HuLP: Human-in-the-Loop for Prognosis

Muhammad Ridzuan[0000−0003−0935−8466], Mai A. Shaaban[0000−0003−1454−6090], Numan Saeed[0000−0002−6326−6434], Ikboljon Sobirov[0000−0002−0476−6359], and Mohammad Yaqub[0000−0001−6896−1105]

Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{Muhammad.Ridzuan, Mai.Kassem, Numan.Saeed, Ikboljon.Sobirov, Mohammad.Yaqub}@mbzuai.ac.ae

Abstract. This paper introduces HuLP, a Human-in-the-Loop for Prognosis model designed to enhance the reliability and interpretability of prognostic models in clinical contexts, especially when faced with the complexities of missing covariates and outcomes. HuLP offers an innovative approach that enables human expert intervention, empowering clinicians to interact with and correct models' predictions, thus fostering collaboration between humans and AI models to produce more accurate prognosis. Additionally, HuLP addresses the challenges of missing data by utilizing neural networks and providing a tailored methodology that effectively handles missing data. Traditional methods often struggle to capture the nuanced variations within patient populations, leading to compromised prognostic predictions. HuLP imputes missing covariates based on imaging features, aligning more closely with clinician workflows and enhancing reliability. We conduct our experiments on two real-world, publicly available medical datasets to demonstrate the superiority and competitiveness of HuLP. Our code is available at https://github.com/BioMedIA-MBZUAI/HuLP.

Keywords: Prognosis · Survival analysis · Interactive · Human-in-the-loop

1 Introduction

Diagnosis and prognosis play pivotal roles in oncology, yet prognosis presents a unique challenge due to its heightened uncertainty and complex nature. Unlike diagnosis, which primarily focuses on confirming the presence of cancerous cells or tumors [20], prognosis entails predicting the trajectory of the disease, including survival time and likelihood of recurrence [7]. This complexity arises from various factors that influence disease progression and outcome, ranging from tumor characteristics to patient demographics and treatment efficacy [21], making prognosis more challenging for clinicians to assess accurately.

While deep learning models are emerging as clinical assistants in prognosis, current approaches face two significant problems in the clinical setting. First, the models leave no space for clinicians to intervene, even when the models are incorrect or less confident, thus limiting the clinicians' ability to provide valuable inputs or improve the models' predictions. Related works allowing human intervention [1,5] are applied to natural images but not used for prognosis. During inference, the models can benefit from such feedback to improve their overall performance, mimicking how doctors collaborate and refine their assessment based on collective expertise. Presently, there is a gap in established methodologies (e.g., [13,12,18]) that enable active human interaction and intervention to refine the model's predictions of clinical features to improve prognosis.

Second, in cancer prognosis, dealing with incomplete data and censored patient outcomes (i.e., instances for which we do not know the exact event time) is challenging. Missing covariates may result from incomplete collection [11], noncompliance [14], or technical errors [11], while missing outcomes may arise due to patients discontinuing follow-up visits [18], relocating [19], or withdrawing from a study [19]. Standard practice in AI research typically uses naive imputation methods such as statistical measures of central tendency (i.e., mean, median, mode), k-nearest neighbor, or more algorithmic approaches, such as multiple imputation by chained equations (MICE). However, in reality, oncologists rely on radiological images to gain more insights into the patients' conditions [9].

The use of electronic health records (EHR) alone in prognosis often falls short of capturing the complex variability among individuals within and across different medical contexts, especially in static non-temporal EHR datasets. For instance, consider two individuals with identical clinical profiles, both diagnosed with lung cancer; despite sharing similar clinical information, their survival outcomes can exhibit significant disparities. Table 1 highlights several such real-world cases from the ChAImeleon [4] lung cancer dataset. Traditional models trained solely on EHR data struggle to reliably distinguish such variations in survival. Notably, the integration of radiological images – which provides a richer manifestation of temporal information, including age, smoking status, and tumor texture and characteristics – offers a promising avenue for capturing data dynamics that are often overlooked in static clinical data.

