Multi-task Deep Segmentation and Radiomics for Automatic Prognosis in Head and Neck Cancer

V. Andrearczyk, P. Fontaine, V. Oreiller, J. Castelli, M. Jreige, J.O. Prior, A. Depeursinge

Published in Predictive Intelligence in Medicine : 4th International Workshop, PRIME 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings, which should be cited to refer to this work.
DOI: 10.1007/978-3-030-87602-9_14

Multi-Task Deep Segmentation and Radiomics for Automatic Prognosis in Head and Neck Cancer

Vincent Andrearczyk[1], Pierre Fontaine[1,2], Valentin Oreiller[1,3], Joel Castelli[4], Mario Jreige[3], John O. Prior[3], and Adrien Depeursinge[1,3]

[1] University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
[2] Univ Rennes, CLCC Eugene Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
[3] Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[4] Radiotherapy Department,Cancer Institute Eugène Marquis, Rennes, France
vincent.andrearczyk@hevs.ch

Abstract. We propose a novel method for the prediction of patient prognosis with Head and Neck cancer (H&N) from FDG-PET/CT images. In particular, we aim at automatically predicting Disease-Free Survival (DFS) for patients treated with radiotherapy or both radiotherapy and chemotherapy. We design a multi-task deep UNet to learn both the segmentation of the primary Gross Tumor Volume (GTVt) and the outcome of the patient from PET and CT images. The motivation for this approach lies in the complementarity of the two tasks and the shared visual features relevant to both tasks. A multi-modal (PET and CT) 3D UNet is trained with a combination of survival and Dice losses to jointly learn the two tasks. The model is evaluated on the HECKTOR 2020 dataset consisting of 239 H&N patients with PET, CT, GTVt contours and DFS data (five centers). The results are compared with a standard Cox PET/CT radiomics model. The proposed multi-task CNN reaches a C-index of 0.723, outperforming both the deep radiomics model without segmentation (C-index of 0.650) and the standard radiomics model (C-index of 0.695). Besides the improved performance in outcome prediction, the main advantage of the proposed multi-task approach is that it can predict patient prognosis without a manual delineation of the GTVt, a tedious and time-consuming process that hinders the validation of large-scale radiomics studies. The code will be shared for reproducibility on our GitHub repository.

Keywords: Head and Neck Cancer · radiomics · automatic segmentation · deep learning

1 Introduction

Radiomics is the quantitative analysis of radiological images to obtain prognostic patient information [1]. The standard approach for radiomics involves the extraction of hand-crafted visual features from radiology images followed by prognostic

modeling. This standard approach generally requires manual annotations of Volumes Of Interest (VOIs), i.e. tumor region, to spatially localize the extraction of features.

In the past seven years, deep Convolutional Neural Networks (CNNs) (mostly variants of the UNet model [2]) have reached excellent results in medical image segmentation, including tumor segmentation [3, 4]. The prediction of patient outcome (e.g. survival) using CNNs, however, has received less attention or success. For this task, the number of observations is smaller than for segmentation tasks, in which each pixel/voxel is an observation, when compared to patient-wise observations for radiomics. The level of abstraction required to predict the patient outcome is also higher than in a segmentation task. For survival tasks, moreover, the loss requires comparing several pairs of observations to estimate concordance, thereby requiring many observations as compared to other classical losses (e.g. cross-entropy and Dice). Survival losses are therefore not particularly suited for deep CNNs training with mini-batches. As a consequence, training a (3D) deep CNN from scratch or based on a pretrained network with a few hundred patients for the prediction of a complex outcome with censored data is generally not as efficient and successful as extracting hand-crafted radiomics features and training a simple survival model, e.g. the Cox Proportional Hazards (CPH) model.

Early works on Artificial Neural Networks (ANNs) for survival problems were proposed to learn nonlinear relationships between prognostic features (not imaging features as used in radiomics) and the risk for a given outcome [5]. More recently, deep ANNs [6, 7] were shown to successfully model complex relationships between the radiomics features and their risk of failure. An extensive review is proposed in [8].

