Head and Neck Tumor and Lymph Node Segmentation and Outcome Prediction from 18F-FDG PET/CT Images: Simplicity is All You Need

Preprint · September 2022

DOI: 10.13140/RG.2.2.30709.04328

5 authors, including:

Thibault Escobar
Université Paris-Saclay
6 PUBLICATIONS **1** CITATION
SEE PROFILE

Fahad Khalid
Institut Curie
4 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Kibrom Berihu Girum
Institut Curie
21 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Irene Buvat
Institut Curie - Inserm
385 PUBLICATIONS **13,475** CITATIONS
SEE PROFILE

Head and Neck Tumor and Lymph Node Segmentation and Outcome Prediction from 18F-FDG PET/CT Images: Simplicity is All You Need

Louis Rebaud[1,2] *, Thibault Escobar[1,3] *,
Fahad Khalid[1], Kibrom Girum[1],
and Irène Buvat[1]

[1] Laboratory of Translational Imaging in Oncology, U1288, Institut Curie, Inserm, Université Paris-Saclay, Orsay, France
[2] Siemens Healthcare SAS, Saint Denis, France
[3] DOSIsoft SA, Cachan, France

louis.rebaud@gmail.com, thibescobar@gmail.com

*co-first authors

Abstract. Automated lesion detection and segmentation might assist radiation therapy planning and contribute to the identification of prognostic image-based biomarkers towards personalized medicine. In this paper, we propose a pipeline to segment the primary and metastatic lymph nodes from fluorodeoxyglucose (FDG) positron emission tomography and computed tomography (PET/CT) head and neck (H&N) images and then predict recurrence free survival (RFS) based on the segmentation results. For segmentation, an out-of-the-box nnUNet-based deep learning method was trained and labelled the two lesion types as primary gross tumor volume (GTVp) and metastatic nodes (GTVn). For RFS prediction, 2421 radiomic features were extracted from the merged GTVp and GTVn using the pyradiomics package. The ability of each feature to predict RFS was measured using the C-index. Only the features with a C-index greater than C_{min}, hyperparameter of the model, were selected and assigned a +1 or -1 weight as a function of how they varied with the recurrence time. The final RFS probability was calculated as the mean across all selected feature z-scores weighted by their +/-1 weight. The fully automated pipeline was applied to the data provided through the HECKTOR 2022 MICCAI challenge. On the test data, the fully automated segmentation model achieved 0.777 and 0.763 Dice scores on the primary tumor and lymph nodes respectively (0.770 on average). The binary-weighted radiomic model yielded a 0.682 C-index. These results allowed us to rank first for outcome prediction and fourth for segmentation in the challenge. We conclude that the proposed fully-automated pipeline from segmentation to outcome prediction using a binary-weighted radiomic model competes well with more complicated models. Team: LITO.

Keywords: Medical imaging · Survival prediction · Segmentation · FDG PET/CT · Head and neck · Machine learning

1 Introduction

Quantitative medical image analysis assists in patient staging, treatment planning and monitoring, and overall patient management. In head and neck (H&N) cancer, fluorodeoxyglucose (FDG) positron emission tomography combined with computed tomography (PET/CT) is a modality of choice for initial staging and patient follow-up and contributes to radiation therapy planning. Indeed, H&N cancer primary treatment mostly relies on radiotherapy and requires target volume delineation of the gross primary tumor volume (GTVp) and cancer node volumes (GTVn) on PET/CT images, which is time-consuming and prone to intra/inter-observer variabilities. Automated segmentation might allow radiation oncologists to optimize the treatment plan in a shorter time while improving reproducibility. In addition, the prediction of the risk of relapse based on medical images could help identify patients for whom treatment intensification and close monitoring might be needed.

In the recent years, machine learning (ML) and radiomics have been instrumental in advancing automated image segmentation and building predictive models. Yet, the diversity of datasets on which methods are designed and tested makes it difficult to compare their performance and determine which one is best suited in a particular context. Given the possible sensitivity of automated segmentation and predictive models to image quality, multi-center evaluation of these methods is absolutely needed before considering clinical deployment.