In response to these challenges, we introduce **Human-in-the-Loop for Prognosis (HuLP)**, a deep learning architecture inspired by [1,5] designed to enhance the reliability and interpretability of prognostic models in clinical settings. Our main contributions are twofold:

- *Allowing user interaction and expert intervention at inference time:* HuLP facilitates human expert intervention during model inference, empowering clinicians to provide input and guidance based on their domain expertise. This capability significantly enhances the model's decision-making process, particularly in complex prognostic scenarios where expert knowledge is invaluable.
- *Capability of handling both missing covariates and outcomes and extraction of meaningful vector representations for prognosis:* HuLP is equipped with a robust mechanism for handling missing data, ensuring end-to-end reliability in prognostic predictions. By leveraging patients' clinical information as intermediate concept labels, our model generates richer representations of clinical features, thereby enhancing prognostic accuracy.

Table 1. Variability in patient survival times from the ChAImeleon [4] lung cancer dataset for a given set of identical covariates. Event "1" signifies the patient's death at the given time (*uncensored*); event "0" signifies that the patient is alive at (least until) the given time (*censored*). "X" represents unknown or missing data.

Age	**Gender**	**Smoking Status**	**T-stage**	**N-stage**	**M-stage**	**Survival (months)**	**Event**
Patients with the same TNM							
70	Male	Ex-smoker	T4	N3	M1c	3.50	1
70	Male	Ex-smoker	T4	N3	M1c	1.20	1
Patients with missing TN							
72	Male	Ex-smoker	X	X	M1	15.23	0
72	Male	Ex-smoker	X	X	M1	5.17	0
Patients with missing TNM							
57	Male	Smoker	X	X	X	3.50	1
57	Male	Smoker	X	X	X	56.53	0
67	Female	Ex-smoker	X	X	X	4.27	0
67	Female	Ex-smoker	X	X	X	58.27	0

2 Methodology

Figure 1 shows the complete architecture of the proposed HuLP model, which is comprised of four main components: **encoder**, **intervention block**, **classifier**, and **prognosticator**.

The **encoder** $\xi(\cdot)$ is a deep neural network (e.g. CNN or transformer) that processes an image $x \in \mathbb{R}^{H\times W\times D\times C}$, where H, W, D and C are the height, width, depth, and number of channels of the input, respectively, and generates a latent feature embedding $y \in \mathbb{R}^{(K\times M)}$. Here, K represents the embedding space dimension, while M is the total number of unique and discrete patient characteristics (*concepts*) across all clinical features (*parent categories*) P in EHR, i.e. $M = \sum_{j=1}^{|P|} m_j$, where m_j denotes the number of unique concepts m in each feature j. Continuous features are discretized. This embedding output plays a foundational role in capturing essential features from the image and is designed to learn a shared representation. It is then passed through M groups of fully connected (FC) layers $\alpha(\cdot)$ to produce concept embeddings c.

The **intervention block**, a key component of HuLP, enables user interaction and expert intervention at test time. During training, each group of embeddings $c_i = \alpha_i(y)$ undergoes a single-neuron sigmoid weighting function, producing the probability p of a concept being active ($p_i = \sigma(c_i)$). The embeddings are then split arbitrarily into two halves, c_i^+ and c_i^-, and multiplied by p_i and $(1-p_i)$ to represent the latent positive and negative concept embeddings, respectively. To facilitate test-time intervention, p is randomly replaced with the hard ground-truth labels $[0, 1]$ with a probability of 0.25. During inference, a human may impart his/her domain knowledge by replacing p with [0,1] to indicate certainty in the presence or absence of a concept. The final concept embedding is generated as the sum of the positive and negative embeddings ($c_{Fi} = p_i c_i^+ + (1-p_i)c_i^-$).

