Multi-Level Multi-Modality Fusion Radiomics: Application to PET and CT Imaging for Prognostication of Head and Neck Cancer

Wenbing Lv, Saeed Ashrafinia, Jianhua Ma, *Member, IEEE*, Lijun Lu, and Arman Rahmim, *Senior Member, IEEE*

***Abstract*—To characterize intra-tumor heterogeneity comprehensively, we propose a multi-level fusion strategy to combine PET and CT information at the image-, matrix- and feature-levels towards improved prognosis. Specifically, we developed fusion radiomics in the context of 3 prognostic outcomes in a multi-center setting (4 centers) involving 296 head & neck cancer patients. Eight clinical parameters were first utilized to build a (1) clinical model. We also built models by extracting 127 radiomics features from (2) PET images alone; (3–8) PET and CT images fused via wavelet-based fusion (WF) using CT-weights of 0.2, 0.4, 0.6 and 0.8, gradient transfer fusion (GTF), and guided filtering-based fusion (GFF); (9) fused matrices (sumMat); (10–11) fused features constructed via feature averaging (avgFea) and feature concatenation (conFea); and finally, (12) CT images alone; above models were also expanded to include both clinical and radiomics features. Seven variations of training and testing partitions were investigated. Highest performance in 5, 6 and 5 partitions was achieved by image-level fusion strategies for RFS, MFS and OS prediction, respectively. Among all partitions, WF0.6 and WF0.8 showed significantly higher performance than CT model for RFS (C-index: 0.60 ± 0.04 vs. 0.56 ± 0.03, p-value: 0.015) and MFS (C-index: 0.71 ± 0.13 vs. 0.62 ± 0.08, p-value: 0.020) predictions, respectively. In partition CER 23 vs. 14, WF0.6 significantly outperformed Clinical model for RFS prediction (C-index: 0.67 vs. 0.53, p-value: 0.003); both avgFea and WF0.6 showed C-index of 0.64 and significantly higher than that of PET only (C-index: 0.51, p-value: 0.018 and 0.031, respectively) for OS prediction. Fusion radiomics modeling showed varying improvements compared to single modality models for different outcome predictions in different partitions, highlighting the importance of generalizing radiomics models. Image-level fusion holds potential to capture more useful characteristics.**

***Index Terms*—Image fusion, radiomics, PET/CT, multicenter, Head & Neck Cancer.**

I. Introduction

THERE are approximately 600,000 new cases of head and neck (H&N) cancer every year worldwide, with $40\%-50\%$ of these resulting in death [1]. Aside from conventional prognostic factors such as PET standard uptake values (SUV) [2], tumor site, stage and HPV status [3], intra-tumor heterogeneity is also increasingly recognized to be related to tumor development, response to therapy and metastasis in H&N cancer [4]–[6]. Accurate quantification of heterogeneity has great potential to identify high-risk patients who may benefit from aggressive treatment, and low-risk patients who may be free from toxic side-effects [7].

Radiomics analysis, incorporating measurement of heterogeneity, has been extensively investigated towards improved outcome prediction [8], though has been commonly conducted only on individual modalities (CT, MRI or PET) [5], [9]–[14]. As different modalities convey different aspects of disease [15], texture and histogram features extracted from CT image quantifying the spatial distribution of tissue intensities [16]; ring-shape ^{18}F-fluorodeoxyglucose (FDG) uptake reflecting intratumoral necrosis had prognostic value for patients with H&N squamous cell carcinoma (HNSCC) [17]; SUV of FDG-PET correlates with microvessel density characterized by perfusion CT in H&N cancer [18]. Use of combined PET and CT scans for visual interpretation has shown superiority for initial diagnosis and subsequence surveillance compared with use of PET or CT alone [19], [20].

Manuscript received May 13, 2019; revised August 3, 2019 and November 2, 2019; accepted November 23, 2019. Date of publication December 4, 2019; date of current version August 5, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 81871437 and Grant U1708261, in part by the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme Lijun Lu, 2018, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515011104, and in part by the China Scholarship Council under Grant 201808440464, as well as the BC Cancer Foundation, Vancouver, Canada. *(Corresponding authors: Jianhua Ma and Lijun Lu.)*

J. Ma and L. Lu are with the School of Biomedical Engineering and Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China (e-mail: jhma@smu.edu.cn; ljlubme@gmail.com).

W. Lv is with the School of Biomedical Engineering and Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China, and also with the Department of Integrative Oncology, BC Cancer Research Centre, Vancouver, BC V5Z 1L3, Canada (e-mail: bing179@smu.edu.cn).

S. Ashrafinia is with the Department of Electrical & Computer Engineering and Radiology, Johns Hopkins University, Baltimore, MD 21287 USA (e-mail: s.ashrafinia@gmail.com).

A. Rahmim is with the Department of Integrative Oncology, BC Cancer Research Centre, Vancouver, BC V5Z 1L3, Canada, and also with the Department of Radiology and Physics, University of British Columbia, Vancouver, BC V6T 1Z1, Canada (e-mail: arahmim1@gmail.com).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2019.2956354

Instead of visual evaluation, for the relatively few quantitative multi-modality radiomics studies performed towards more accurate tumor imaging phenotype exploration, the adopted strategy has most commonly been to concatenate significant features from PET and CT for multivariate analysis to improve prediction of outcome in different kinds of cancers [21]–[23]. Some studies instead attempted to integrate multi-modality information via image fusion [24]–[27]. Mu *et al.* [24] used weighted summation of PET and CT images to predict immunotherapy response in non-small cell lung cancer (NSCLC) patients. Riyahi *et al.* [25] also directly adopted the weighted summation of normalized PET and CT images to generate a single blended PET-CT image, and then, feature changes between baseline and follow-up were used for tumor response prediction in esophageal cancer. While simple summation of images may be limited, Vallières *et al.* [26] investigated wavelet-based fusion of PET and MRI images to predict lung metastasis in soft-tissue sarcomas, and concluded that fused features could be superior to using separate features extracted from individual images. Zhou *et al.* [27] also investigated usefulness of radiomics features extracted from wavelet-based fusion of PET and MRI to predict progression of patients with mild cognitive impairment to Alzheimer's disease, and showed that the fusion-modality model (C-index = 0.8039) had higher prediction accuracy than the MRI model (C-index = 0.7627) and the PET model (C-index = 0.7755).

Different from aforementioned studies working on feature concatenation and image integration, Parekh *et al.* [28] designed a tissue signature co-occurrence matrix by merging matrices constructed from multiparametric MRI images, and demonstrated improved performance for diagnosis of breast cancer and brain stroke compared to single parameter radiomics. As such, compared to single-modality radiomics features providing limited information corresponding to each imaging modality, multi-modality fusion radiomics features may produce more meaningful features and textural visualization of the underlying tumor regions, and have the potential to provide improved prognosis.

Obviously, different levels of fusion (image-, matrix- and feature-levels) may lead to different radiomics features, which may finally affect the model performance. To the best of our knowledge, multi-modality fusion radiomics features from different levels have not been comprehensively investigated and compared in prognostic tasks. Thus, in the present study, we proposed a multi-level fusion strategy for radiomics analysis, combining the information provided by PET and CT at the image-, matrix- and feature-levels towards improved prognosis (3 outcomes) of multi-center (4 centers) H&N cancer patients (n = 296). (1) In image-level fusion, in addition to wavelet-based fusion (WF) [26], we additionally introduce two advanced fusion methods, namely gradient transfer fusion (GTF) [29] and guided filtering-based fusion (GFF) [30] to fuse PET and CT images arriving at a single fused image, from which features were extracted to fully characterize the information from PET and CT images. (2) In matrix-level fusion, we construct a summed individual enriched matrix (noted as sumMat) by considering the voxel relationships in PET and CT simultaneously. (3) In feature-level fusion, we investigate the concatenation of features from the two imaging modalities (noted as conFea), and the mean values of PET and CT features (noted as avgFea) were also investigated.