Challenges offer unique opportunities for testing and comparing the performance of different methods on a common database using large multi-center datasets. The HEad and neCK TumOR (HECKTOR) challenges organized as part of MICCAI aims at establishing best-performing methods for segmentation and prediction tasks [1,2]. In 2022, the HECKTOR challenge first task was to automatically segment the H&N GTVp and GTVn from FDG PET/CT images. The second task consisted in automatically predicting patient outcomes from a PET/CT image, with or without clinical information, with PET/CT images and clinical information collected from nine different centers.

Several contributions to the automated segmentation in the context of H&N cancer have already been published. Guo et al. proposed a modified U-net approach using dense blocks and reached 0.71 average Dice score on a public multi-center dataset of 250 PET/CT H&N patients [3]. Their study also showed that combining PET and CT in two channels substantially increased the segmentation performance compared to using PET (0.64 average Dice score) or CT (0.31 average Dice score) alone. Ren et al. compared several modality combinations including PET, CT, and magnetic resonance imaging (MRI) on a multi-center dataset of 153 patients for deep learning tumor segmentation using a U-net approach [4]. All combinations including PET provided similar results (0.72 to 0.74 Dice score), while the anatomic-only combination (CT and MRI) led to a lower score (0.58). More generally, automated medical image segmentation is currently dominated by deep convolutional neural networks (CNN) [5,6,7]. Most methods rely on U-net based approaches with several context-specific changes in model architecture, training scheme, and data pre- or post-processing. In HECKTOR 2021 challenge, the best-performing segmentation method used a tuned nnUNet with squeeze and excitation (SE) layers on fused PET and CT images, yielding a 0.779 Dice score on primary tumor [7,8].

Similarly, models have been proposed to predict patient outcome from PET/CT images in H&N cancer (eg, [9,10]). In HECKTOR 2021, two different methods performed

best at predicting the progression free survival [11,12]. Both were based on a CNN trained on unsegmented images using large bounding boxes, and achieved 0.720 and 0.694 C-index on the test data respectively. A logistic model based on radiomic features calculated from the segmented tumor region also performed well with a 0.683 C-index [13].

This paper presents our simple and efficient pipeline for fully automatic segmentation and outcome prediction method and its performance on the HECKTOR 2022 challenge data. For the segmentation task, we adapted the publicly available nnUNet deep learning framework to detect and segment the H&N primary tumor (GTVp) and nodal gross tumor volumes (GTVn) [7]. For the prediction task, we introduce a novel binary-weighted model operating on radiomic features calculated from the tumor regions automatically segmented in the previous step. The evaluation was conducted on the HECKTOR 2022 challenge data and the models are publicly available.

2 Materials and methods

Here, we describe our proposed fully-automatic end-to-end framework to segment lesions and predict outcome from 18F-FDG PET/CT images (Fig.1). First, a well established out-of-the-box nnUNet deep learning method was trained to segment and label the GTVp and GTVn [7]. From the segmented GTVp and GTVn regions, we extracted radiomic features. We then applied the binary-weighted model to rank the patients as a function of their recurrence free survival.

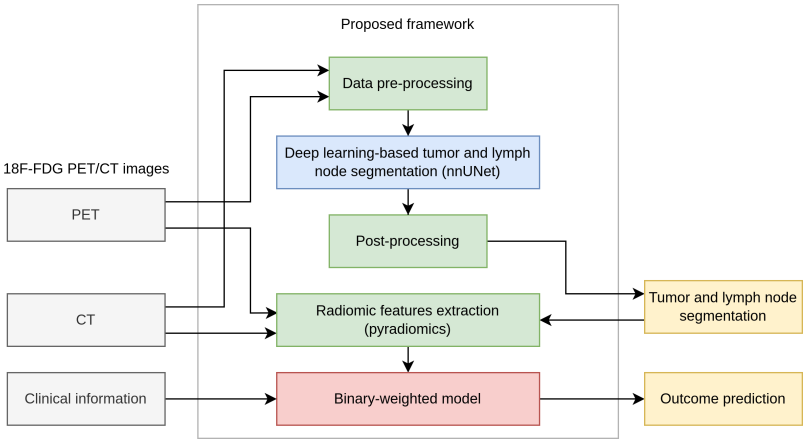

Fig. 1. Proposed framework: schematic representation of the fully-automatic pipeline from segmentation to outcome prediction.