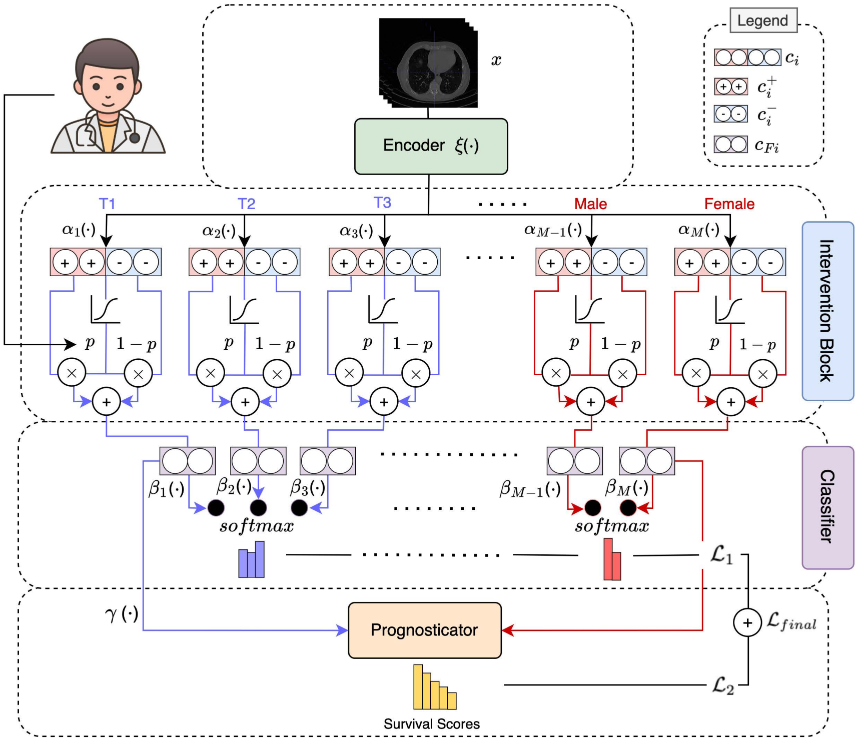

Fig. 1. HuLP is composed of (1) a deep learning **encoder** that extracts features from medical images; (2) an **intervention block** that allows human intervention during test time; (3) a **classifier** that ensures concept alignment of the feature embeddings; and (4) a **prognosticator** that performs survival prediction. Here, *(T1, T2, T3)* and *(Male, Female)* are example clinical concepts obtained from the parent categories *T-stage* and *gender*, respectively. Our loss is a combination of the concept loss L_1 applied on the classifier and prognosis loss L_2 applied on the prognosticator.

The **classifier** is an FC layer $\beta(\cdot)$ that encourages concept alignment of the embeddings by enforcing each embedding layer to predict only one concept (i.e. $\beta_i(c_{Fi}) \in \mathbb{R}^1$). This is followed by a softmax and cross-entropy loss for the concepts under each parent category. The classifier is designed to allow missing data (see *Loss function* for details).

Finally, the **prognosticator** processes the concepts $w = [c_{F1}, c_{F2}, ..., c_{FM}]$ through an FC layer $\gamma(\cdot)$ with n discrete time bins. The softmax outputs of the layer represent the estimated hazard for each patient, indicating the instantaneous rate of an event (e.g., death or cancer recurrence) conditioned on surviving up to time t for a given patient concept vector w. The prognosticator layer is responsible for predicting the progression of clinical outcomes over time. The pseudo-code of the model is described in Algorithm 1 in the *Appendix*.