Overall, our study investigates the prognostic performance of PET-CT radiomics features extracted by using image-level, matrix-level and feature-level fusion strategies. The rest of this article is structured as follows: Section II describes the dataset, multi-level fusion strategy, feature extraction and statistical analysis in detail. Section III provides the experiment results. Section IV discusses the main finding, limitations and some future directions, followed by conclusions in Section V.

II. Materials and Methods

A. Dataset

The dataset used in this study was from The Cancer Imaging Archive (TCIA) http://www.cancerimagingarchive.net, containing FDG-PET/CT imaging data, clinical data, outcome data and radiotherapy contours (RTstruct) of 296 patients from four different institutions in Québec, Canada. Of these, 65 were from Centre hospitalier de l'Université de Montréal (CHUM, noted as CER1), 100 from Centre hospitalier universitaire de Sherbrooke (CHUS, noted as CER2), 90 from Hôpital général juif (HGJ, noted as CER3) de Montréal, and 41 from Hôpital Maisonneuve-Rosemont (HMR, noted as CER4) de Montréal. All patients were histologically confirmed with H&N cancer, and had pre-treatment FDG-PET/CT scans between April 2006 and November 2014. PET images had varying pixel sizes of 3.52 to 5.47 mm, slice thicknesses of 3.27 to 4 mm, slice spacings of 3.27 to 4 mm with matrix sizes of 128 or 144, while CT images have varying pixel sizes of 0.68 to 1.37 mm, slice thicknesses of 1.5 to 3.75 mm, slice spacings of 1.5 to 3.27 mm (except for 2 cases at 5 mm) with matrix size of 512. There are 47 patients received radiation therapy alone and 249 patients received chemo-radiation therapy. The median follow-up period of all patients was 44 months (range: 6–113). Three outcomes: recurrence-free survival (RFS), metastasis-free survival (MFS) as well as overall survival (OS) were considered. The characteristics of the 296 patients are listed in Table I.

B. Multi-Level Fusion

In order to characterize tumor more comprehensively, we proposed a multi-level fusion strategy to combine the information provided by PET and CT at image-, matrix-and feature-levels. A flowchart of our efforts is illustrated in Fig. 1. As for image-level fusion, three public available fusion methods, namely wavelet-based fusion (WF, https://github.com/mvallieres/radiomics), gradient transfer fusion (GTF, https://github.com/jiayi-ma/GTF) and guided filtering-based fusion (GFF, https://github.com/funboarder13920/image-fusion-guided-filtering) were adopted. PET and CT images were combined into a single fused image, aiming to preserve and extract useful information from both PET and CT images. The WF method first decomposes PET and CT images into 8 sub-bands wavelet coefficients by using 3D discrete wavelet transform (DWT), and then corresponding wavelet coefficients

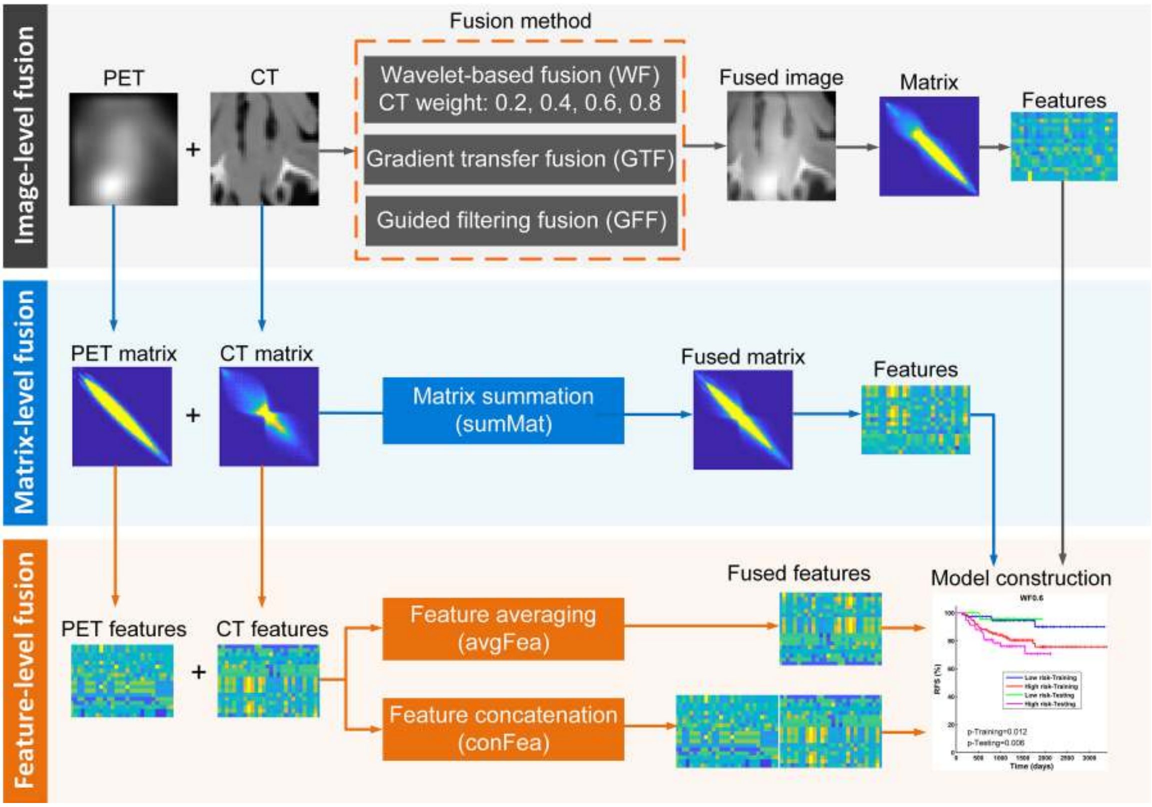


Fig. 1. The flowchart of multi-level fusion strategy (image-, matrix-to feature-level fusion) to combine PET and CT information for the prediction of RFS, MFS and OS for head and neck (H&N) cancer patients.

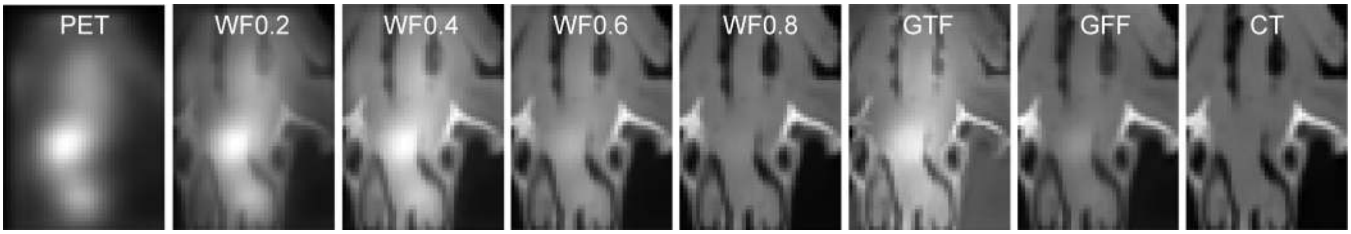


Fig. 2. An example of PET, CT and fused images obtained by wavelet-based fusion with CT weights of 0.2, 0.4, 0.6 and 0.8 (WF0.2, WF0.4, WF0.6, WF0.8), gradient transfer fusion (GTF) and guided filtering-based fusion (GFF) methods.