Radiomics studies that make use of CNNs have often focused on using automatic segmentations or deep features, obtained from a trained CNN, in a standard radiomics pipeline, e.g. in the context of brain tumor [9]. In lung cancer, CNNs trained to perform tumor segmentation in FluoroDeoxyGlucose-Positron Emission Tomography (FDG-PET) and Computed Tomography (CT) were shown to identify a rich set of survival-related features with remarkable prognostic value [10]. CNNs have also been trained end-to-end for classification radiomics tasks, e.g. in Head and Neck cancer (H&N) [11]; see [12] for a review on deep learning-based radiomics. A few recent works proposed the use of CNNs trained with a Cox loss for survival tasks. Zheng *et al.* [13] proposed to train a CNN with a Cox loss for the prediction of patient survival in the context of pancreatic ductal adenocarcinoma. A survival CNN was used in [14] to predict the outcome of patients with brain cancer from histopathology images and genomic biomarkers, as well as in [15] for a survival analysis of rectal cancer based on PET/CT images.

Multi-task learning is a well-studied sub-field, notably in robotics and autonomous driving [16, 17]. It was shown that learning related tasks in parallel while using a shared representation can improve the individual learning efficiency and prediction accuracy, as features learned for each task can help the learning of other tasks [16]. In medical imaging, multi-task training was recently

used for several tasks and modalities. A multi-task method was proposed in [18] for brain tumor segmentation in Magnetic Resonance Imaging (MRI), leveraging fully annotated images (segmentation task) and weakly annotated images (classification of presence or absence of a tumor). Again for brain tumor segmentation, Weninger *et al.* [19] combined three tasks that share the encoder, namely tumor segmentation, image reconstruction (auto-encoder), and the classification of presence/absence of enhancing tumor. Another multi-task deep learning model was developed in [20] to jointly identify COVID-19 patients and segment the lesions from chest CT images. In histopathology, multi-task was used to improve tumor tissue classification by combining the prediction of auxiliary clinically relevant features as well as a domain adversarial task [21]. We are, however, not aware of previous works that make use of multi-task deep learning combining segmentation and survival losses in a model that can be trained end-to-end, neither multi-task learning applied to PET-CT radiomics nor to H&N cancer.

We base our work on the idea that deep features learned for tumor segmentation may also be useful for the prediction of patient outcome. In particular, neuron activations triggered by the segmentation task are expected to spatially guide the network to extract prognostically relevant patterns in the tumor area. In addition, training the model for tumor segmentation as an auxiliary task to outcome prediction allows the network to better exploit the scarce and precious radiomics training data where one observation is one patient. In this work, we show the potential of our multi-task approach for the prediction of outcome for patients with H&N cancer from PET/CT images. The main task is the prediction of Disease-Free Survival (DFS), to which we combine the auxiliary task of segmentation of the primary Gross Tumor Volume (GTVt).

To summarize, the contribution of this work is to propose a fully automatic bi-modal 3D deep segmentation and prognostic model able to deal with survival data and to learn from relatively small datasets, thanks to the multi-task paradigm.

2 Methods

In this section, we present the different methods developed in our work and the dataset used for the experiments. The main contribution, i.e. multi-task deep radiomics, is presented in Section 2.2. All models are bi-modal as they take CT and PET images as input. An overview of the multi-task hybrid segmentation-deep radiomics architecture is depicted in Fig. 1. The code to reproduce the experiments will be shared on our GitHub repository[5].

2.1 Segmentation

As a first model, we use a multi-modal 3D UNet for segmentation based on the model developed in [22], winner of the HECKTOR 2020 challenge[6] [4]. The

[5] https://github.com/vandrearczyk

[6] The main difference is the reduced number of filters to be able to use a larger batch-size for the survival loss in Section 2.2.