Loss function. The proposed loss is a combination of the concept loss $\mathcal{L}_1$ and prognosis loss $\mathcal{L}_2$. The concept loss $\mathcal{L}_1$ is applied on the classifier layer using cross-entropy and is computed as an average over all non-missing covariates (Eq. 1). This loss is backpropagated for patients with non-missing covariates but skipped for patients with missing covariates, thus providing the advantage of avoiding hard imputation of missing data prior to training.

$$\mathcal{L}_1 = -\frac{1}{\hat{N}} \sum_{i=1}^{M} \sum_{j=1}^{\hat{N}} y_{ij} \log(\beta_i(c_{Fij})) \tag{1}$$

where M is the number of discrete clinical categories (see *encoder*) and $\hat{N}$ is the number of patients with non-missing data.

The prognosis loss $\mathcal{L}_2$ is applied to the prognosticator. It is a slightly modified version of the DeepHit [13] loss function for a single non-competing risk. Given time T, event indicator E, and concept vector w, we convert the outputs of HuLP into an estimated survival function $\hat{S}$ using

$$\hat{S}(T \mid w) = exp(-\hat{H}(T \mid w)) = exp(-\sum_{t=1}^{T} \hat{h}(t \mid w)) \tag{2}$$

where $\hat{H}$ is the cumulative hazard function and $\hat{h}$ is the estimated hazard from the softmax outputs of the model. $\mathcal{L}_2$ is thus defined as a weighted average of the discrete log-likelihood and rank losses:

$$\mathcal{L}_2 = a\text{loss}_{\mathcal{LL}} + (1-a)\text{loss}_{\text{rank}} \tag{3}$$

where

$$loss_{\mathcal{LL}} = -\sum_{i=1}^{N} \Big[E_i \log(\hat{h}_{e_i}(T_i \mid w_i) + (1-E_i)\log(\hat{S}(T_i \mid w_i)) \Big] \tag{4}$$

and

$$loss_{\text{rank}} = \sum_{i,j} E_i \, \mathbb{1}\{T_i < T_j\} \exp\left(\frac{\hat{S}(T_i \mid w_i) - \hat{S}(T_j \mid w_j)}{c} \right) \tag{5}$$

N is the total number of patients, e_i is the index of the event time for observation, and c is set to a constant 0.1 following [13].

The log-likelihood (Eq. 4) captures information regarding the time of the event and its occurrence for uncensored patients, and the time at which the patient was lost to follow-up (indicating that the patient was alive up to that time) for censored patients. The ranking loss (Eq. 5) compares the survival scores between possible pairs i, j of patients to incentivize the correct ordering of pairs.

The final loss is calculated using:

$$\mathcal{L}_{final} = b\mathcal{L}_1 + (1-b)\mathcal{L}_2 \tag{6}$$

a (in Eq. 3) and b are weighting hyperparameters.

3 Experimental Setup

3.1 Datasets

The prognostic ability of HuLP is assessed by comparing it with conventional benchmarks in analyzing two real-world medical datasets: ChAImeleon [4] and HECKTOR [2,16]. Below, we provide a brief overview of each.

The ChAImeleon [4] lung cancer dataset consists of 320 patient CT scans with EHR. The clinical features include age, gender, smoking status, clinical category (T-stage), regional nodes category (N-stage), and metastasis category (M-stage), with up to 26% missing and 59% censored data. For preprocessing, we combined all missing data labels, i.e. "Unknown", nan, "TX", and "NX" into the same category "X" for each feature. All cancer sub-stages are combined into their parent stage to increase the number of samples per category (e.g. T1a, T1b, T1c are combined into T1). We use a publicly available segmentation model [8] to restrict the ROI to the lung areas.

HECKTOR [2,16] is a multi-modal head-and-neck cancer dataset comprising 224 CT and PET scans with EHR. The PET and CT scans are registered to a common origin. The clinical features include center, age, gender, TNM 7/8th edition staging and clinical stage, tobacco and alcohol consumption, performance status, HPV status, and treatment (chemoradiotherapy or radiotherapy only), with up to 90% missing and 75% censored data. Features with over 80% missing data are dropped, and all cancer sub-stages are combined into their parent stage. To standardize the inputs, the scans are preprocessed in the same manner for each dataset via resampling, cropping, and resizing.