2.1 Data

To develop and evaluate the proposed method, we used the HECKTOR 2022 data that included FDG PET/CT images, clinical and survival data of 524 patients from 7 centers for training and PET/CT and clinical data only of 359 patients from 3 centers for blind testing of the models [1,2]. In the training data, reference segmentations of the primary

tumor (GTVp) and metastatic nodes (GTVn) were provided. Train and test PET/CT scans were provided with 9 clinical features with some missing values: gender, age, weight (1.23% missing values), tobacco (0 = *no*, 1 = *yes*) (61.1% missing), alcohol (0 = *no*, 1 = *yes*) (68.5% missing), performance status (56.0% missing), human papillomavirus (HPV) status (0 = *no*, 1 = *yes*) (35.2% missing), surgery (0 = *no*, 1 = *yes*) (38.7% missing), and chemotherapy (0 = *no*, 1 = *yes*). RFS was provided for 488 patients in the train set, and 339 patients of the test set for whom the outcome was known were concerned by the outcome prediction (task 2).

Data pre-processing: The training CT images had an original median voxel-size of $0.976 \times 0.976 \times 2.798\ mm^3$ and the PET images had median voxel-size of $4.000 \times 4.000 \times 3.270\ mm^3$. All PET/CT images and corresponding segmentations were resampled to $2.0 \times 2.0 \times 2.0\ mm^3$. CT and PET images were resampled using a third-order spline. The segmentation mask was resampled using nearest neighbor interpolation.

2.2 Tumor and lymph node segmentation

Deep learning model: All CT images were clipped between 0.5^{th} and 99.5^{th} percentile of the Hounsfield Units (HU) intensity values and normalized using z-score based on all training images. PET Standardized uptake values (SUV) were normalized using z-score patient-wise on the whole image. We used a nnUNet in "3D full resolution" mode to detect and segment the tumor and lymph nodes [7]. The pre-processed PET/CT images were given to the model as two-channel input images (PET and CT). Each PET/CT image was decomposed in random patches of $160 \times 160 \times 96 \times 2$ voxels before input into model training. The architecture of the 3D model was not modified except for the output channel. The output was a $1 \times 1 \times 1$ convolution of size $160 \times 160 \times 96 \times 2$, where 2 corresponds to the tumor and lymph nodes channels. A softmax non-linear activation was used at the output layer of the 3D nnUNet model.

Training scheme: The train set consisting of 524 patients was randomly divided into training and validation subsets using a five-fold cross-validation technique. Each fold contained data from 104 or 105 validation patients and 420 or 419 training patients. The nnUNet model was trained using the sum of Dice and cross-entropy losses. The initial number of feature maps in the architecture was 32. Performance assessment and post-processing strategy were determined based on the five-fold cross-validation with 1000 epochs training, with an initial learning rate of 0.01 and a scheduler weight decay of $3e^{-5}$. We selected a batch size of two. Other hyper-parameter settings, including data augmentation techniques, were the default settings of nnUNet. Implementation was done in Pytorch and training was performed using four GPUs: three NVIDIA Quadro RTX 5000 with 16GB and one NVIDIA RTX A6000 with 49GB GPU memory. On average, the training time was 141s per epoch on NVIDIA Quadro RTX 5000 and 82s on NVIDIA RTX A6000.