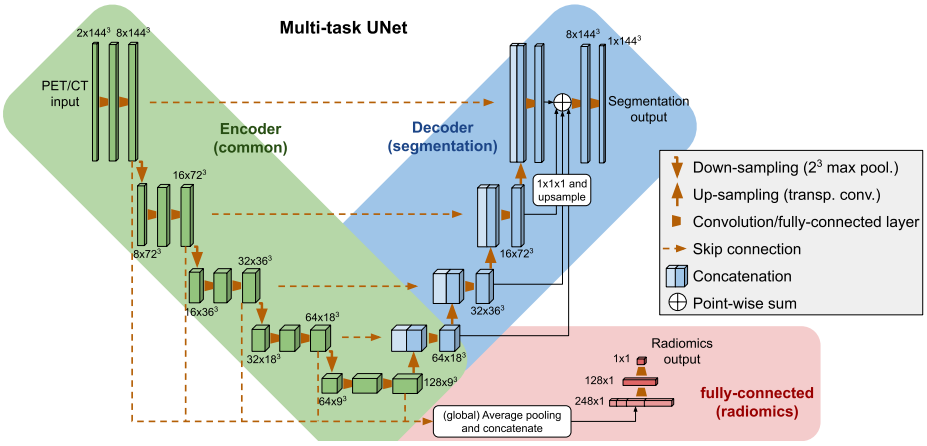

Fig. 1. 3D multi-modal (PET/CT) and multi-task architecture with a common down-sampling branch (green), an up-sampling segmentation branch (blue) and a radiomics branch (red). Residual convolutional layers are used in the down-sampling part.

architecture is presented in Fig. 1, including down-sampling (green) and up-sampling (blue) parts. The probabilities of the softmax activated outputs are thresholded at 0.5 to obtain a binary mask. More details on the implementation can be found in [22]. The model is trained with a Dice loss, computed as

$$\mathcal{L}_{Dice} = -2\frac{\sum_k \hat{y}_k y_k}{\sum_k \hat{y}_k + \sum_k y_k}, \tag{1}$$

where $\hat{y}_k \in [0,1]$ is the softmax output for a voxel k, $y_k \in \{0,1\}$ is the value of this voxel in the 3D ground truth mask and the sum is computed over all voxels.

2.2 Multi-task Segmentation and Radiomics

The multi-task architecture is composed of the normal 3D segmentation with an additional radiomics branch (red in Fig. 1) at the bottleneck of the network. This radiomics branch is connected to multiple layers of the downsampling path using skip connections to gather information at multiple scales and complexity. It is composed of a global average pooling layer to aggregate the spatial information, a densely connected layer (ReLU activated) with 128 neurons and a prediction layer with a single neuron. Dropout with 0.5 probability is added before the dense layers for regularization. For the segmentation task, we use the Dice loss defined in Eq. (1). For the radiomics task, we use a Cox loss [7] computed as

$$\mathcal{L}_{Cox} = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\hat{h}(x_i) - \log \sum_{j \in \mathcal{H}(T_i)} e^{\hat{h}(x_j)} \right), \tag{2}$$

where $\hat{h}(x_i)$ is the estimated log-hazard for input x_i. E_i, T_i, x_i are the event indicator (0 for censored data), the observed duration, and the baseline covariate data in the ith observation respectively. $\mathcal{H}(T_i)$ is the hazard indicator for individuals that have not experienced the event of interest before time T_i. The Cox loss is motivated by clinical needs such as group stratification based on risk. It alleviates the need to model the underlying hazard, a strength of the Cox model in this scenario.

For learning the radiomics task only, we use the same network as the multi-task one, without the up-sampling part of the UNet (i.e. only green and red parts in Fig. 1). The model is trained with the Cox loss defined in Eq. (2). Note that in this third method, the manually annotated contours are used neither during training nor testing.

2.3 Training Scheme

All networks are trained with an Adam optimizer based on a cosine decay learning rate (initial 10^{-3}) and a batch size of 10, for 60 epochs. The three training scenarios are obtained by modifying the pair of loss weights (w_{rad}, w_{seg}) in the combined loss $\mathcal{L} = w_{rad}\mathcal{L}_{Cox} + w_{seg}\mathcal{L}_{Dice}$: $(0, 1)$ for the segmentation, $(1, 1)$ for the multi-task, and $(1, 0)$ for the radiomics. For each of the three settings, the best model based on the combined loss obtained on the validation data is used for testing. The training data is resampled to balance the proportion of censored data for the Cox loss (82% of censored data on average before resampling and 50% after). Data augmentation is applied to the training data including random shifts (maximum 30 voxels), mirroring (sagittal axis) and rotations (maximum five degrees). The models are implemented in TensorFlow 2 and trained on a Nvidia V-100 GPU (32GB). The average training time of the multi-task model for one fold of the Cross-Validation (CV, see Section 2.6) is 148 min.