3.2 Implementation Details

HuLP is run for 100 epochs using DenseNet-121 [10] as the encoder. We use positive/negative embeddings of size 64, combined to form a final concept embedding of size 32. The prognosticator outputs 12 discrete time bins for ChAImeleon [4] and 16 for HECKTOR [2,16], obtained as the square root of the number of observations corresponding to the quantiles of the survival time distribution. We use a batch size of 32, AdamW [15] optimizer with a learning rate of 1×10^{-3}, and a cosine annealing scheduler with a warmup of 5 epochs. All experiments are implemented using PyTorch [17].

We compare HuLP against three deep survival methods: DeepHit [13], Deep-MTLR [6], and Fusion [18]. DeepHit [13] and Deep-MTLR [6] are chosen because they are both top-performing discrete survival methods in prognosis; similarly, our HuLP implementation is also discrete. Fusion [18] is chosen as a multimodal baseline and also because it won the HECKTOR [2,16] competition, the same dataset used in this work.

DeepHit [13] and Deep-MTLR [6] (EHR) are run using mode imputation with two FCs of size 64 each followed by a ReLU activation, batch normalization and dropout with probability 0.1, and a prognosticator using a batch size of 96 and learning rate of 1×10^{-2}. We also compare our method against a variant of

DeepHit [13] and DeepMTLR [6] using imaging data as inputs and DenseNet-121 [10] as the encoder to directly predict survival outcomes. Finally, we implement the idea of Fusion [18] by extracting imaging features using DenseNet-121 [10], concatenated with EHR through a late fusion technique. We maintain a constant ratio of patients who experienced each event and those who were censored in each fold. The experiments are repeated with two seeds and five-fold cross-validation.

4 Results

We report the time-dependent concordance (C-index) of Antolini et al. [3] from the survival curves. Table 2 summarizes our results. EHR without images presents the problem of limited depth and richness of static information. Images without EHR leave the model unguided. HuLP consistently demonstrates statistically significant improvements (p-value<0.05) over these methods and remains competitive with Fusion [18]; however, the learning of fusion from EHR and image embeddings was disjoint. HuLP distinguishes itself by integrating EHR as an intermediate concept labeling that guides the model towards the relevant features, thus producing rich, disentangled embeddings of the clinical features from the images with two added advantages: it allows human expert intervention during test time and is robust to missing data.

To emulate human intervention, p is fully replaced with ground-truth labels for non-missing data, while p is retained for missing data. We run inference on the validation set with and without test-time intervention. Notably, the integration of user interaction and expert intervention of the clinical concepts significantly enhances the model's prognostic capabilities (Table 3), yielding an improvement of about 0.11 C-index on ChAImeleon [4].

To investigate HuLP's robustness to missing data, we create a challenging 8:2 train-validation split stratified by gender where each patient in the validation split has identical or similar EHR as at least one other patient and at least one missing data. We randomly mask entries in the training EHR with increasing probabilities to emulate situations with missing data on the ChAImeleon

Table 2. Average concordance indices on two seeds and five-fold cross-validation. The highest scores per dataset are bolded. (*) is shown for statistically significant experiments (p-value $<$ 0.05) based on the average performance of HuLP and the best-performing baseline.

	Modality	ChAImeleon Lung Cancer [4]	HECKTOR [2,16]
DeepHit [13]	EHR	0.6522* ± 0.0371	0.6054* ± 0.1047
DeepMTLR [6]	EHR	0.6624* ± 0.0643	0.6085* ± 0.0985
DeepHit [13]	Image	0.6328* ± 0.0559	0.7144 ± 0.0269
DeepMTLR [6]	Image	0.6400* ± 0.0361	0.6222* ± 0.0788
Fusion [18]	Image+EHR	**0.7399** ± 0.0534	0.7012 ± 0.0457
HuLP (ours)	Image+EHR	0.7124 ± 0.0533	**0.7329** ± 0.0415

Table 3. Effect of test-time concept interventions on concordance index scores for ChAImeleon [4]. The results shown are for five-fold cross-validation with two seeds.

