard head, our framework employs dual mixtures of experts: one over feature encoders and another over hazard networks, where hazard experts are shared across time bins and dynamically routed by patient and time embeddings.

C-index over conventional single-network models, with further gains when incorporated into the ConSurv framework.

2 Method

2.1 Preliminary

We formulate survival prediction in discrete-time setting. Each patient $i \in \{1, ..., N\}$ is represented by covariates x_i, an observed time $\tau_i \in \{0, ..., T_{max}\}$, and an event indicater $\delta_i \in \{0, 1\}$, where $\delta_i = 1$ denotes an obersved event at τ_i and $\delta_i = 0$ indicates that the observation is right-censored at τ_i, i.e., the patient was event-free up to τ_i but their subsequent status is unobserved.

The conditional hazard function specifies the instantaneous event probability at time t:

$$\lambda(t \mid x) = \mathbb{P}(T = t \mid T \geq t, x).$$

It induces the survival function

$$S(t \mid x) = \prod_{t' \leq t} \big(1 - \lambda(t' \mid x)\big),$$

which captures the probability of remaining event-free after t. The probability mass for an event at time t is then

$$p(t \mid x) = \lambda(t \mid x)\, S(t - 1 \mid x).$$

Model estimation proceeds via maximum likelihood. The negative log-likelihood objective combines information from observed and censored cases:

$$\mathcal{L}_{NLL} = -\sum_{i=1}^{N} \Big[\delta_i \log \hat{p}(\tau_i \mid x_i) + (1 - \delta_i) \log \hat{S}(\tau_i \mid x_i)\Big].$$

2.2 Mixture of Feature Encoders

We first enhance the representation learning stage by introducing a mixture of feature encoder architecture. An initial encoder $g(\cdot)$ extracts patient-level representations, which are then routed into multiple expert encoders $\{f_k\}_{k=1}^{K}$. The router takes patient features as input and produces π_k^{feat} via softmax function, which represents the routing probability (or mixing weight) of expert k for patient x. The final encoded representation for patient x is computed as

$$z(x) = \sum_{k=1}^{K} \pi_k^{feat} \cdot f_k(g(x))$$

By conditioning routing decisions on patient features, this design encourages the encoder to discover hidden subgroups and produce subgroup-aware representations. To prevent collapse into a single expert, we incorporate a load balancing loss $\mathcal{L}_{LB}^{feat} = \alpha\Big(K \sum_k \bar{\pi}_k^{feat^2} - 1\Big)$, where $\bar{\pi}_k^{feat} = \frac{1}{B} \sum_{i=1}^{B} \pi_{k,i}^{feat}$ denotes the batch wise averaged assignment probability for routing logit for expert k. This regularizer promotes healthy utilization of all experts by penalizing excessive reliance on a subset of them.

2.3 Mixture of Hazard Networks

On top of the encoded patient representations, we further introduce a mixture of hazard network for hazard prediction. Unlike feature encoder MoE, the hazard MoE conditions routing on both patient features and temporal embeddings. Each hazard expert $\{h_l\}_{l=1}^{L}$ predicts hazards for all discrete time bins, while the router conditions by concatenation of patient features and time embeddings to produce routing probability $\pi_{t,l}^{haz}$. The final hazard prediction at t is expressed as

$$\lambda(t|x) = \sum_{l=1}^{L} \pi_{t,l}^{haz} \cdot h_l(z(x), e_t)$$

where e_t denotes the learnable time embedding for time-bin t. Joint conditioning on patient features and time enables experts to specialize across both patient heterogeneity and temporal dynamics. This enables experts to capture finer-grained survival patterns (e.g., subgroups that differ not only in patient profiles but also in how risks evolve over time). As with the feature encoder MoE, we apply a load balancing loss $\mathcal{L}_{LB}^{haz} = \beta\big(T_{max} \sum_t L \sum_l \bar{\pi}_{t,l}^{haz^2} - 1\big)$ to encourage balanced usage of hazard experts across all time-bin.

