Beyond the Batch: Momentum-Updated Virtual Cohorts for 3D PET-CT Prognosis

Xinglong Liang[1,2], Tianyu Zhang[1,2], Lishan Cai[1,4], Jiaju Huang[3], Chunyao Lu[1,2], Luyi Han[1,2], Yaofei Duan[3], Xin Wang[1,4], Yuan Gao[1,4], Muzhen He[1], Jun Xu[5], Yue Sun[3], Tao Tan[3⋆], and Ritse Mann[1,2]

[1] Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[2] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands
[3] Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao Special Administrative Region of China
[4] GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, P. Debyelaan 25, 6202 AZ, Maastricht, The Netherlands
[5] School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen, 518055, China taotan@mpu.edu.mo

Abstract. Deep learning-based survival analysis with the Cox proportional hazards loss is challenged in 3D medical imaging due to GPU memory constraints, which enforce small mini-batches and thus yield an insufficient risk set and noisy optimization. To address this issue, we propose a Momentum-Contrastive Survival Framework (MCSF) that maintains a momentum-updated virtual cohort bank, decoupling the effective risk set size from the mini-batch size and enabling a larger global risk pool across iterations. Beyond enlarging the risk pool, we introduce a time-aware ranking objective that explicitly leverages event-time information: each event sample serves as an anchor and is compared against other event samples with temporal proximity-based weighting, providing a more stable and informative ranking signal than batch-limited Cox optimization. MCSF can be seamlessly integrated with various 3D backbones for end-to-end learning. Experiments on Breast and Head & Neck cancer cohorts demonstrate consistent improvements, achieving C-indices of 0.716 and 0.654, respectively, outperforming strong baseline backbones and highlighting the robustness of MCSF for 3D PET-CT survival prediction. Code is available at https://github.com/XinglongLiang08/MCSF.

Keywords: Survival Analysis · Momentum Contrast · PET-CT.

1 Introduction

Positron emission tomography/computed tomography (PET-CT) plays a central role in modern oncology by coupling anatomical context with tumor metabolism,

⋆ Corresponding author

most commonly through FDG uptake patterns [6, 5]. Beyond diagnosis and staging, FDG PET-CT has demonstrated substantial prognostic value across multiple cancers, including breast cancer [20, 15], Head & Neck cancer [1], lung cancer [10] and lymphoma [18]. Most of these approaches rely on predefined parameters, particularly metabolic metrics such as MTV, TLG, and SUVmax; however, such predefined parameters may lead to information loss, parameter selection bias, and an inability to capture complex spatial relationships and high-level semantic features [17, 15]. These limitations highlight the need to develop more robust imaging-driven risk representations that can directly leverage large-field PET-CT volumes to support individualized risk stratification.

Among deep survival approaches, CoxPH-based (Cox proportional hazards) objectives remain a popular choice because they naturally handle right-censoring and focus on risk ranking rather than explicit time discretization [4, 12]. The Cox partial likelihood for an observed event at time t_i compares the patient's predicted risk against a risk set R_i containing all individuals still at risk at t_i. This "global" comparison is exactly what makes CoxPH effective—but it also creates a practical bottleneck in 3D medical imaging. Training 3D backbones on volumetric data is GPU-memory intensive. In such small-batch regimes, the effective risk set inside each iteration becomes severely under-sampled, weakening the ranking signal and amplifying gradient noise. This mismatch between the global nature of Cox partial likelihood and the local nature of small mini-batches affects consistency and asymptotic variance [13, 23].

A straightforward response is to increase batch size, yet this is typically infeasible for whole-body 3D PET-CT because scans cover a large anatomical field-of-view with high-resolution volumetric data from two co-registered modalities, resulting in massive voxel counts and GPU-memory footprints that constrain training to very small mini-batches; gradient accumulation is often ineffective because it does not increase the risk-set size. Moreover, lesions are often spatially dispersed, so aggressive downsampling, cropping, or patching can miss relevant disease burden and patient-level context, undermining the capture of global risk signals. Alternative strategies include time-discretized models [22] or approximations to the Cox loss [14, 13]; however, discretization may introduce binning bias and additional design choices, while approximations can still suffer when batch diversity is low. Therefore, an effective solution should expand the risk pool seen by each event while maintaining the memory footprint compatible with 3D training.