Post-processing: The segmentation output of the deep learning model had a $2 \times 2 \times 2$ mm^3 voxel spacing. It was then resampled into the corresponding original CT spacing. Then, a median filter with a $3 \times 3 \times 3$ voxel kernel size was applied to smooth out the staircase effect.

Prediction on the test set: For predictions on the test set, three strategies were used. First we ensembled the five models trained during cross-validation. Second, a bagging strategy was adopted increasing the number of ensembled models to nine. Finally, we increased the number of epochs to 1500 and trained only one model on the whole dataset.

2.3 Outcome prediction

Our prediction model was based on engineered radiomic features extracted from the tumor regions segmented using the automated approach described in Section 2.2. These features were then analyzed using an original approach yielding what we call a binary-weighted model.

Radiomic features extraction: We used the segmentation mask produced by the deep learning model described in section 2.2. Primary tumor and lymph node regions were merged as a single "lesion" mask. To make the model less sensitive to potential segmentation errors, multiple masks were created from this binary lesion mask:

- Original lesion mask
- Smallest bounding box enclosing all the lesions
- Lesion mask refined by removing all voxels in which SUV was less than 2.5
- Lesion mask refined by removing all voxels in which SUV was less than 4
- Lesion mask re-segmented with a threshold of 40% of global SUVmax
- Lesion mask dilated by 1mm (resp 2, 4, 8 and 16 mm)
- A 2mm (resp 4, 8 mm) thick shell surrounding each connected component of the lesion mask

For each of these 13 masks, 93 radiomic features were computed on the PET image and 93 on the CT image with pyradiomics [14]. Three additional handcrafted features (tumors and lymph nodes number, whole-body or H&N scanassessment). Used together with the provided nine clinical ones, this pipeline produced 2430 features.

Binary-weighted model: From the literature and our experience, we hypothesize that it is difficult to accurately estimate biomarker importance in outcome prediction. Indeed, noise in the data, censoring of the target, e.g. progression free survival, and relatively low number of training samples might increase the risk of biased estimation of the feature weights. To mitigate this effect, we propose to reduce the learned information to the bare minimum and only estimate a sign to be assigned to each feature for estimating the target. This is the core mechanism of the introduced binary-weighted model.

Definition: Our training dataset includes N samples and M features. Many radiomic features are highly correlated. To comply with the basic assumption of our binary-weighted model, only one among a set of correlated features should be kept because if they are all input to the model, this will artificially give a large weight to the information reflected by the feature. We thus perform feature selection by calculating the absolute value of the Pearson correlation coefficient for all pairs of features. A threshold

ρ is used to set the value above which two features are deemed too correlated. In such case, one of the two features is randomly selected and dropped.

Let's C_{index} be the Harrell's concordance index [15]. Each feature x_i is evaluated on its ability to correctly predict the target value y with:

$$c_i = C_{index}(x_i, y) \tag{1}$$

To reduce the risk of wrong estimation of the sign, the features with $|c_i| < C_{min}$ are dropped, where $|c_i| = max\{1 - c_i, c_i\}$ and C_{min} is a hyperparameter in $[0.5, 1]$. The remaining features are assigned a sign as follows:

$$s_i = \begin{cases} +1, & \text{if } c_i \geq 0.5 \\ -1, & otherwise \end{cases} \tag{2}$$

A normalization step is necessary to scale the feature values to the same range. Otherwise, features with large absolute values would have a higher weight in the final prediction. To do so, the model computes the z-score of each feature:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3}$$

where μ_i and σ_i are the mean and standard deviation of x_i in the train set. The estimate $\hat{y}$ of the target y is computed with:

$$\hat{y} = \frac{1}{M} \sum_{i}^{M} s_i \times z_i \tag{4}$$

The computation of $\hat{y}$, μ_i and σ_i are done by ignoring the missing values of the dataset. This allows the model to use features with missing values.

Here, C_{min} and ρ are the only two hyperparameters of the model.