2.4 Overall Training Objective

The overall training objective combines the discrete-time negative log-likelihood (NLL) loss with the load balancing regularizers applied at both stages:

$$\mathcal{L} = \mathcal{L}_{NLL} + \mathcal{L}_{LB}^{feat} + \mathcal{L}_{LB}^{haz}.$$

This formulation ensures that the model not only fits observed survival outcomes but also maintains balanced expert utilization across both representation and prediction stages, leading to more robust subgroup- and time-aware survival modeling.

3 Experiment

Method	Dual MoE	C-index	Time-dependent C-index								
			10%	20%	30%	40%	50%	60%	70%	80%	90%
Metabric											
CoxPH Cox [1972]	-	0.663 ± 0.017	0.658 ± 0.032	0.667 ± 0.022	0.665 ± 0.018	0.665 ± 0.019	0.659 ± 0.013	0.651 ± 0.015	0.648 ± 0.015	0.639 ± 0.020	0.646 ± 0.026
Naïve impl.	✗	0.646 ± 0.021	0.670 ± 0.050	0.660 ± 0.032	0.644 ± 0.019	0.644 ± 0.021	0.638 ± 0.022	0.629 ± 0.015	0.621 ± 0.017	0.611 ± 0.024	0.606 ± 0.022
Naïve impl.	✓	0.654 ± 0.015	0.669 ± 0.032	0.667 ± 0.022	0.657 ± 0.013	0.653 ± 0.017	0.646 ± 0.015	0.638 ± 0.016	0.628 ± 0.018	0.621 ± 0.027	0.623 ± 0.022
ConSurv Lee et al. [2024]	✗	0.657 ± 0.020	0.656 ± 0.044	0.668 ± 0.030	0.658 ± 0.018	0.657 ± 0.021	0.649 ± 0.018	0.639 ± 0.011	0.629 ± 0.010	0.616 ± 0.024	0.617 ± 0.026
ConSurv Lee et al. [2024]	✓	0.668 ± 0.018	0.696 ± 0.034	0.689 ± 0.024	0.676 ± 0.021	0.669 ± 0.022	0.657 ± 0.017	0.647 ± 0.015	0.642 ± 0.016	0.632 ± 0.021	0.634 ± 0.019
GBSG											
CoxPH Cox [1972]	-	0.659 ± 0.012	0.739 ± 0.046	0.709 ± 0.018	0.681 ± 0.017	0.676 ± 0.014	0.670 ± 0.013	0.662 ± 0.012	0.658 ± 0.011	0.655 ± 0.011	0.652 ± 0.011
Naïve impl.	✗	0.662 ± 0.012	0.744 ± 0.039	0.706 ± 0.017	0.678 ± 0.018	0.674 ± 0.015	0.669 ± 0.014	0.662 ± 0.012	0.657 ± 0.011	0.655 ± 0.012	0.652 ± 0.011
Naïve impl.	✓	0.667 ± 0.010	0.751 ± 0.033	0.717 ± 0.018	0.689 ± 0.016	0.684 ± 0.016	0.677 ± 0.014	0.670 ± 0.011	0.666 ± 0.011	0.663 ± 0.011	0.659 ± 0.010
ConSurv Lee et al. [2024]	✗	0.665 ± 0.011	0.742 ± 0.039	0.709 ± 0.017	0.682 ± 0.016	0.679 ± 0.014	0.674 ± 0.013	0.667 ± 0.011	0.663 ± 0.011	0.661 ± 0.011	0.658 ± 0.010
ConSurv Lee et al. [2024]	✓	0.668 ± 0.011	0.752 ± 0.036	0.715 ± 0.018	0.689 ± 0.017	0.684 ± 0.016	0.677 ± 0.014	0.670 ± 0.012	0.666 ± 0.011	0.663 ± 0.011	0.659 ± 0.010

Table 1. Comparison of both overall and time-dependent C-index on the Metabric and GBSG datasets. We report the average performance over 10 random seeds.[1]

3.1 Experimental Settings

Dataset. We evaluate our method on two widely used breast cancer survival datasets, which are Metabric [Curtis et al., 2012] and GBSG [Schumacher et al., 1994] . Metabric contains clinical and gene expression information from 1,981 patients, with 21 variables in total; 55.2% of the cases are censored and 44.8% uncensored. GBSG includes 2,232 patients with 21 clinical and tumor-related variables, originally collected to study the impact of hormone therapy on recurrence-free survival. In this dataset, 43.2% of the cases are censored and 56.8% uncensored.