In this work, we propose the Momentum-Contrastive Survival Framework (MCSF), a mechanism designed to decouple survival objective optimization from mini-batch size constraints. Inspired by momentum-contrastive learning [7, 3], MCSF maintains a Virtual Cohort Bank (VCB) across iterations to construct a large global risk pool. Instead of restricting each event's comparison to the current mini-batch, we compute the Cox ranking term against a substantially larger set formed by this bank. To address the critical issue of "feature drift" when reusing historical data, we employ an Evolving Reference Encoder that produces slowly evolving representations, stabilizing the relative risk ordering

required for survival analysis. Our framework is model-agnostic and can be integrated with common 3D backbones for end-to-end PET-CT survival prediction. Furthermore, to reduce noisy comparisons introduced by the enlarged VCB, we incorporate a time-aware weighting mechanism that assigns higher weights to pairs with larger survival time differences, thereby mitigating the noise from pairs with indeterminate or proximal outcomes. This allows the model to learn robust global risk stratifications even when training with extremely limited batch sizes. The workflow is shown in Fig. 1.

The primary contributions of this work are as follows: (1) We identify the small-batch bottleneck in 3D PET-CT survival learning and motivate a VCB to approximate the global Cox risk set, alleviating under-sampled comparisons and noisy gradients. (2) We introduce a time-aware weighted ranking objective to downweight ambiguous pairs, enabling reliable global risk ordering. (3) We demonstrate robust gains across clinically relevant breast cancer and Head & Neck survival prediction tasks, highlighting a practical path toward stable and scalable 3D imaging-based risk modeling.

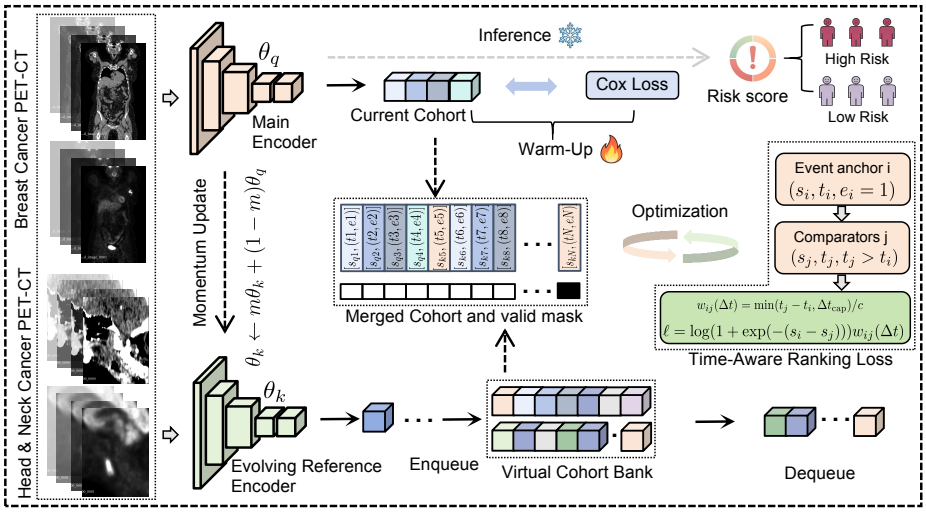

Fig. 1: Overview of the proposed Momentum-Contrastive Survival Framework for 3D PET-CT survival prediction. A main encoder extracts current-cohort features, while an evolving reference encoder is updated by momentum to generate stable representations stored in a virtual cohort bank. The merged current cohort and virtual cohort bank form an enlarged risk pool for survival ranking, where a time-aware weighting mechanism downweights ambiguous pairs and improves robust global risk stratification under small mini-batches.