Curse of dimensionality: The curse of dimensionality is a phenomenon where we observe a loss in performance of ML models when too many features are given as an input. This especially occurs in medical datasets when the data are high-dimensional and the number of samples is low [16]. We hypothesize that the binary-weighted model is resilient to this phenomenon. We tested this hypothesis on the train set of the HECKTOR dataset by gradually increasing the number of features input to the model.

Ensembling: To produce a more precise and stable estimate $\hat{y}$, a bagging strategy was adopted. An ensemble of E binary-weighted models were trained, each model being trained on a random sample of size N of the training data drawn with replacement (ie a bootstrap sample). Each model also randomly selected F features to work with. The models were trained on their bootstrap sample from the train set and predicted $\hat{y}$ on the test set. The E predictions from the E models were then aggregated with a median. F is a hyperparameter of the ensemble model. Our experiments on the train set suggested that the higher E, the better the performance. We used $E = 10^5$ on the test set, a number large enough too ensure good results while keeping computational cost reasonable.

Cross-validation: To evaluate a model from the train set, we used a two-hundred-fold Monte Carlo cross-validation with a validation set of size $0.5 \times N$ (CV). The model prediction on the validation set was evaluated with Harrell's C-index. The average score and its confidence interval were reported.

Hyperparameters optimization: The ensemble model has 3 hyperparameters: F, C_{min} and ρ. To determine the best hyperparameter set, random search was used. 1000 hyperparameter sets were randomly drawn and evaluated using CV. The hyperparameter sets were then ranked by their CV scores. To reduce the risk of overfitting the hyperparameter choice on the train set, the B best hyperparameter sets were selected, and for the prediction on the test set, an ensemble model was trained with each binary-weighted model randomly selecting a hyperparameter set from the selected B. The B value was optimized with an additional CV. Three bagged models were evaluated in the train and test sets of the HECKTOR challenge. While similar, each model used more and more hyperparameter sets in its random search, each time increasing the probability of overfitting on the train set. The number of hyperparameter sets tested was increased gradually through the 3 attempts given to the participating teams.

Feature importance: While the binary-weighted model only gives weights of -1 or $+1$, after bagging, an approximation of feature importance can be computed by taking the average sign of each feature across all models. Feature importance was determined on the train set of HECKTOR.

3 Results

3.1 Segmentation evaluation

Cross-validation: The Dice score across all images through the cross-validation was 0.850 for GTVp and 0.789 for GTVn (0.821 on average). For thorough comparison, Table 1 reports the Dice score across the different centers of acquisition.

Table 1. Dice scores for primary tumor and lymph node segmentation across the different centers evaluated on a five-fold cross-validation on the train set.

Center	**Patient #**	**GTVp Dice**	**GTVn Dice**	**average Dice**
CHUP	72	0.868	0.687	0.778
CHUV	53	0.823	0.781	0.803
MDA	198	0.821	0.813	0.817
HMR	18	0.846	0.811	0.829
CHUS	72	0.865	0.805	0.835
CHUM	56	0.849	0.831	0.840
HGJ	55	0.883	0.829	0.856
All	**524**	**0.850**	**0.789**	**0.821**

Test: Table 2 displays the class-specific Dice scores for our three submitted models for evaluation on the test set. The model trained on all training data for 1500 epochs achieved the highest scores.

Table 2. Dice scores from our 3 methods on the test set of HECKTOR.

Method	**GTVp Dice**	**GTVn Dice**	**average Dice**
Ensembled 5 folds	0.778	0.761	0.769
Bagging 9 samples	0.779	0.759	0.769
Whole train set	**0.777**	**0.763**	**0.770**

3.2 Qualitative assessment

PET/CT images, ground truth and predicted segmentations are shown in Fig.2 for 5 patients. The examples were selected based on the Dice scores. The top two rows display high Dice scoring patients (average Dice 0.922 and 0.910 respectively), the third row a patient with an average score (0.761), while the fourth (0.303) and fifth (0.000) rows display patients with the lowest scores.