Comparison Methods. We evaluated the effectiveness of the proposed dual mixture-of-experts framework under two experimental settings. First, we considered a naïve implementation trained solely with $\mathcal{L}_{NLL}$, consisting of a single feature encoder and a single hazard network, and compared it against our proposed dual MoE framework. Second, we applied our approach on top of the ConSurv framework [Lee et al., 2024], replacing its feature encoder and hazard network with a mixture of

[1]Results may differ from [Lee et al., 2024] since dataset splits vary with different random seeds.

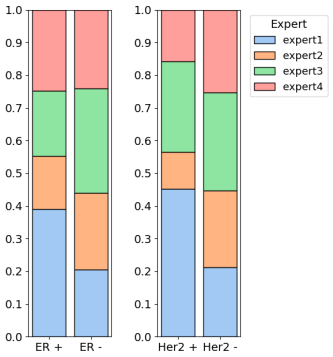

Figure 2. Average routing probabilities of *feature-encoder experts* across ER and HER2 subgroups.

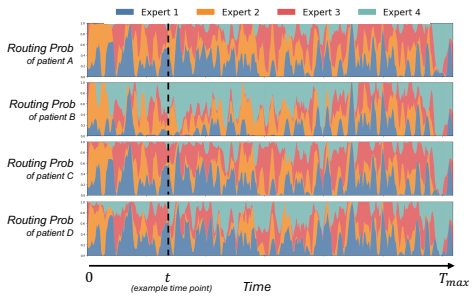

Figure 3. Patient-level routing probabilities of *hazard experts* over time. At any discrete time bin t (a vertical slice), the proportion of each color equals proportion of routing probability to each hazard experts ($\pi_{t,l}^{haz}, l \in \{1, 2, 3, 4\}$), which sum to 1.

feature encoders and a mixture of hazard networks, respectively. In both cases, we set the number of feature-encoder experts (K) and hazard experts (L) to (4, 4) for Metabric and (6, 3) for GBSG. More details can be found in the Appendix.

Evaluation Metrics. We evaluate performance using the concordance index (C-index) [Harrell et al., 1984], which measures concordance between predicted hazards and observed event times across the entire period. However, this global measure may overlook how performance varies across different time intervals. To assess temporal variations, we also report time-dependent C-index, computed at multiple time horizons defined by the 10%–90% percentiles of observed event times [Gerds et al., 2013].

3.2 Results

Table 1 summarizes the main performance comparison of the proposed dual MoE framework. By replacing the single feature encoder and hazard network with dual mixtures, we observed consistent improvements in both overall and time-dependent C-index on the METABRIC and GBSG datasets. Furthermore, integrating our framework with ConSurv leads to additional gains, indicating that the proposed method is easily applicable to other deep learning based discrete-time survival models.

Visualization of Feature Routing Probability. We examined the routing behavior of the feature-encoder MoE across estrogen receptor (ER) and HER2 subgroups using Metabric dataset (Fig. 2). Specifically, for each subgroup, we averaged the feature-encoder routing probabilities across all patients belonging to that subgroup to obtain representative expert assignment distributions. The router exhibited distinct expert preferences between subgroups, indicating that it adapts to patient heterogeneity rather than assigning weights uniformly.

Visualization of Trajectory of Hazard Routing Probability through Time. We visualized patient-level routing probabilities of the hazard router (Fig. 3). Each panel shows the soft assignment probabilities of four hazard experts over time for each of 4 patients (A-D). Routing patterns vary across patients but consistently show shifts in expert dominance between early and late time horizons, indicating adaptation to both individual heterogeneity and temporal structure.

More ablation studies, including effect of each MoE architecture and input of hazard router, are provided in the Appendix.

4 Conclusion & Future Work

In this work, we proposed a dual mixture-of-experts framework for discrete-time survival analysis that integrates mixtures of feature encoders and hazard networks to model patient heterogeneity and temporal risk variation. On the Metabric and GBSG datasets, our method consistently outperformed conventional single-network models and yielded further gains with ConSurv. As future work, we plan to further analyze the role of each expert and extend our framework to multimodal settings, such as mammography-based risk prediction.