are weight-averaged by setting CT weights of 0.2, 0.4, 0.6, and 0.8 (noted as WF0.2, WF0.4, WF0.6 and WF0.8). The corresponding PET weight was 1-(CT weight), and the fused image was thus obtained by inverse DWT. PET and CT images without fusion were also used to construct single modality prognostic models. The GTF method formulates the fusion problem as an $\ell 1$-TV minimization problem, where the data fidelity term maintains the main intensity distribution in the PET image, while the regularization term preserves gradient variations in the CT image:

$$\Phi(\mathbf{x}) = \|\mathbf{x} - \mathbf{u}\|_1 + \lambda\|\nabla_{\mathbf{x}} - \nabla_{\mathbf{v}}\|_1 \tag{1}$$

where $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{v}$ represent fused, PET and CT images, respectively, λ is a positive parameter, ∇ is the image gradient, and the default parameters used in [29] were adopted in this study. The GFF method first uses an average filter to get a two-scale representation of both PET and CT images (base layer containing large scale variations in intensity, and a detail layer capturing small scale details), and then the base and detail layers are fused by using a guided filtering-based weighted averaging method. The default parameters of guided filter fusion in [30] were used in this study. Fig. 2 shows an example of PET, CT, and the fused images obtained by WF0.2, WF0.4, WF0.6, WF0.8, GTF and GFF methods.

As for matrix-level fusion, two texture matrices were first constructed from PET and CT images separately, and then were summed up into a single enriched matrix (noted as sumMat). In other words, the summed matrices considered the voxel relationships in PET and CT simultaneously. Seven types of matrices characterize local regional or global image patterns

TABLE I
THE CHARACTERISTIC OF PATIENTS FROM FOUR CENTERS

Characteristic	CHUM (CER1)	CHUS (CER2)	HGJ (CER3)	HMR (CER4)
Patients No.	65	100	90	41
Age (year)	63 (44-90)	64 (34-88)	61 (18-84)	67 (49-85)
Sex, no M/F	49/16	72/28	74/16	30/11
Site, U/N/O/H/L	5/2/57/1/0	0/6/72/1/21	4/14/54/4/14	0/6/19/7/9
Tx/T1/ T2/T3/T4	5/8/28/19/5	0/8/44/31/17	4/19/20/35/12	1/2/17/9/12
N0/N1/N2/N3	4/8/45/8	37/10/50/3	12/18/57/3	5/4/27/5
Stage I/II/III/IV	0/2/7/54	2/17/21/60	1/5/28/56	0/3/5/33
RT/CRT	4/61	31/69	5/85	7/34
HPV, -/+/n.a.	3/21/41	13/25/62	30/30/30	0/2/39
Recurrence	7	15	12	9
Metastasis	3	10	16	11
Death	5	18	14	19
MATV (cm^3)	43.8±2.8	31.4±3.0	41.3±3.4	61.2±6.4
SUVmax	17.1±7.4	11.1±4.7	18.0±7.1	16.4±8.1
SUVmean	5.7±2.7	4.6±2.1	6.4±2.7	5.0±2.4
HUmean	8.1±28.6	20.0±32.8	39.7±16.7	26.3±35.1

M/F: male/female.
U/N/O/H/L: unknown/nasopharynx/oropharynx/hypopharynx/larynx.
RT/CRT: radiotherapy/chemotherapy.
n.a.: not available.

were considered: gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), neighborhood gray tone difference matrix (NGTDM), gray level gap length matrix (GLGLM), neighboring gray level dependence matrix (NGLDM) and gray level distance zone matrix (GLDZM).

As for feature-level fusion, one popular fusion strategy adopted by multiple studies involves concatenation of features from different modalities (noted as conFea) was used in this study, we also use the mean value of PET and CT features (noted as avgFea) to investigate whether it is useful for prognosis. In addition, 8 clinical features (Sex, Age, Site, T stage, N stage, TNM stage, Therapy, HPV status) were collected for model construction. Above models were also expanded to include both clinical and radiomcis features, thus, a total of 23 strategies (one clinical model, 4 single modality model: PET or CT model and 18 fusion models with or without clinical features) were investigated.

C. Feature Extraction

The tumor volume was defined as GTVprimary + GTVlymph nodes, and was applied to the FDG-PET image, CT image and fused images, which evaluated the overall tumor burden including primary tumor and lymph node metastasis. Images were discretized into 64 bins, voxel size was interpolated to an isotropic $1 \times 1 \times 1$ mm^3, and 127 radiomics features were extracted from each region. This included 9 shape features, 11 intensity features, 6 histogram features, 26 GLCM features, 13 GLRLM features, 13 GLSZM features, 5 NGTDM features, 13 GLGLM features, 5 NGLDM features, 16 GLDZM features, and 10 moment features from the Standardized Environment for Radiomics Analysis (SERA) package (https://rahmimlab.com/software/sera/) [31], which in compliance with imaging biomarker standardization initiative (IBSI) guidelines [32].

D. Statistical Analysis

Since four centers were involved in the dataset, apart from one partition adopted in previous studies [33]–[35] (center 2 and center 3 used for training, and center 1 and center 4 used for testing, noted as CER 23 vs. 14), we additionally investigated six other kinds of partitions (CER 12 vs. 34, CER 13 vs. 24, CER 123 vs. 4, CER 124 vs. 3, CER 134 vs. 2 and CER 234 vs 1) by ensuring training patients to be more than testing patients, while those partitions having less training patients than testing patients were excluded. Three endpoints recurrence-free survival (RFS), metastasis-free survival (MFS) and overall survival (OS) were considered. Thus, this study was carried out for 7 partitions × 23 strategies × 3 outcomes. For each set, univariate Cox analysis was conducted by performing 50 repetitions of 3-fold cross validation in the training cohort, the prognostic performance was measured using the concordance index (C-index), from which features were sorted in descending order of the mean validation C-index (in the 150 validation rounds), and the top 10 features were then selected. For feature pairs with Spearman's correlation higher than 0.8, the one with lower C-index was further removed, resulting in a non-redundant candidate feature subset for subsequent multivariate analysis. We considered all possible combinations of the candidate features (in sets of 2 up to 5 features) to perform multivariate Cox analysis in the training cohorts. Let k denote the number of candidate features, then the number of all possible combinations will be $C_k^2 + C_k^3 + C_k^4 + C_k^5$, thus there were a total of 627 combinations/models for $k = 10$ in this study. The final optimal model in the training set was identified by Akaike information criteria (AIC). AIC rewards goodness of fit (as assessed by the log-likelihood) and penalizes number of features simultaneously, requiring increases in log-likelihood (LogL2-LogL1) to be higher than increases in the number of features (FeaNum2-FeaNum1) between two models. We have discussed AIC in comparison to more conservative criteria elsewhere [36]. For each model, the median value of the prognostic score generated in the training cohort was used untouched in the testing cohort as a threshold to separate patients into high- and low-risk subsets, and the difference between the two Kaplan-Meier curves was evaluated by log-rank test. Statistically significant differences between individual C-indices were evaluated by using the R package "compareC" (Version 1.3.1), while statistically significant differences between two series of C-indices were evaluated by using paired Student's t test. Significance level was set as $p < 0.01$ in order to set stricter acceptance given multiple testing. Feature extraction and all other statistical analyses were conducted on Matlab R2018b (The MathWorks Inc.). Since HPV status was only available for 124 patients, HPV status was excluded from multivariate

TABLE II
THE MEAN AND SD OF C-INDEX IN TESTING COHORT OF EACH STRATEGY AMONG ALL 7 DIFFERENT TRAINING AND TESTING PARTITIONS FOR RFS, MFS AND OS