Results for patients (1) and (2) were very satisfactory. In patient (3), the model accurately identified the two nodes and the tumor but missed some voxels, especially at the sharp edges. In patient (4), false positive node voxels were labeled by the model (not shown in the figure because not in the slice). Last, patient (5) shows an example of accurate detection and segmentation but with complete class mismatch. The green contour representing the tumor is precisely delineated by the model but labelled as a node, as shown by the pink predicted contour, yielding a Dice equal to zero.

3.3 Performance of the outcome prediction model

Table 3 shows the results of the different models tested during the challenge. A binary-weighted model without bagging was evaluated only on the train set and not submitted because its performances were below the bagged models on the train set. The performance of the three submitted bagged models is correlated with the number of hyperparameter sets evaluated on the train set. The best model was the one which had the most extensive search of hyperparameters.

Table 3. C-index and number of hyperparameters searched for the prediction models evaluated on the train and test set of the HECKTOR challenge. On the train set, the mean C-index over the CV is reported as well as the confidence interval (CI).

Model	**CV C-index train set (CI)**	**C-index test set**	**Nb tested sets of hyperparameters**
Binary-weighted	0.645 (0.585 - 0.707)		10
Binary-weighted bagged	0.668 (0.605 - 0.730)	0.670	10
Binary-weighted bagged	0.675 (0.613 - 0.731)	0.673	100
Binary-weighted bagged	**0.688 (0.642 - 0.732)**	**0.682**	1000

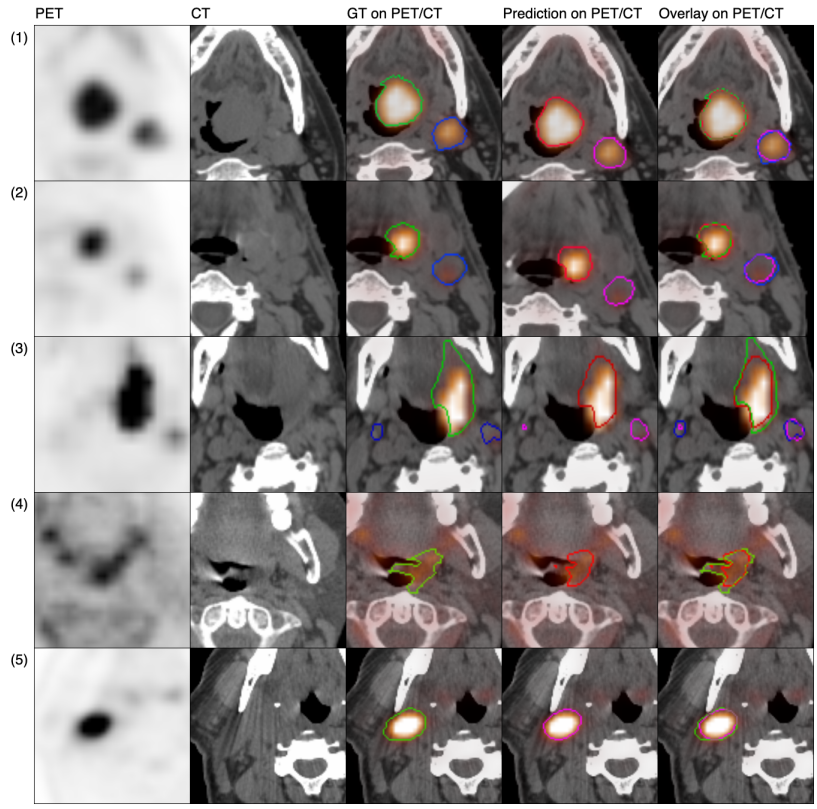

Fig. 2. Examples of PET/CT images, ground truth and predicted segmentation for five patients from the validation sets of the five-fold cross-validation. Green and blue ground truth contours correspond to tumor and lymph node respectively. Red and pink contours correspond to the predicted segmentation for tumor and lymph node.