2 Methodology

2.1 Problem Definition and the Small-Batch Challenge

Let $\mathcal{D} = \{(X_i, t_i, e_i)\}_{i=1}^{N}$ denote a dataset of N patients, where $X_i \in \mathbb{R}^{C\times D\times H\times W}$ represents the 3D PET-CT volume, $t_i \in \mathbb{R}^+$ is the observed time, and $e_i \in \{0, 1\}$ is the event indicator ($e_i = 1$ denotes an event occurrence, e.g., death/recurrence; $e_i = 0$ denotes censoring). The goal of a deep survival model f_θ is to predict a risk score $s_i = f_\theta(X_i)$, such that patients with shorter survival times are assigned higher risk scores, i.e., if $t_i < t_j$ and $e_i = 1$, then $s_i > s_j$.

The standard objective for training such models is the CoxPH loss. For an anchor patient i who experiences an event at t_i, the loss maximizes the partial likelihood against a risk set $\mathcal{R}_i = \{j \mid t_j \geq t_i\}$:

$$\mathcal{L}_{\text{Cox}} = -\frac{1}{N_e} \sum_{i:e_i=1} \left(s_i - \log \sum_{j\in\mathcal{R}_i} \exp(s_j) \right), \tag{1}$$

where N_e is the number of events. In 3D medical imaging, computing the global $\mathcal{R}_i$ is computationally intractable. Instead, $\mathcal{R}_i$ is approximated by the current mini-batch $\mathcal{B}$. However, due to the massive memory footprint of dual-modality PET-CT volumes, the batch size is severely constrained. Under such constraints, the in-batch risk set is often sparse or empty, leading to high variance in gradient estimation and poor convergence. To address this, we propose the MCSF to decouple the risk set size from physical memory limitations.

2.2 Momentum-Contrastive Survival Framework

As illustrated in Fig. 1, the MCSF framework leverages a momentum-contrastive architecture to maintain a large, consistent representation of the patient population without re-computing historical features. The framework consists of two parallel streams: a Main Encoder and an Evolving Reference Encoder, along with a virtual cohort bank.

Backbone We adopt a 3D ResNet18 [8] backbone tailored for PET-CT representation learning. Each input volume $X_i \in \mathbb{R}^{3\times D\times H\times W}$ is composed of three channels: PET, CT, and a binary lesion segmentation mask M_i. To enhance the contribution of anatomically relevant regions, we apply a segmentation-guided multiplicative modulation before feeding the data into the backbone. Specifically, PET and CT are re-weighted voxel-wise by an affine transform of the mask:

$$\tilde{M}_i = 1 + b\, M_i, \tag{2}$$

$$\text{PET}'_i = \text{PET}_i \odot \tilde{M}_i, \quad \text{CT}'_i = \text{CT}_i \odot \tilde{M}_i, \tag{3}$$

where b is a learnable scaling parameter and $\odot$ denotes element-wise multiplication. The final network input is formed by concatenating the modulated PET and CT channels:

$$X'_i = \text{Concat}(\text{PET}'_i, \text{CT}'_i) \in \mathbb{R}^{2\times D\times H\times W}. \tag{4}$$

The backbone encoder then maps X'_i to the corresponding feature representation.

Momentum Update and Virtual Cohort Bank To ensure the stability of risk scores utilized from previous iterations, the parameters θ_k are updated via a momentum average of θ_q:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{5}$$

where $m \in [0, 1)$ is a momentum coefficient. We maintain a First-In-First-Out (FIFO) queue $\mathcal{Q}$ of size K ($K \gg |\mathcal{B}|$). At the end of each iteration, the tuples $\{(s_{k,j}, t_j, e_j)\}$ generated by the Reference Encoder are enqueued, while the oldest entries are removed. This queue serves as a virtual cohort, providing a dense and diverse risk pool for the current training batch.

2.3 Time-Aware Pairwise Ranking Loss

Standard Cox loss treats all valid pairs in the risk set equally. However, in a survival context, the confidence of the ranking relationship depends on the time separation between patients. A pair with a large survival difference (e.g., 5 years vs. 1 month) provides a stronger learning signal than a pair with proximal outcomes. To exploit the expanded Virtual Cohort, we introduce a Time-Aware Pairwise Ranking Loss.