			Clinical+radiomics										
C-index		Clinical	PET	WF0.2	WF0.4	WF0.6	WF0.8	GTF	GFF	sumMat	avgFea	conFea	CT
RFS	Mean	0.58	0.59	0.57	0.57	0.60	0.54	0.56	0.58	0.57	0.60	0.56	0.56
	SD	0.05	0.04	0.05	0.04	0.04	0.03	0.06	0.08	0.05	0.05	0.05	0.03
MFS	Mean	0.61	0.67	0.64	0.64	0.65	0.71	0.61	0.61	0.62	0.63	0.64	0.61
	SD	0.05	0.08	0.11	0.07	0.10	0.13	0.09	0.07	0.10	0.08	0.08	0.09
OS	Mean	0.62	0.65	0.60	0.60	0.62	0.61	0.61	0.62	0.60	0.62	0.60	0.62
	SD	0.06	0.05	0.06	0.04	0.04	0.05	0.06	0.07	0.07	0.05	0.07	0.07
			Radiomics										
			PET	WF0.2	WF0.4	WF0.6	WF0.8	GTF	GFF	sumMat	avgFea	conFea	CT
RFS	Mean		0.59	0.57	0.59	0.61	0.54	0.55	0.56	0.58	0.58	0.56	0.57
	SD		0.05	0.05	0.05	0.06	0.04	0.05	0.06	0.04	0.04	0.04	0.03
MFS	Mean		0.68	0.65	0.63	0.68	0.70	0.62	0.64	0.64	0.64	0.64	0.62
	SD		0.07	0.10	0.08	0.09	0.11	0.08	0.08	0.09	0.05	0.08	0.08
OS	Mean		0.59	0.55	0.56	0.59	0.58	0.57	0.62	0.57	0.61	0.60	0.59
	SD		0.05	0.03	0.04	0.04	0.04	0.04	0.07	0.06	0.03	0.06	0.05

analysis, though the prognostic performance of HPV was later evaluated and reported for these 124 patients.

III. RESULTS

A. Feature Popularity in Univariate Analysis

Supplementary Fig. S1 shows top 10 features with best performance after 50 repetitions of 3-fold cross validation as selected for each partition, each fusion strategy and each outcome; also, feature popularity among the partitions and fusion strategies for each outcome are illustrated. Features selected more than 20 times are listed in supplementary Table S1. Four clinical features (Age, Site, T stage and Therapy) and 3 texture features (IMC2_GLCM, InVar_GLCM and SGLGE_GLGLM) were mostly highly selected towards prediction of RFS. Two clinical features (N stage and T stage), 3 shape features (MV, Compactness2 and Irregularity) and 5 texture features (LGLGE_GLGLM, LGHGE_GLGLM, oJ3, GLN_GLSZM, and ICM1_GLCM) showed high popularity in MFS prediction. Three clinical features (T stage, Site and TNM stage), 3 shape features (Sphericity, Eccentricity and Irregularity) and 3 texture features (RLV_GLRLM, LGHGE_GLGLM and GLN_GLSZM) were most commonly selected for OS prediction.

For the 124 patients whose HPV status are available, as shown in Supplementary Fig. S2, HPV status was only significantly associated with RFS (C-index: 0.76, $p < 0.0001$), HPV + showed lower risk (HR: 0.12, 95% CI: [0.03–0.45]) of recurrence compared with HPV-, while it was not predictive of MFS (C-index: 0.57, p: 0.605) or OS (C-index: 0.63, p: 0.316).

B. Comparison Between different Fusion Strategies

Box plot of Fig. 3 shows the C-index in the testing cohort for each strategy among all 7 different training/testing partitions for RFS, MFS and OS prediction (also detailed in supplementary Table S2); the mean and SD of C-indices are also listed in Table II; corresponding statistically significant difference comparison between strategies can be found in supplementary Fig. 5(a–c) and Fig. S3. For RFS prediction (Fig 5a), WF0.6 (C-index: 0.60 ± 0.04) showed significantly higher performance relative to CT (C-index: 0.56 ± 0.03) with p-value of 0.015. For MFS prediction (Fig. 5b), the C-index of WF0.8 model (0.71 ± 0.13) was significantly higher than that of CT only (0.62 ± 0.08, p-value: 0.020), WF0.2 (0.64 ± 0.11, p-value: 0.003) and GFF (0.61 ± 0.07, p-value: 0.019). For OS prediction

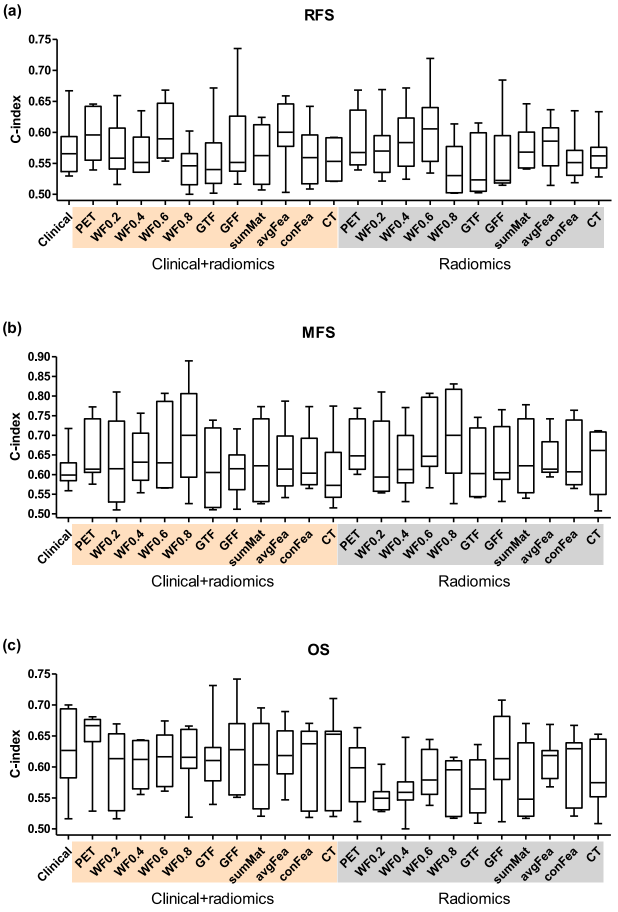


Fig. 3. Box plots of the C-index in testing cohort for each strategy among all 7 different partitions for prediction of (a) RFS, (b) MFS and (c) OS.