With test-time interv.	Without test-time interv.
0.7124 ± 0.0533	0.6060 ± 0.0441

Table 4. Effect of different imputation methods on concordance index scores on ChAImeleon [4]. The results shown are the averages of three seeds. The highest scores per column are bolded.

	Missing data percentage			
Imputation	30%	40%	50%	70%
Mode	0.5817	0.6563	0.6401	0.6430
kNN (k=1)	0.6068	0.6526	0.6556	0.6585
MICE	0.6068	0.6275	0.6541	0.6371
HuLP (ours)	**0.6297**	**0.6748**	**0.6740**	**0.6593**

[4] dataset. Table A1 in the *Appendix* details the distribution of the validation split. We run our experiments for three seeds and compare our method against conventional benchmarks for imputation, including mode, kNN, and MICE. Table 4 presents the results of our experiments, showing our method's robustness to missing data, particularly in the low missing-data regime. At high missing-data regime, the improvement becomes less significant, likely because the model receives inadequate feedback from $\mathcal{L}_1$ to capture the semantic meaning of the concepts. However, the results remain competitive with the baselines.

5 Discussion

To our knowledge, HuLP is the first prognostic model that allows human interaction and intervention of known concepts for prognosis. This innovation represents a significant advancement particularly in prognosis, where predicting future outcomes can often be more challenging than diagnosing present conditions. Compared to methods where human experts are passive users, HuLP empowers clinicians to actively engage with the model, refining its concept predictions based on their domain expertise. This collaborative approach fosters a synergistic relationship between humans and computers, allowing each to leverage their strengths. Clinicians, with their deep understanding of clinical features, can provide refined adjustments to the model's predictions regarding the presence or absence of certain concepts, while HuLP dynamically incorporates these inputs to enhance the accuracy of prognostic assessments. This active collaboration not only improves the interpretability and reliability of prognostic models but also instills confidence in their use in clinical decision-making.

Additionally, in addressing the challenge of missing data, HuLP presents a tailored methodology that surpasses conventional approaches like mode, kNN, and MICE. These methods, while widely used, often oversimplify the complexity of clinical datasets and may introduce bias, thereby compromising the validity of prognostic predictions. In contrast, HuLP harnesses the power of neural networks to better accommodate the nuances of missing data in prognostic modeling. In particular, during test time, HuLP implicitly imputes the missing covariates based on the imaging features rather than relying on a simplistic hard imputation. This aligns more closely with clinician workflows and enhances the reliability and trustworthiness of prognostic assessments.

6 Conclusion

This paper presents HuLP, Human-in-the-Loop for Prognosis, an innovative approach that allows clinicians to interact with and intervene in model predictions at test time, enhancing prognostic model reliability and interpretability in clinical settings. HuLP extracts meaningful representations from imaging data and can effectively handle missing covariates and outcomes. Experimental results on two medical datasets demonstrate HuLP's superior and competitive performance. Future work should focus on validating HuLP in clinical settings with clinical inputs and exploring the usability of the disentangled feature embeddings.

Acknowledgments. De-identified data used in the development of this solution were prepared and provided by the ChAImeleon Project within the Open Challenges organized by its Consortium, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172, entitled ‘Accelerating the lab to market transition of AI tools for cancer management.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Promises and Pitfalls of Black-Box Concept Learning Models, vol. 1 (2021)
2. Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C.C.L., Elhalawani, H.M., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., Hatt, M., Depeursinge, A.: Overview of the hecktor challenge at miccai 2021: Automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In: HECKTOR@MICCAI (2022), `https://api.semanticscholar.org/CorpusID:245877569`
3. Antolini, L., Boracchi, P., Biganzoli, E.: A time-dependent discrimination index for survival data. Stat. Med. **24**(24), 3927–3944 (Dec 2005)
4. Bonmatí, L.M., Miguel, A., Suárez, A., Aznar, M., Beregi, J.P., Fournier, L., Neri, E., Laghi, A., França, M., Sardanelli, F., Penzkofer, T., Lambin, P., Blanquer, I., Menzel, M.I., Seymour, K., Figueiras, S., Krischak, K., Martínez, R., Mirsky, Y., Yang, G., Alberich-Bayarri, Á.: CHAIMELEON project: Creation of a pan-european repository of health imaging data for the development of AI-powered cancer management tools. Front. Oncol. **12** (Feb 2022)

5. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lio, P., Jamnik, M.: Concept embedding models: Beyond the accuracy-explainability trade-off. Advances in Neural Information Processing Systems **35** (2022)
6. Fotso, S.: Deep neural networks for survival analysis based on a multi-task framework. ArXiv **abs/1801.05512** (2018), `https://api.semanticscholar.org/CorpusID:13482950`
7. Glare, P., Sinclair, C., Downing, M., Stone, P., Maltoni, M., Vigano, A.: Predicting survival in patients with advanced disease. European Journal of Cancer **44**(8), 1146–1156 (2008)
8. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Exp. **4**(1), 50 (Aug 2020)
9. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. Nature Reviews Cancer **18**(8), 500–510 (2018)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017). `https://doi.org/10.1109/CVPR.2017.243`
11. Janssen, K.J., Donders, A.R.T., Harrell Jr, F.E., Vergouwe, Y., Chen, Q., Grobbee, D.E., Moons, K.G.: Missing covariate data in medical research: to impute is better than to ignore. Journal of clinical epidemiology **63**(7), 721–727 (2010)
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 Med. Res. Methodol. **18**(1) (Dec 2018)
13. Lee, C., Zame, W., Yoon, J., Van der Schaar, M.: DeepHit: A deep learning approach to survival analysis with competing risks. Proc. Conf. AAAI Artif. Intell. **32**(1) (Apr 2018)
14. Levy, D.E., O'Malley, A.J., Normand, S.L.T.: Covariate adjustment in clinical trials with non-ignorable missing data and non-compliance. Statistics in Medicine **23**(15), 2319–2339 (2004)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), `https://openreview.net/forum?id=Bkg6RiCqY7`
16. Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Vallières, M., Zhu, S., Xie, J., Peng, Y., Iantsen, A., Hatt, M., Yuan, Y., Ma, J., Yang, X., Rao, C., Pai, S., Ghimire, K., Feng, X., Naser, M.A., Fuller, C.D., Yousefirizi, F., Rahmim, A., Chen, H., Wang, L., Prior, J.O., Depeursinge, A.: Head and neck tumor segmentation in pet/ct: The hecktor challenge. Medical Image Analysis **77**, 102336 (2022). `https://doi.org/https://doi.org/10.1016/j.media.2021.102336`, `https://www.sciencedirect.com/science/article/pii/S1361841521003819`
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Saeed, N., Al Majzoub, R., Sobirov, I., Yaqub, M.: An ensemble approach for patient prognosis of head and neck tumor using multimodal data. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) Head and Neck Tumor Segmentation and Outcome Prediction. pp. 278–286. Springer International Publishing, Cham (2022)

19. Sparr, L.F., Moffitt, M.C., Ward, M.F.: Who returns and who stays away. Am J Psychiatry **150**, 801–805 (1993)
20. Thakor, A.S., Gambhir, S.S.: Nanooncology: the future of cancer diagnosis and therapy. CA: a cancer journal for clinicians **63**(6), 395–418 (2013)
21. Zugazagoitia, J., Guedes, C., Ponce, S., Ferrer, I., Molina-Pinelo, S., Paz-Ares, L.: Current challenges in cancer treatment. Clinical therapeutics **38**(7), 1551–1566 (2016)