Merged Cohort Construction For every event anchor i in the current batch $\mathcal{B}$ (where $e_i = 1$), we construct a comparison pool $\mathcal{P}$ by merging the current batch and the virtual cohort bank: $\mathcal{P} = \mathcal{B} \cup \mathcal{Q}$. A valid comparator $j \in \mathcal{P}$ is any subject who is still at risk at time t_i, i.e., $t_j \geq t_i$ (including censored subjects).

Time-Aware Weighting We introduce a weighting function $w_{ij}(\Delta t)$ to modulate the contribution of each pair based on their time difference $\Delta t = t_j - t_i$. Pairs with small Δt are down-weighted to mitigate noise arising from clinical variability or indeterminate progression:

$$w_{ij}(\Delta t) = \min(t_j - t_i, \Delta t_{\text{cap}})/c, \tag{6}$$

where Δt_{cap} is a predefined upper bound on Δt, and c is a temporal threshold. This linear ramp function ensures that the model focuses on learning clear, distinguishable survival patterns.

Table 1: Comparison of survival prediction performance (best in **bold**).

Dataset	Metric	ResNet18	ResNet34	ResNet50	DenseNet121	SEResNet50	SEResNeXt50	NLL	Ours
Breast	C-index	0.677 ± 0.019	0.669 ± 0.048	0.599 ± 0.065	0.709 ± 0.027	0.644 ± 0.072	0.675 ± 0.036	0.685 ± 0.013	$\mathbf{0.716 \pm 0.011}$
	HR	2.684 ± 0.545	3.052 ± 1.226	1.416 ± 0.356	2.584 ± 0.468	2.302 ± 1.107	2.724 ± 0.601	1.094 ± 0.1418	$\mathbf{3.396 \pm 0.922}$
H & N	C-index	0.601 ± 0.035	0.625 ± 0.007	0.613 ± 0.022	0.596 ± 0.017	0.631 ± 0.013	0.608 ± 0.023	0.630 ± 0.037	$\mathbf{0.654 \pm 0.015}$
	HR	2.166 ± 0.475	1.690 ± 0.441	1.660 ± 0.289	1.476 ± 0.291	1.656 ± 0.227	1.914 ± 0.436	1.250 ± 0.276	$\mathbf{2.794 \pm 0.429}$

Table 2: Ablation study of the proposed framework components.

Dataset	Components			Metric	Backbones		
	Baseline	VCB	TAPRL		ResNet18	ResNet34	ResNet50
Breast Cancer	✓			C-index	0.677 ± 0.019	0.669 ± 0.048	0.599 ± 0.065
				HR	2.684 ± 0.545	$\mathbf{3.052 \pm 1.226}$	1.416 ± 0.356
	✓	✓		C-index	0.701 ± 0.029	$\mathbf{0.695 \pm 0.042}$	0.633 ± 0.061
				HR	1.924 ± 0.336	2.862 ± 0.840	1.920 ± 0.564
	✓	✓	✓	C-index	$\mathbf{0.716 \pm 0.011}$	0.672 ± 0.005	$\mathbf{0.654 \pm 0.071}$
				HR	$\mathbf{3.396 \pm 0.922}$	1.574 ± 0.217	$\mathbf{2.302 \pm 0.654}$
Head & Neck Cancer	✓			C-index	0.601 ± 0.035	0.625 ± 0.007	0.613 ± 0.022
				HR	2.166 ± 0.475	1.690 ± 0.441	1.660 ± 0.289
	✓	✓		C-index	0.637 ± 0.015	0.637 ± 0.026	0.629 ± 0.017
				HR	2.622 ± 0.015	$\mathbf{2.848 \pm 1.077}$	1.430 ± 0.355
	✓	✓	✓	C-index	$\mathbf{0.654 \pm 0.015}$	$\mathbf{0.649 \pm 0.017}$	$\mathbf{0.634 \pm 0.016}$
				HR	$\mathbf{2.794 \pm 0.429}$	2.460 ± 0.441	$\mathbf{1.684 \pm 0.323}$