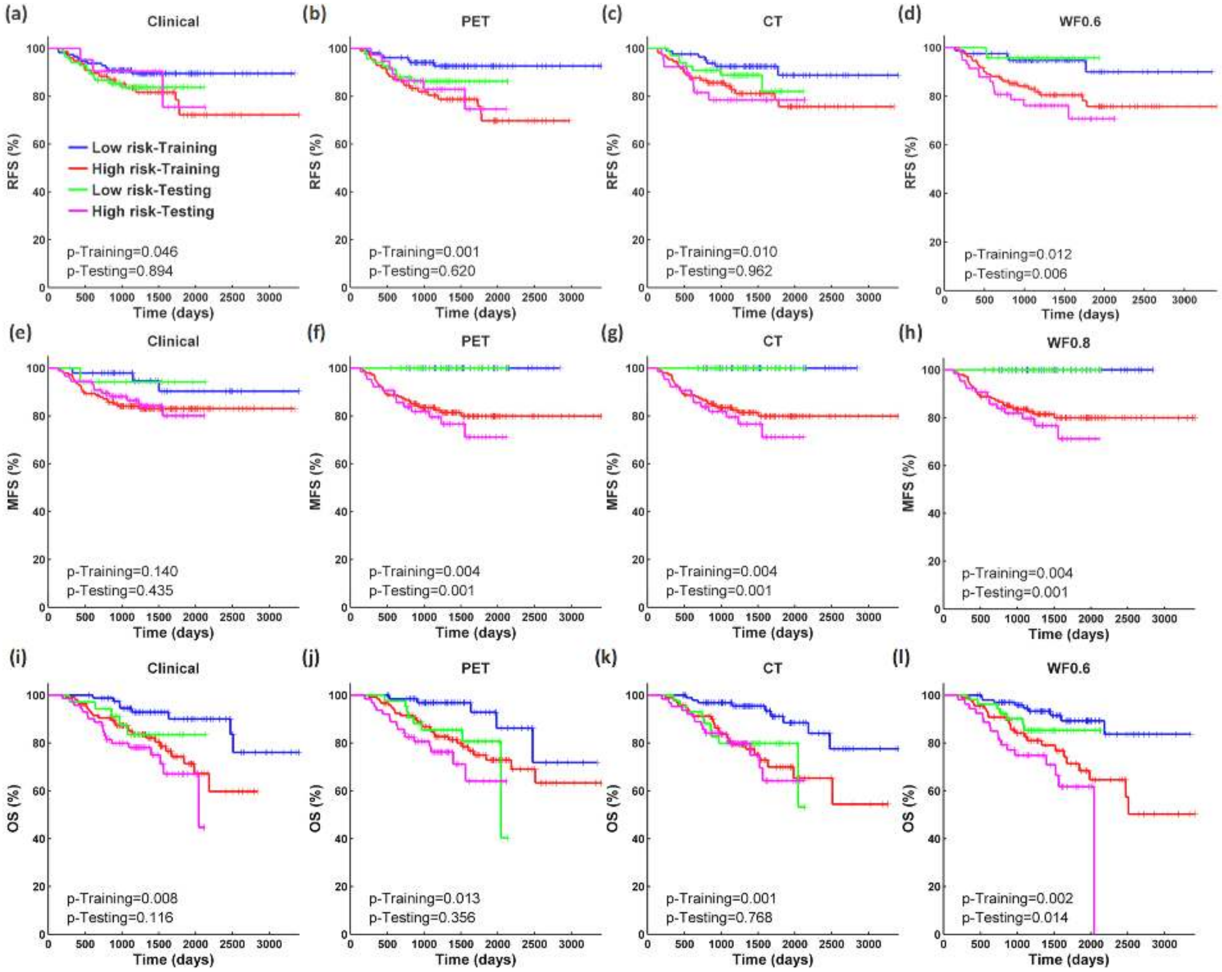
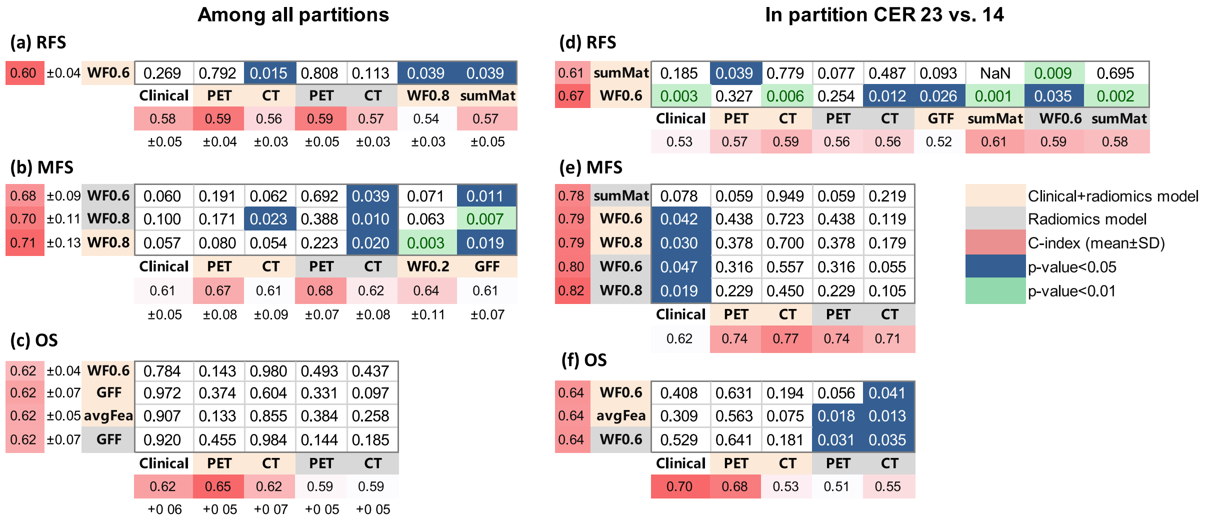


Fig. 4. The Kaplan–Meier curves shown only for partition CER 23 vs. 14. The models shown include the Clinical-only model, PET model and CT model for (a-c) RFS, (e-g) MFS and (i-k) OS prediction; (d) WF0.6 model (wavelet-based fusion with CT weight of 0.6) for RFS, (h) WF0.8 model (wavelet-based fusion with CT weight of 0.8) for MFS and (l) WF0.6 model (wavelet-based fusion with CT weight of 0.6) for OS prediction. The curves for training vs. testing are also color-coded as shown in the legend.

(Fig. 5c), no fusion model significantly outperformed Clinical only, PET only or CT only models. Overall, on average (multiple training and testing in multi-center scenario), fusion model showed limited superiority relative to single modality, though it outperformed CT model in RFS and MFS predictions, highlighting the potential of generalizing radiomics models when utilized in a multi-center setting. More results and discussions below are aimed to provide greater insights into our findings, and we note that in specific partition CER 23 vs. 14, several fusion models indeed demonstrated better performance compared to Clinical model, CT model and PET model.

For specific partition CER 23 vs. 14, Fig. 4 shows the Kaplan–Meier curves of models constructed by using only clinical parameter, PET or CT single modality, and fusion strategy (WF0.6 model, WF0.8 model and WF0.6 model for RFS, MFS, and OS prediction, respectively); corresponding lists of feature combinations are shown in Supplementary Table S3. Statistically significant difference comparisons between strategies are detailed in Fig. 5(d f) and further detailed in supplementary Fig. S4. WF0.6 model which involved clinical and radiomics features showed significantly higher performance than Clinical, CT and sumMat models (regardless of with clinical features or not) for RFS prediction (C-index: 0.67 vs. 0.53, 0.59/0.56 and 0.61/0.58 with p-values of 0.003, 0.006/0.012 and 0.001/0.002). Besides, sumMat model marginally significantly outperformed PET model for both RFS and MFS prediction (C-index: 0.61 vs. 0.57, p-value: 0.039, and 0.78 vs. 0.74, p-value: 0.059, respectively). WF0.8 model without clinical feature marginally significantly outperformed Clinical model (C-index: 0.82 vs. 0.62, p-value: 0.019) for MFS prediction (Fig. 5e). Amongst the fusion models, for OS prediction (Fig. 5f), WF0.6 without clinical feature and avgFea with clinical feature both statistically significantly outperformed PET only (C-index: 0.64 vs. 0.51, with p-value of 0.031 and 0.018, respectively) and CT only (C-index: 0.64 vs. 0.55, with p-value of 0.035 and 0.013, respectively); WF0.6 with clinical feature can statistically significantly separate patients into high-risk and low-risk groups (Fig. 4) in comparison to Clinical only or PET only models, though showing lower C-index of 0.64 (p-value: 0.014) relative to Clinical model of 0.70 (p-value: 0.116) or PET model of



Among all partitions

(a) RFS

0.60	±0.04	WF0.6	0.269	0.792	0.015	0.808	0.113	0.039	0.039
			Clinical	PET	CT	PET	CT	WF0.8	sumMat
			0.58	0.59	0.56	0.59	0.57	0.54	0.57
			±0.05	±0.04	±0.03	±0.05	±0.03	±0.03	±0.05