3.4 Resilience to the curse of dimensionality

Fig. 3 shows the result of the experiment using the train set to test our hypothesis stating that binary-weighted models do not suffer from the curse of dimensionality. The performance plateaued when increasing the number of features used by the model up to the maximum number of available features.

3.5 Feature importance

The importance of the clinical and some representative radiomic features evaluated on the train set is presented in Fig. 4. The error bars are not shown because by construction of the model, they are unnecessary (the higher the absolute value, the lower the standard deviation).

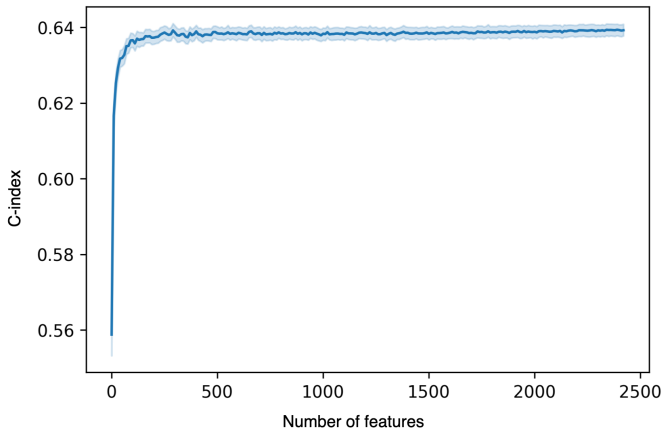

Fig. 3. Cross-validated C-index of a binary-weighted model (not bagged) when increasing the number of features.

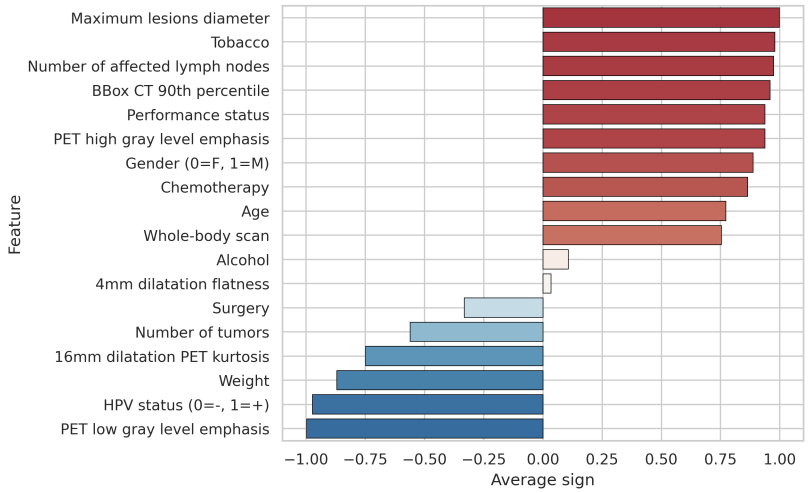

Fig. 4. Importance of the clinical and representative radiomic features. A positive value (red) shows a positive correlation with the risk and a negative value (blue) is a negative correlation. The higher the absolute value of the average sign, the more important the feature. "Whole-body scan" is 1 if the scan is whole-body or 0 if only H&N.

4 Discussion

4.1 Segmentation

Our segmentation method was inspired by Xie and Peng [8] using Isensee et al. [7] framework. Our choice of not using the SE layers and keep PET and CT separated as two channels was based on the intuition that approaching the problem in a straightforward way would increase its robustness. Overall, our segmentation results were satisfactory, ranking fourth in the challenge with 0.770 average Dice, compared to the 0.788 Dice achieved by the winner.

Although the centers had different numbers of patients, Dice scores were consistently lower for lymph nodes than for primary tumor in all centers, demonstrating they are more difficult to segment. Mislabelling of node regions as seen in Fig. 2 decreased Dice value although contours were accurately delineated. One way to address this mislabelling could be to set higher weight to the lymph node class in the loss function.

According to our test results, the deployment strategy did not have a big impact on performance. Indeed, ensembling the cross-validation models, using a bagging strategy while increasing the number of models, or training only one model on the whole dataset, led to very similar performance.