Optimization Objective The final objective is formulated as a weighted logistic ranking loss (referred to as TAPRL), following learning-to-rank formulations such as RankNet [2]. For the current batch, we minimize:

$$\mathcal{L}_{\text{Rank}} = \frac{1}{N_e} \sum_{i \in \mathcal{E}} \frac{1}{Z_i} \sum_{j \in \mathcal{P}} w_{ij}(t_j - t_i) \cdot \log\left(1 + \exp\left(-(s_i - s_j)\right)\right), \tag{7}$$

where $\mathcal{E} = \{i \in \mathcal{B} \mid e_i = 1\}$ is the set of events in the current batch with cardinality $N_e = |\mathcal{E}|$, s_i is the score from the Main Encoder, s_j is the score from the Reference Encoder (if $j \in \mathcal{Q}$) or Main Encoder (if $j \in \mathcal{B}$), and $Z_i = \sum_j w_{ij}$ is a normalization factor.

Training Strategy To prevent cold-start instability, we employ a warm-up strategy. For the initial epochs, the model is trained using the standard Cox loss on the local mini-batch only; then, the Reference Encoder and Virtual Cohort mechanism are activated to refine global risk stratification. At this stage, TAPRL is applied on the merged cohort to refine global risk stratification.

3 Experiments and Results

3.1 Datasets and Evaluation Metric

Datasets The first dataset consists of 1,210 breast cancer patients from our private collection(I–III stage patients, the maximum follow-up time was 13 years,

Table 3: Ablation study on the VCB size K using ResNet18 backbone.

Dataset	Metric	$K = 32$	$K = 64$	$K = 128$
Head & Neck Cancer	C-index	0.641 ± 0.022	$\mathbf{0.654 \pm 0.015}$	0.608 ± 0.008
	HR	2.156 ± 0.639	$\mathbf{2.794 \pm 0.429}$	1.606 ± 0.238

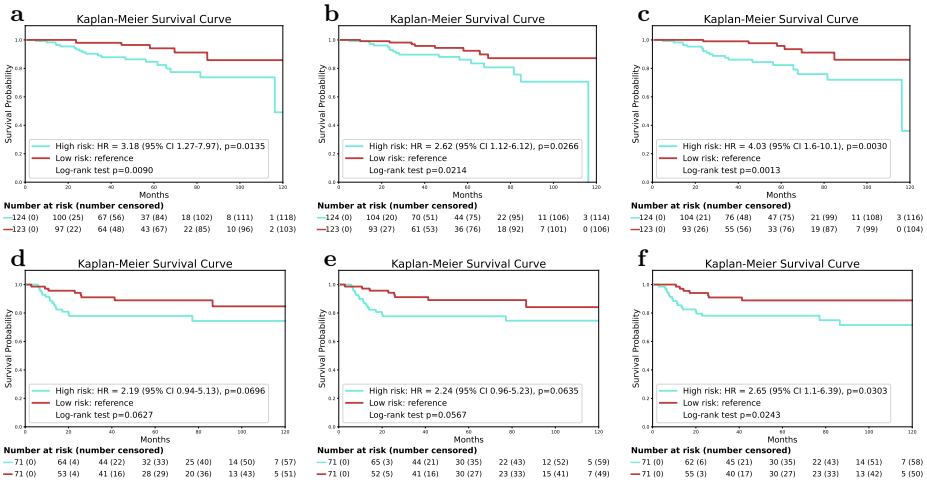

Fig. 2: Kaplan–Meier survival curves for risk stratification. Panels a–c show results on the breast cancer cohort, and panels d–f on the Head & Neck cohort. Columns correspond to models: ResNet34, SEResNeXt50, and our method.

the event rate was 9.8%), while the second is a public Head & Neck cancer dataset comprising 678 patients [19]. For each cohort, we randomly held out 20% of patients for testing. The remaining data was split 80% for training and 20% for validation. All compared models used the same mask-guided input modulation, ensuring fair comparison.