(b) MFS

0.68	±0.09	WF0.6	0.060	0.191	0.062	0.692	0.039	0.071	0.011
0.70	±0.11	WF0.8	0.100	0.171	0.023	0.388	0.010	0.063	0.007
0.71	±0.13	WF0.8	0.057	0.080	0.054	0.223	0.020	0.003	0.019
			Clinical	PET	CT	PET	CT	WF0.2	GFF
			0.61	0.67	0.61	0.68	0.62	0.64	0.61
			±0.05	±0.08	±0.09	±0.07	±0.08	±0.11	±0.07

(c) OS

0.62	±0.04	WF0.6	0.784	0.143	0.980	0.493	0.437
0.62	±0.07	GFF	0.972	0.374	0.604	0.331	0.097
0.62	±0.05	avgFea	0.907	0.133	0.855	0.384	0.258
0.62	±0.07	GFF	0.920	0.455	0.984	0.144	0.185
			Clinical	PET	CT	PET	CT
			0.62	0.65	0.62	0.59	0.59
			+0 06	+0 05	+0 07	+0 05	+0 05

In partition CER 23 vs. 14

(d) RFS

0.61	sumMat	0.185	0.039	0.779	0.077	0.487	0.093	NaN	0.009	0.695
0.67	WF0.6	0.003	0.327	0.006	0.254	0.012	0.026	0.001	0.035	0.002
		Clinical	PET	CT	PET	CT	GTF	sumMat	WF0.6	sumMat
		0.53	0.57	0.59	0.56	0.56	0.52	0.61	0.59	0.58

(e) MFS

0.78	sumMat	0.078	0.059	0.949	0.059	0.219
0.79	WF0.6	0.042	0.438	0.723	0.438	0.119
0.79	WF0.8	0.030	0.378	0.700	0.378	0.179
0.80	WF0.6	0.047	0.316	0.557	0.316	0.055
0.82	WF0.8	0.019	0.229	0.450	0.229	0.105
		Clinical	PET	CT	PET	CT
		0.62	0.74	0.77	0.74	0.71

(f) OS

0.64	WF0.6	0.408	0.631	0.194	0.056	0.041
0.64	avgFea	0.309	0.563	0.075	0.018	0.013
0.64	WF0.6	0.529	0.641	0.181	0.031	0.035
		Clinical	PET	CT	PET	CT
		0.70	0.68	0.53	0.51	0.55

▨	Clinical+radiomics model
▨	Radiomics model
🟥	C-index (mean±SD)
🟦	p-value<0.05
🟩	p-value<0.01

Fig. 5. Statistically significant p-value comparisons of C-indices between different models (a–c) among all partitions and (d-f) in partition CER 23 vs. 14; Clinical + radiomics models are marked as orange, while radiomics models are marked as gray; p-values lower than 0.05 are marked as blue, while p-values lower than 0.01 are marked as green; higher C-index values are marked as red.

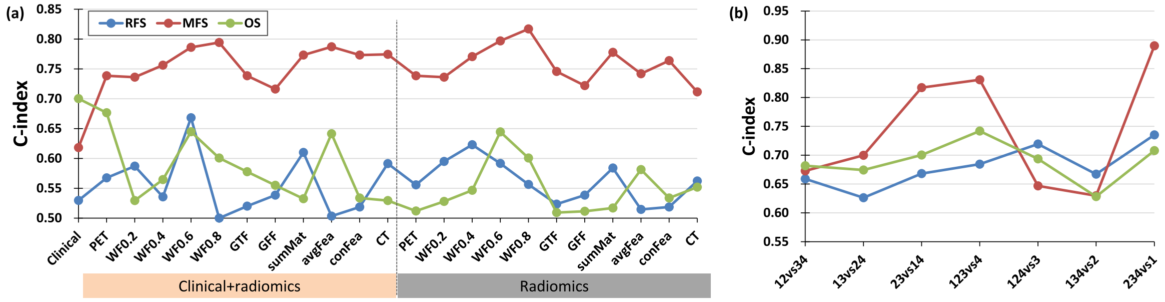


Fig. 6. (a) The C-index in testing cohort of each strategy and each outcome in partition CER 23 vs. 14 (centers 2 and 3 used for training, centers 1 and 4 used for testing); (b) the highest C-index values in the test set in each partition (amongst different strategies).

TABLE III
The Highest C-Index in Each Partition and for Each Outcome Amongst Different Fusion Strategies. Fusion Models Which Involving Radiomics Features Only Without Clinical Feature Are Marked With *, Otherwise, Fusion Models Involved With Both Clinical and Radiomics Features

Partition	RFS			MFS			OS		
	C-index	p-value	Strategy	C-index	p-value	Strategy	C-index	p-value	Strategy
CER 12 vs. 34	0.66	0.122	avgFea	0.67	0.140	WF0.8	0.68	0.019	GFF*
CER 13 vs. 24	0.63	0.118	GFF	0.70	0.266	WF0.8*	0.67	0.002	WF0.6
CER 23 vs. 14	0.67	0.006	WF0.6	0.82	0.001	WF0.8*	0.70	0.116	Clinical
CER 123 vs. 4	0.68	0.225	GFF*	0.83	0.005	WF0.8*	0.74	0.001	GFF
CER 124 vs. 3	0.72	0.017	WF0.6	0.65	0.099	WF0.6*	0.69	0.250	Clinical
CER 134 vs. 2	0.67	0.262	Clinical	0.63	0.194	Clinical	0.63	0.823	WF0.6*
CER 234 vs. 1	0.74	0.481	GFF	0.89	0.228	WF0.8	0.71	0.469	GFF*

0.68 (p-value: 0.356). Fig. 6(a) shows the C-index in testing cohort of each strategy and each outcome in partition CER 23 vs. 14. WF0.6 showed highest C-index of 0.67 for RFS prediction, WF0.8 showed highest C-index of 0.82 for MFS prediction, and Clinical model showed highest C-index of 0.70 for OS prediction.

Results demonstrated varying influences of different fusion strategies depending on the predicted outcome; this is likely due to the fact that different fusion strategies can highlight different aspects of biology, which could be more specifically related to specific outcomes. Thus, specific models instead of a general model need to be developed for particular application.

C. Comparison Between Different Partitions

Table III and Fig. 6(b) show the highest testing C-index in each partition under specific strategy for each outcome; the corresponding feature combinations are detailed in supplementary Table S4. Highest performance in 5, 6 and 5 partitions was achieved by image-level fusion strategies for RFS, MFS and

OS prediction, respectively, revealing the potential of fusion radiomics for outcome prediction. For RFS prediction, only two partitions CER 234 vs. 1 and CER 124 vs. 3 showed C-index>0.7 (0.74 and 0.72 by GFF and WF0.6 strategies, respectively) while all other 5 partitions showed C-index<0.7 (0.63-0.68). For MFS prediction, partitions CER 234 vs. 1, CER 123 vs. 4 and CER 23 vs. 14 showed higher C-index (0.89, 0.83 and 0.82 all by WF0.8, respectively) than other partitions (0.63-0.70). For OS prediction, CER 123 vs. 4 showed highest C-index of 0.74 by GFF among all partitions, followed by CER 234 vs. 1 and CER 23 vs. 14 both with C-indices of 0.71 and 0.70 by GFF and Clinical model, respectively. Prognostic performance was strongly influenced by data partitioning; as such, batch effect removal [37], caution in interpretation and use of extensive validation are necessary for translation of radiomics into clinical practice.