References

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x. URL https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. SurvITE: Learning heterogeneous treatment effects from time-to-event data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=f0_tkoEJV88.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. doi: 10.1038/nature10983. URL https://doi.org/10.1038/nature10983. PMCID: PMC3440846.

Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. doi: 10.7717/peerj.6257. URL https://doi.org/10.7717/peerj.6257.

Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Chang Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, June 2013. doi: 10.1002/sim.5681. Epub 2012 Nov 22.

Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1498. URL https://aclanthology.org/D18-1498/.

Frank E. Jr Harrell, Kerry L. Lee, Robert M. Califf, David B. Pryor, and Robert A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152, April–June 1984. doi: 10.1002/sim.4780030207.

Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23695–23704, June 2023.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11842. URL https://doi.org/10.1609/aaai.v32i1.11842.

Dongjoon Lee, Hyeryn Park, and Changhee Lee. Toward a well-calibrated discrimination via survival outcome-aware contrastive learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=UVjuYBSbCN.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep Recurrent Survival Analysis. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 4798–4805. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33014798. URL https://doi.org/10.1609/aaai.v33i01.33014798.

M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. Neumann, and H. F. Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12(10):2086–2093, October 1994. doi: 10.1200/JCO.1994.12.10.2086.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. ISSN 1941-0042. doi: 10.1109/tip.2021.3112012. URL http://dx.doi.org/10.1109/TIP.2021.3112012.

A Training Details

Table 2 summarizes the key hyperparameters used in our experiments.

Item	METABRIC	GBSG
Initial encoder	MLP (depth=4)	MLP (depth=3)
Feature-encoder router	MLP (depth=1)	MLP (depth=1)
Feature-encoder expert	MLP (depth=1)	MLP (depth=2)
Number of feature-encoder experts (K)	4	6
Hazard router	MLP (depth=1)	MLP (depth=1)
Hazard expert	MLP (depth=1)	MLP (depth=1)
Number of hazard experts (L)	4	3
Time embedding dim (d_{time})	8	8
Load-balancing coef. (feature) α	0.3	0.3
Load-balancing coef. (hazard) β	0.5	0.5

Table 2. Hyperparameter details. Here, depth refers to the number of hidden layers in each MLP.

B More Ablation Studies

Feat. MoE	Haz. MoE	C-index
✗	✗	0.646 $\pm$ 0.021
✓	✗	0.649 $\pm$ 0.023
✗	✓	0.650 $\pm$ 0.025
✓	✓	0.654 $\pm$ 0.015

Table 3. Ablation on each MoE architecture.

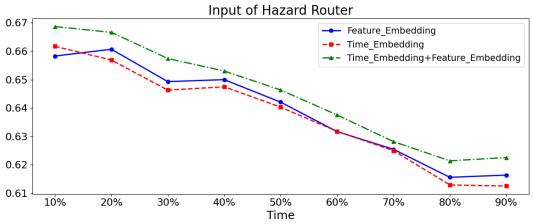

Figure 4. Ablation on input of hazard router.

For simplicity, we restrict our ablation experiments using Naïve implementation using the Metabric dataset.

Effect of each MoE Architecture. We conducted ablation studies to evaluate the contribution of each MoE component, the mixture of feature encoders and the mixture of hazard networks. As shown in Table 3, introducing the mixture of feature encoders or the mixture of hazard networks individually improves performance over the conventional single-network models When combined, the full dual mixture achieves the best performance, demonstrating that the two components are complementary.

Input of Hazard Router. We evaluate how the choice of router inputs affects performance. Specifically, we measured time-dependen C-index for three variants: patient features only, time embeddings only, and both. Note that in all variants the hazard experts themselves still take both patient features and time embeddings as input; only the inputs to the router are modified. As shown in Figure 4, using either patient features or time embeddings alone is suboptimal; their combination consistently achieves the best performance, highlighting the importance of jointly conditioning the router on both patient heterogeneity and temporal variation.