Evaluation Metric We evaluated the survival prediction performance using the Concordance Index (C-index). To assess risk stratification, we employed Kaplan-Meier (KM) survival curves, Log-rank tests (p-value), and Hazard Ratios (HR) with 95% Confidence Intervals (CI).

3.2 Implementation Details

Input volumes were standardized to $192 \times 192 \times 192$ for the breast cancer dataset and $200 \times 200 \times 300$ for the H&N cancer dataset. During training, we adopted the AdamW [16] optimizer with a batch size of 4, an initial learning rate of 1×10^{-4}, and a weight decay of 0.1. The learning rate was decayed by a factor of 0.1 every 3 epochs over 10 total epochs. For the MCSF framework, the VCB size was set to $K = 64$ with a momentum coefficient $m = 0.995$. All experiments

were conducted on a single NVIDIA RTX A6000 GPU using PyTorch. Each experiment was repeated five times and results were averaged.

3.3 Results

Comparison with Representative Models We compared our proposed MCSF with several widely adopted and strong baselines that remain prevalent in 3D medical imaging. The models compared include ResNet [8], DenseNet [11], and SEResNet [9], as well as a ResNet18 backbone trained with the negative log-likelihood loss (NLL) following previous works [22, 21, 24].

As shown in Table 1, our method consistently outperforms the baseline models in both datasets. On the Breast Cancer dataset, our model achieves a C-index of 0.716, surpassing the second-best DenseNet121. In the H&N Cancer dataset, our model achieves a C-index of 0.654, outperforming the best baseline (SEResNet50 with 0.631). For each method, we obtained a continuous risk score from the model and stratified patients into high- and low-risk groups using the median score as the cutoff. KM curves were plotted for the two groups and compared using the log-rank test. As shown in Fig. 2, our method yields clearer separation between the high- and low-risk survival curves. Notably, unlike ResNet34 and SEResNeXt50 which show non-significant results on the H&N dataset, our framework demonstrates superior risk stratification capabilities, with KM curves further providing supportive qualitative evidence. Compared to the ResNet-18 baseline, MCSF increases parameters (8 M to 17 M), forward FLOPs (285 G to 570 G), and GPU memory (6 G to 8 G). However, inference requires only the main encoder, incurring the same computational cost as the baseline. These results highlight the effectiveness of the MCSF in improving survival prediction performance by expanding the global risk pool, even with limited batch sizes.

Ablation Study We conducted an ablation study to evaluate the contributions of VCB and TAPRL across three backbones (Table 2). Overall, adding VCB consistently improves performance over the baseline, and incorporating TAPRL further boosts it, yielding the best results in most settings. The strongest gains are observed with ResNet18, where the full model reaches a C-index of 0.716 on Breast Cancer and 0.654 on H&N Cancer. We further study the effect of the VCB size K (Table 3). While enlarging the bank generally improves the approximation of the global Cox risk set, the benefit is not monotonic: $K = 64$ achieves the best C-index and HR, whereas a larger bank ($K = 128$) degrades performance. This suggests that an overly large bank induces feature drift by introducing stale prototypes, which dilutes informative comparisons and amplifies ranking noise.

4 Conclusion

In this study, we present the MCSF for 3D PET-CT survival prediction under the small-batch regime. Motivated by the mismatch between the global risk-set formulation of CoxPH and the severely limited in-batch risk set induced by GPU

memory constraints, MCSF maintains a momentum-updated VCB to construct a large and consistent risk pool. We also introduce a TAPRL that leverages event-time separation to downweight ambiguous pairs and emphasize informative comparisons, thereby stabilizing optimization and improving risk ordering. Experiments on two datasets demonstrate strong risk stratification across multiple 3D backbones. Overall, MCSF provides a practical and model-agnostic solution for scalable 3D imaging-based survival learning, and may facilitate more reliable prognostic modeling to support individualized oncology decision making.