IV. Discussion

In this work, we first investigated and compared multi-modality radiomics combining the information provided by PET and CT at the image-, matrix- and feature-levels towards improved prognosis of head and neck cancer patients in a multi-center manner. PET and CT images are routinely generated from a single examination in clinical practice. Merging anatomical tissue density provided by CT and functional glucose metabolism reflected by PET can be critical to the quantification of intra-tumor heterogeneity and the subsequent outcome prediction. Low attenuation regions in CT images are well known to correlate with necrosis change [38]. Higher metabolism regions in PET images result from sufficient oxygen, glucose and nutrients provided by neovascularization. Our results demonstrated that image-level fusion of PET and CT provides potential prognostic performance compared with matrix-level fusion strategy as well as use of clinical features, PET features or CT features alone only for a specific partition of training and testing while not for all partitions on average. Our findings point out future directions for investigating advanced multi-modality image fusion methods in application of radiomics analyses on outcome prediction, also highlighting the potential of generalizing radiomics models when utilized in a multi-center scenario.

Three previous studies [33]–[35] used center HGJ and CHUS for training, and center HMR and CHUM for testing (i.e., CER 23 vs. 14 in our study); thus, below comparison is based on this partition. In comparison with the model reported by Vallières *et al.* [33], our model showed lower C-index for prediction of MFS (0.82 vs. 0.88) and OS (0.70 vs. 0.76) and comparable C-index (0.67 vs. 0.67) for prediction of RFS. We note that in that work, the authors performed random forest for final model construction, while the multivariate Cox analysis was adopted in our study considering time-to-event continuous variable instead of binary variable. Besides, they extracted a total of 1615 radiomics features by considering several isotropic voxel size and gray level discretization methods; however, only 1 mm voxel size and 64 bins gray level were adopted in our study, as more features involvement may result in higher risk of false discovery rate. The difference in methodology may result in variations in performance [39].

In the study by Diamant *et al.* [34], deep learning was only applied to CT images while clinical parameters were not integrated, and only AUC values were reported (requiring a relatively *ad hoc* time-to-event cut-off threshold), while the C-index (which preserves the continuous time-to-outcome information) was not reported. Thus, our results are not directly comparable; but to summarize, the authors obtained AUC values of 0.65, 0.88 and 0.70 for RFS, MFS and OS, in comparison to our C-index values of 0.67, 0.82 and 0.70. Their study only made use of CT images and not PET images, and thus our study sheds light on the relative value of and optimal approaches to fusing information from both modalities. As for the study by Bizzego *et al.* [35], they concluded that combining radiomics and deep learning features from both PET and CT images outperformed using only one feature type or single modality. The authors only considered prediction of recurrence, and reported Matthews Correlation Coefficient (MCC) of 0.748 instead of AUC or C-index, which is not directly comparable with our C-index of 0.67. Besides, the final model contained 261 radiomics features and 239 deep learning features, which are much more than our maximum of 5 features.

There are a number of metrics (such as entropy, standard deviation and fusion mutual information etc.) that are used in other applications for evaluation of fused image quality [40], which are beyond the scope of the current study. We directly evaluated the prognostic performance of each fusion strategy aiming to identify the ones with higher prognostic performance instead of evaluating the fusion quality by aforementioned metrics. At the same time, we performed feature difference comparisons between different strategies. Since (i) shape features are irrelevant for fusion strategies, (ii) the sumMat strategy was only concerned with matrix-based texture features, and (iii) the conFea strategy identically contained PET and CT features, we conducted feature difference comparisons only on 91 texture features from sumMat and 108 intensity and texture features from remaining strategies on all patients. As shown in supplementary Fig. S5, only 5/108 features from PET and CT were highly correlated, consistent with the fact that these two modalities capture complementary aspect of tumors. Most features from WF0.8, GFF and avgFea were weakly correlated with both PET features and CT features, providing an insight towards their good prognostic performance. Most features from WF0.2, WF0.4, WF0.6 and GTF were highly correlated with PET features while not with CT features, meaning that fused images generated by these four fusion strategies keep more information from PET, and are more similar to PET compared to CT. Most features from sumMat were highly correlated with CT features, indicating that CT intensity distributions make higher contribution to the construction of fused matrices. Thus, these features are surrogates to PET or CT features, and may not be able to generate more useful information, thus hindering their added value to use of PET or CT alone.

Four features in the WF0.6 model showed highest prognostic value for RFS prediction in partition CER 23 vs. 14, and the prognostic score was formulated as 0.87 × Age + 0.91 × SZHGE_GLSZM-0.71 × LRHGE_GLRLM + 1.06 × B3. This indicates that older age, higher

SZHGE_GLSZM and higher B3 to be associated with poor prognosis, while LRHGE_GLRLM is negatively correlated with recurrence. Small zone high gray-level emphasis from gray level size zone matrix (SZHGE_GLSZM), describing scattered high intensity small regions, may represents specific habitats that are resistant to therapy. Long run high gray level emphasis from gray level run length matrix (LRHGE_GLRLM) describes long strip high intensity vessel patterns, which may imply tumors with good oxygen supply, which are sensitive to therapy thus with good prognosis as hypoxia is associated with poorer prognosis [41]. Moment invariants feature B3 describes the complexity of both shape and intensity distribution, may also reveal the intratumor heterogeneity.

We considered all possible combinations of the candidate features to perform multivariate Cox analysis, and the final optimal model in the training set was identified by Akaike information criteria (AIC). This procedure alleviates the impact of different model initializations in forward stepwise selection [36], while it's only suitable for small number of candidate features, since large number of candidate features required large search space and will be time consuming. In this study, we tried 627 combinations (up to 5 features) among top 10 features of 155–255 training patients within few seconds, providing a trade-off between higher model performance without over fitting [42] and acceptable running time. When using other stricter criteria (e.g., Bayesian information criterion-BIC) for model selection, less features will be retained in final model and will result in lower performance, thus, in this study, we preferred AIC. As for partition CER 234 vs. 1 (Table III), three C-indices of RFS, MFS and OS prediction were higher than 0.70, while the corresponding log-rank p-values (0.228–0.481) were not significant, most probably because only 7, 3 and 5 patients had recurrence, metastasis and death among 65 testing patients in center 1, respectively.

We investigated fusion models by involving radiomics features with or without clinical features. When using only radiomics features, the performance for OS prediction was seen to be lower compared to using both radiomics and clinical features. As shown in Supplementary Table S5, WF0.2, GTF and sumMat without clinical feature showed significantly lower mean C-index compared to these models when involving clinical feature (C-index: 0.55 vs. 0.60, p-value: 0.043, 0.57 vs. 0.61, p-value: 0.028, 0.57 vs. 0.60, p-value: 0.043, respectively). This suggests that these two kinds of features are complementary to one another. Our study was conducted in a multi-institution data setting. As shown in Supplementary Fig S6 and Table S6, clinical features such as Age, Site, T stage, N stage and TNM stage, along with several radiomics features from GLCM, GLGLM and GLSZM were popular in the final multivariate model, demonstrating their importance across different institutions. Furthermore, most popular features in univariate analysis (Supplementary Table S1) were also highly selected in multivariate analysis.

This study has some limitations: HPV status was not available for all patients and was not invoked in model construction, potentially limiting model performance. We applied GTF and GFF with default parameters to our dataset, while parameter adjustment in different fusion strategies may result in variation of prognostic performance. Future directions including validation on larger cohorts of patients and investigation of more image fusion methods on radiomics analysis.

V. Conclusion

In this study, we proposed multi-level fusion strategies to combine the information provided by PET and CT at the image-, matrix- and feature-levels towards improved prediction of outcome (RFS, MFS and OS) in multi-center (4 centers) head and neck cancer patients (296 subjects). Fusion radiomics model showed varying improvements compared to single modality models for different outcome prediction in different training and testing partitions, highlighting the potential of generalizing radiomics models when utilized in a multi-center scenario. Integrating information at image level (i.e., merging metabolic information in PET and anatomic information in CT voxel by voxel) holds potential to capture more useful characteristics.