2.4 Standard Radiomics

We compare the deep algorithms with a state-of-the-art standard radiomics models based on hand-crafted features. We implemented the feature extraction and survival model in Python 3.9 with the following libraries: SimpleITK (2.0.2), PyRadiomics (3.0.1), scikit-learn (0.22.2) and scikit-survival (0.12.0). The computation time of the feature extraction and training times for one fold on an AMD Ryzen 7 3700X with 8 cores are 13 min and 3 seconds respectively.

Feature extraction We extract features inside the GTVt from the CT and PET images using PyRadiomics [23]. A total of 274 features are extracted for each patient, including 18 first-order and 112 texture features extracted per modality[7] as well as 14 additional shape-based features. Note that this approach, as

[7] 24 GLCM, 16 GLRLM, and 16 GLDZM features. A Fixed Bin Number (FBN) of 64 and a Fixed Bin Size (FBS) of 50 are used for CT. A FBN of 8 and a FBS of 1 are used for PET.

opposed to the deep learning counterparts presented above, requires test-time ground truth annotations of the primary tumors as VOIs to extract the features.

Survival model We first select features based on the univariate C-index estimated on the validation set (see Section 2.6 for the clarification of training, validation and test splits). This selection is recommended e.g. in [24] and is similar to an F1-score univariate selection. More precisely, we use a shifted version (i.e. |C-index-0.5|) to account for both concordant and anti-concordant features. The resulting top 20 features are kept. Second, correlated features are removed when higher than a given threshold value $t \in [0.6, 0.65, 0.70, 0.75, 0.80]$, optimized on the validation set using grid-search, as recommended e.g. in [25]. The resulting feature set is used by a CPH model [26] to predict the hazard score of the DFS outcome. The best performing model on the validation data is kept and used to predict hazards on the test set.

2.5 Evaluation

The radiomics results, i.e. predictions of DFS, are evaluated using the Concordance index (C-index) ranging from zero to one [27]. A C-index of one corresponds to the best model prediction, while a value of 0.5 is equivalent to a random prediction. The segmentation results are evaluated using the Dice Similarity Coefficient (DSC) ranging from zero to one, with one reflecting a perfect similarity between predicted and ground-truth contours. The DSCs are averaged across multiples cases as specified in the results section.

2.6 Dataset and Experimental Setup

For the experiments, we use the HECKTOR 2020 data including 239 cases from five centers [4]. This dataset was used for tumor segmentation and we propose for the first time to use it for prediction of patient outcome. Each case includes a CT and a PET image (inputs), a ground truth annotation of the GTVt and the DFS patient outcome information, i.e. time to recurrence of the GTVt following the treatment (outputs). The number of events is 43, i.e. 18% of non-censored cases, whereas the remaining 82% cases did not encounter the event during the follow-up. The average follow-up time is 1182 days.

The PET and CT images are resampled to $1 \times 1 \times 1$ mm ($2 \times 2 \times 2$ mm for the standard radiomics pipeline) using trilnear interpolation. The volumes are cropped to $144 \times 144 \times 144$ voxels (after augmentation described in Section 2.3) using the bounding boxes of the HECKTOR 2020 challenge [28]. The PET images are standardized individually to zero mean and unit variance. The CT images are clipped to [-1024,1024] and mapped to [-1,1]. A 5-fold CV is used for all experiments. For each fold, 20% of the dataset used for testing and the remainder is split again randomly as 80% training, 20% validation.

3 Results

Performance results for radiomics, i.e. comparison of models on the DFS prediction task, and segmentation are reported in Tables 1 and 2, respectively. It is worth noting that the radiomics results of the segmentation-only model and the segmentation results of the radiomics-only model are only reported as sanity checks as they both achieve (expected) random performance on these tasks. There is no precedent work on this dataset nor task.