Acknowledgements

This study is supported by Shenzhen Medical Research Fund (D2501013). This work is supported by Macao Polytechnic University Grant (RP/FCA-13/2026). This work was supported by the Science and Technology Development Fund, Macau SAR (File no. 0004/2025/ASJ) under the FDCT-FAPESP Joint Funding Scheme. Xinglong Liang was funded by Chinese Scholarship Council scholarship.

Disclosure of Interests

The authors declare no competing interests.

References

1. Andrearczyk, V., Oreiller, V., Abobakr, M., Akhavanallaf, A., Balermpas, P., Boughdad, S., Capriotti, L., Castelli, J., Cheze Le Rest, C., Decazes, P., et al.: Overview of the hecktor challenge at miccai 2022: automatic head and neck tumor segmentation and outcome prediction in pet/ct. In: 3D Head and Neck Tumor Segmentation in PET/CT Challenge, pp. 1–30. Springer (2022)
2. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on Machine learning. pp. 89–96 (2005)
3. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
4. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
5. Endo, K., Oriuchi, N., Higuchi, T., Iida, Y., Hanaoka, H., Miyakubo, M., Ishikita, T., Koyama, K.: Pet and pet/ct using 18 f-fdg in the diagnosis and management of cancer patients. International journal of clinical oncology **11**, 286–296 (2006)
6. Groheux, D., Espié, M., Giacchetti, S., Hindié, E.: Performance of fdg pet/ct in the clinical management of breast cancer. Radiology **266**(2), 388–405 (2013)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Huang, B., Sollee, J., Luo, Y.H., Reddy, A., Zhong, Z., Wu, J., Mammarappallil, J., Healey, T., Cheng, G., Azzoli, C., et al.: Prediction of lung malignancy progression and survival with machine learning based on pre-treatment fdg-pet/ct. EBioMedicine **82** (2022)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology **18**(1), 24 (2018)
13. Kvamme, H., Borgan, Ø., Scheel, I.: Time-to-event prediction with neural networks and cox regression. Journal of machine learning research **20**(129), 1–30 (2019)
14. Lee, C., Zame, W., Yoon, J., Van Der Schaar, M.: Deephit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
15. Liang, X., Zhang, T., Braga, M., Han, L., Donswijk, M., Huang, J., Song, J., Lu, C., Wang, X., Gao, Y., et al.: Multi-omics deep learning improves fdg pet-ct-based long-term prognostication of breast cancer. npj Precision Oncology (2026)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Meng, M., Gu, B., Bi, L., Song, S., Feng, D.D., Kim, J.: Deepmts: Deep multitask learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment pet/ct. IEEE Journal of Biomedical and Health Informatics **26**(9), 4497–4507 (2022)
18. Qian, C., Jiang, C., Xie, K., Ding, C., Teng, Y., Sun, J., Gao, L., Zhou, Z., Ni, X.: Prognosis prediction of diffuse large b-cell lymphoma in ^{18}F-FDG f-fdg pet images based on multi-deep-learning models. IEEE Journal of Biomedical and Health Informatics **28**(7), 4010–4023 (2024)
19. Saeed, N., Hassan, S., Hardan, S., Aly, A., Taratynova, D., Nawaz, U., Khan, U., Ridzuan, M., Andrearczyk, V., Depeursinge, A., et al.: A multimodal and multicentric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. arXiv preprint arXiv:2509.00367 (2025)
20. Weber, M., Kersting, D., Umutlu, L., Schäfers, M., Rischpler, C., Fendler, W.P., Buvat, I., Herrmann, K., Seifert, R.: Just another "clever hans"? neural networks and fdg pet-ct to predict the outcome of patients with breast cancer. European journal of nuclear medicine and molecular imaging **48**(10), 3141–3150 (2021)
21. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21241–21251 (2023)
22. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. IEEE transactions on pattern analysis and machine intelligence **43**(9), 3126–3137 (2020)
23. Zeng, L., Tang, W., Ren, Z., Ding, Y.: Optimizing cox models with stochastic gradient descent: Theoretical foundations and practical guidances. arXiv preprint arXiv:2408.02839 (2024)
24. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In: The Twelfth International Conference on Learning Representations (2024)