Based on the qualitative visual assessment, our model tends to perform better on smooth connected components. Complex structures and sharp contours are more prone to errors. Processing and training methods adapted to higher resolution input images might have reduced these errors.

4.2 Binary-weighted model

Our results suggest that the binary-weighted model is a competitive and robust method. This implies that it might indeed be challenging to accurately estimate feature weights. The more degrees of freedom in a model, the higher the risk of overfitting. In problems with weak and noisy targets and low number of training samples, reducing the training to the bare minimum could be of utmost interest. For the HECKTOR challenge, it probably helped mitigate the overfitting.

Fig. 3 shows that the binary-weighted model does not suffer from the curse of dimensionality. The vast majority of ML algorithms need some feature selection to avoid a drop in performance due to too many features. We hypothesized that in our binary-weighted model, the features would work together to cancel their noise and biases, analogous to the wisdom of the crowd phenomenon where errors of individuals cancel each other out. Adding more features does not result in loss in performance as in other traditional ML methods.

Features importance shed light on the model interpretation (Fig. 4). For instance, a high performance status is associated with worse prognosis. Tobacco is also associated with a higher risk in our model. Large tumor diameter and high SUV values in the lesions are associated with increased risk. Other features, such as chemotherapy, can be interpreted as indirect measure of the patient condition. Interestingly, the number of affected lymph nodes appears to be a strong prognostic factor.

5 Conclusions

We proposed a new, fully automated framework to predict outcomes in H&N patients from a given PET/CT image and clinical information. It involves deep learning-based GTVp and GTVn segmentation, radiomic feature extraction, and outcome prediction. Our pipeline including the novel binary-weighted radiomic model outperformed other methods for outcome prediction while providing accurate segmentation, ranking first for prediction and fourth for segmentation in the HECKTOR 2022 challenge. The number of lymph nodes was one of the prognostic features, highlighting the importance of lymph node segmentation for predicting the outcome in H&N cancer.

We created an easy-to-use package for the binary-weighted model, called Individual Coefficient Approximation for Risk Estimation (ICARE). The code is publicly available at: github.com/Lrebaud/ICARE.

References

1. Oreiller, V., et al.: Head and Neck Tumor Segmentation in PET/CT: The HECKTOR Challenge. Medical Image Analysis. 77:102336 (2022).
2. Andrearczyk, V., et al.: Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT. in: Head and Neck Tumor Segmentation and Outcome Prediction (2023).
3. Guo, Z., et al.: Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol. 64.20:205015 (2019)
4. Ren, J., et al.: Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol. 60.11:1399-1406 (2021).
5. Menze, B. H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 34.10:1993-2024 (2014).
6. Antonelli, M., et al.: The medical segmentation decathlon. Nat Commun. 13.1:1-13 (2022).
7. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 18.2:203-211 (2021).
8. Xie, J. and Peng, Y.: The head and neck tumor segmentation based on 3D U-Net. In: LNCS Challenges (2022).
9. Vallières, M., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 7.1:1-14 (2017).
10. Diamant, A., et al.: Deep learning in head & neck cancer outcome prediction. Sci Rep. 9.1:1-10 (2019).
11. Saeed, N., et al.: An ensemble approach for patient prognosis of head and neck tumor using multimodal data. In: LNCS Challenges (2022).
12. Naser, M.A., et al.: Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET-CT imaging data. In: LNCS Challenges (2022).
13. Salmanpour, M.R., et al.: Advanced automatic segmentation of tumors and survival prediction in head and neck cancer. In: LNCS Challenges (2022).
14. Van Griethuysen, et al.: Computational radiomics system to decode the radiographic phenotype. Cancer research. 77.21:e104-e107 (2017).
15. Harrell Jr, F.E., et al.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 15.4:361-387 (1996).
16. Berisha, V., et al.: Digital medicine and the curse of dimensionality. NPJ Digit Med. 4.1:1-8 (2021).