Regarding the comparison with state of the art, standard radiomics methods have been applied to other H&N datasets and other survival outcomes in e.g. [29], yet direct comparison is not possible. We therefore compare against the standard radiomics method described in Section 2.4.

Table 1. Performance comparison for the radiomics task. w_{rad} and w_{seg} are the radiomics and segmentation loss weights, respectively. We report the C-index for each fold of the CV as well as the average ± standard-error of the C-index.

model (w_{rad}, w_{seg})	fold-1	fold-2	fold-3	fold-4	fold-5	**Mean**
Deep radiomics (1,0)	0.703	0.599	0.578	0.687	0.684	0.650 ± 0.026
Deep multi-task (1,1)	0.713	0.702	0.803	0.615	0.783	**0.723** ± 0.033
Deep segmentation (0,1)	0.413	0.377	0.473	0.348	0.570	0.416 ± 0.024
Standard radiomics	0.827	0.710	0.627	0.687	0.624	0.695 ± 0.0826

Table 2. Performance comparison for the segmentation task. w_{rad} and w_{seg} are the radiomics and segmentation loss weights, respectively. We report the average Dice score for each fold of the CV as well as the global average ± standard-error for the Dice.

model (w_{rad}, w_{seg})	**fold-1**	**fold-2**	**fold-3**	**fold-4**	**fold-5**	**Mean**
Deep radiomics (1,0)	0.006	0.001	0.000	0.034	0.000	0.008 ± 0.007
Deep multi-task (1,1)	0.700	0.589	0.361	0.595	0.677	0.584 ± 0.060
Deep segmentation (0,1)	0.713	0.681	0.696	0.639	0.685	**0.683** ± 0.012

4 Discussion and Conclusions

The network design proposed in this work relies on the assumption that teaching to segment the tumoral volume will benefit the prediction of patient prognosis. The observed results seem to validate this hypothesis, as the most concordant DFS prediction is achieved when combined with the segmentation task (see Table 1) culminating to an average C-index of 0.723. The combination of Dice and survival losses allowed to spatially guide neuron activations with the former, and learn localized prognostically relevant patterns with the latter. As a reminder,

only a large bounding box encompassing the spatially extended oropharyngeal region (automatically determined as of [28]) is provided to the network. Therefore, this approach does not require tumoral contours at test time, as opposed to the standard radiomics approach. Remarkably, this fully automatic approach outperformed the standard radiomics pipeline relying on manual contouring (C-index of 0.695). This result opens avenues for very large scale clinical studies to validate the prognostic models on patients for which we only have the outcomes but no manual annotations of the tumors.

We believe that this work contributes to the state of the art by proposing a fully automatic bi-modal 3D deep prognostic model able to deal with survival data and to learn from relatively small datasets. It can do so by optimally leveraging training data via the use of the highly related and observation-rich segmentation task.

A surprisingly high prognostic performance is achieved even without using the Dice loss (only deep radiomics), which is highlighted by a C-index of 0.650, 7.3% lower than the top result. This demonstrates the efficacy of the survival loss combined with an appropriate encoding architecture. Note that in this fully-radiomics method, the GTVt contours are provided neither during training nor testing.

The best-performing segmentation method achieved a Dice score of 0.683. For this GTVt segmentation, the segmentation task did not benefit from the radiomics task, where the Dice loss provided best results when used on its own (see Table 2). This observation may be due to an optimization issue, the PFS task adding noise to the gradients. Other loss weights could be used to favor the segmentation task, yet the main task of interest in this work is the outcome prediction and the segmentation is only used to boost the performance on the latter. Note that the segmentation performance is far from the winner of the HECKTOR 2020 challenge (average DSC of 0.759), in which the main and only task was tumor segmentation, and the model was based on a complex ensemble of UNets.

A limitation of the proposed work is the use of binary weights (w_{rad}, w_{seg}) given to the two losses in the multi-task model (i.e. unweighted sum of losses). In future work, we will explore other types of loss weighting such as geometric [30] and epistemic uncertainty losses [31]. As another interesting future work, one could also consider adding an auxiliary task of domain adversariability to the training with a branch similar to the radiomics one and with a gradient reversal to create domain invariant features and ensure good generalization to new scanners and image acquisition protocols [32]. We also plan to study activation maps to reveal the most prognostically relevant regions and patterns used by the deep radiomics model.

Acknowledgements

This work was partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069) and the Swiss Personalized Health Network (SPHN via the IMAGINE and QA4IQI projects).

References

1. Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.
2. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
3. Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.
4. Vincent Andrearczyk, Valentin Oreiller, Mario Jreige, Martin Vallières, Joel Castelli, Hesham Elhalawani, Sarah Boughdad, John O. Prior, and Adrien Depeursinge. Overview of the HECKTOR challenge at MICCAI 2020: Automatic head and neck tumor segmentation in PET/CT. In *Lecture Notes in Computer Science (LNCS) Challenges*, 2021.
5. David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995.
6. Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114. PMLR, 2016.
7. Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.
8. Jon Arni Steingrimsson and Samantha Morrison. Deep learning for survival outcomes. *Statistics In Medicine*, 39(17):2339–2349, 2020.
9. Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II*, volume 11384. Springer, 2019.
10. Stephen Baek, Yusen He, Bryan G. Allen, John M. Buatti, Brian J. Smith, Ling Tong, Zhiyu Sun, Jia Wu, Maximilian Diehn, Billy W. Loo, et al. Deep segmentation networks predict survival of non-small cell lung cancer. *Scientific Reports*, 9(1):1–10, 2019.
11. Vishwa S. Parekh and Michael A. Jacobs. Deep learning and radiomics in precision medicine. *Expert Review of Precision Medicine and Drug Development*, 4(2):59–72, 2019.
12. André Diamant, Avishek Chatterjee, Martin Vallières, George Shenouda, and Jan Seuntjens. Deep learning in head & neck cancer outcome prediction. *Scientific Reports*, 9(1):1–10, 2019.
13. Yucheng Zhang, Edrise M. Lobo-Mueller, Paul Karanicolas, Steven Gallinger, Masoom A. Haider, and Farzad Khalvati. CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging. *BMC Medical Imaging*, 20(1):1–8, 2020.
14. Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat, and Lee A.D. Cooper.

Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.
15. Hongming Li, Pamela Boimel, James Janopaul-Naylor, Haoyu Zhong, Ying Xiao, Edgar Ben-Josef, and Yong Fan. Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 846–849. IEEE, 2019.
16. Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
17. Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.
18. Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3):034002, 2019.
19. Leon Weninger, Qianyu Liu, and Dorit Merhof. Multi-task learning for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 327–337. Springer, 2019.
20. multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation.
21. Mara Graziani, Sebastian Otálora, Henning Muller, and Vincent Andrearczyk. Guiding cnns towards relevant concepts by multi-task and adversarial learning. *arXiv preprint arXiv:2008.01478*, 2020.
22. Andrei Iantsen, Dimitris Visvikis, and Mathieu Hatt. Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. In *Lecture Notes in Computer Science (LNCS) Challenges*, 2021.
23. Joost J.M. Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.
24. Yannick Suter, Urspeter Knecht, Mariana Alão, Waldo Valenzuela, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. Radiomics for glioblastoma survival analysis in pre-operative mri: exploring feature robustness, class boundaries, and machine learning techniques. *Cancer Imaging*, 20(1):1–13, 2020.
25. Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017.
26. Cox R. David et al. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.
27. Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics multivariable prognostic models. *Statistics in Medicine*, 15:361–387, 1996.
28. Vincent Andrearczyk, Valentin Oreiller, and Adrien Depeursinge. Oropharynx detection in PET-CT for tumor segmentation. In *Irish Machine Vision and Image Processing*, 2020.
29. Martin Vallieres, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo JWL Aerts, Nader Khaouam, Phuc Felix Nguyen-Tan, Chang-Shu Wang, Khalil Sultanem, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific reports*, 7(1):1–14, 2017.

30. Sumanth Chennupati, Ganesh Sistu, Senthil Yogamani, and Samir A. Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
31. Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.
32. Vincent Andrearczyk, Adrien Depeursinge, and Henning Müller. Neural network training for cross-protocol radiomic feature standardization in computed tomography. *Journal of Medical Imaging*, 6(3):024008, 2019.