References

[1] J. Ferlay *et al.*, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, Mar. 2015.

[2] V. Paidpally, A. Chirindel, S. Lam, N. Agrawal, H. Quon, and R. M. Subramaniam, "FDG-PET/CT imaging biomarkers in head and neck squamous cell carcinoma," *Imag. Med.*, vol. 4, no. 6, pp. 633–647, Dec. 2012.

[3] K. K. Ang *et al.*, "Human papillomavirus and survival of patients with oropharyngeal cancer," *New England J. Med.*, vol. 363, no. 1, pp. 24–35, Jul. 2010.

[4] E. A. Mroz, A. D. Tward, R. J. Hammon, Y. Ren, and J. W. Rocco, "Intra-Tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the Cancer Genome Atlas," *PLOS Med.*, vol. 12, no. 2, Feb. 2015, Art. no. e1001786.

[5] W. Lv *et al.*, "Robustness versus disease differentiation when varying parameter settings in radiomics features: Application to nasopharyngeal PET/CT," *Eur. Radiol.*, vol. 28, no. 8, pp. 3245–3254, Aug. 2018.

[6] M. Hatt *et al.*, "18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort," *J. Nucl. Med.*, vol. 56, no. 1, pp. 38–44, Jan. 2015.

[7] S. Marur and A. A. Forastiere, "Head and neck Squamous Cell Carcinoma: Update on Epidemiology, diagnosis, and treatment," *Mayo Clinic Proc.*, vol. 91, no. 3, pp. 386–396, Mar. 2016.

[8] Z. Liu *et al.*, "The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges," *Theranostics*, vol. 9, no. 5, pp. 1303–1322, 2019.

[9] A. J. Wong, A. Kanwar, A. S. Mohamed, and C. D. Fuller, "Radiomics in head and neck cancer: from exploration to application," *Transl. Cancer Res.*, vol. 5, no. 4, pp. 371–382, Aug. 2016.

[10] L. Lu *et al.*, "Robustness of radiomic features in [C-11] choline and [F-18] FDG PET/CT imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization," *Mol. Imag. Biol.*, vol. 18, no. 6, pp. 935–945, Dec. 2016.

[11] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, and H. J. Aerts, "Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer," *Frontiers Oncol.*, vol. 5, 2015, Art. no. 272.

[12] M. Bogowicz *et al.*, "Computed tomography radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 99, no. 4, pp. 921–928, Nov. 2017.

[13] A. Jethanandani *et al.*, "Exploring applications of radiomics in magnetic resonance imaging of head and neck cancer: A systematic review," *Frontiers Oncol.*, vol. 8, 2018, Art. no. 131.

[14] M. Bogowicz *et al.*, "Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models," *Radiotherapy Oncol.*, vol. 125, no. 3, pp. 385–391, Dec. 2017.

[15] B. F. Branstetter *et al.*, "Head and neck malignancy: is PET/CT more accurate than PET or CT alone?," *Radiology*, vol. 235, no. 2, pp. 580–586, May 2005.

[16] H. Zhang *et al.*, "Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy," *Radiology*, vol. 269, no. 3, pp. 801–809, Dec. 2013.

[17] S. Koyasu *et al.*, "Prognostic value of pretreatment ^{18}F-FDG PET/CT parameters including visual evaluation in patients with head and neck squamous cell carcinoma," *Amer. J. Roentgenol.*, vol. 202, no. 4, pp. 851–858, Apr. 2014.

[18] S. Bisdas, K. Spicer, and Z. Rumboldt, "Whole-tumor perfusion CT parameters and glucose metabolism measurements in head and neck squamous cell carcinomas: a pilot study using combined positron-emission tomography/CT imaging," *Amer. J. Neuroradiol.*, vol. 29, no. 7, pp. 1376–1381, Aug. 2008.

[19] H. Schoder, H. W. Yeung, M. Gonen, D. Kraus, and S. M. Larson, "Head and neck cancer: Clinical usefulness and accuracy of PET/CT image fusion," *Radiology*, vol. 231, no. 1, pp. 65–72, Apr. 2004.

[20] M. Bogowicz *et al.*, "Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma," *Acta Oncol.*, vol. 56, no. 11, pp. 1531–1536, Nov. 2017.

[21] W. Lv *et al.*, "Radiomics analysis of PET and CT components of PET/CT imaging integrated with clinical parameters: Application to prognosis for nasopharyngeal carcinoma," *Mol. Imag. Biol.*, vol. 21, pp. 954–964, Jan. 2019.

[22] M. Vaidya, K. M. Creach, J. Frye, F. Dehdashti, J. D. Bradley, and I. El Naqa, "Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer," *Radiotherapy Oncol.*, vol. 102, no. 2, pp. 239–245, Feb. 2012.

[23] C. Lartizien, M. Rogez, E. Niaf, and F. Ricard, "Computer-aided staging of lymphoma patients with FDG PET/CT imaging based on textural information," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 3, pp. 946–955, May 2014.

[24] W. Mu *et al.*, "Radiomic biomarkers from PET/CT multi-modality fusion images for the prediction of immunotherapy response in advanced non-small cell lung cancer patients," *Proc. SPIE*, vol. 10575, 2018, Art. no. 105753S.

[25] S. Riyahi *et al.*, "Quantification of local metabolic tumor volume changes by registering blended PET-CT images for prediction of pathologic tumor response," *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*. Berlin, Germany: Springer, 2018. pp. 31–41.

[26] M. Vallieres, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Phys. Med. Biol.*, vol. 60, no. 14, pp. 5471–5496, Jul. 2015.

[27] H. Zhou *et al.*, "Dual-Model radiomic biomarkers predict development of mild cognitive impairment progression to Alzheimer's disease," *Frontiers Neurosci.*, vol. 12, 2018, Art. no. 1045.

[28] V. S. Parekh and M. A. Jacobs, "MPRAD: A multiparametric radiomics framework," 2018, *arXiv:1809.09973*.

[29] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[30] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[31] S. Ashrafinia *et al.*, "Quantitative Nuclear Medicine Imaging using Advanced Image Reconstruction and Radiomics," Ph.D. dissertation, Dept. Elect. Comput. Eng. Radiol., Johns Hopkins Univ., Baltimore, MD, USA, 2019.

[32] A. Zwanenburg, *et al.*, "Image biomarker standardisation initiative," 2016, *arXiv:1612.07003*.

[33] M. Vallieres *et al.*, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Sci. Rep.*, vol. 7, no. 1, Aug. 2017, Art. no. 10117.

[34] A. Diamant, A. Chatterjee, M. Vallieres, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Sci. Rep.*, vol. 9, no. 1, Feb. 2019, Art. no. 2764.

[35] A. Bizzego *et al.*, "Integrating deep and radiomics features in cancer bioimaging," in *Proc. Conf. Comput. Intell. Bioinformat. Comput. Biol.*, 2019, pp. 1–8.

[36] A. Rahmim *et al.*, "Prognostic modeling for patients with colorectal liver metastases incorporating FDG PET radiomic features," *Eur. J. Radiol.*, vol. 113, pp. 101–109, Apr. 2019.

[37] C. Chen *et al.*, "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods," *PLOS One*, vol. 6, no. 2, Feb. 2011, Art. no. e17238.

[38] R. M. Subramaniam, M. Truong, P. Peller, O. Sakai, and G. Mercier, "Fluorodeoxyglucose-positron-emission tomography imaging of head and neck squamous cell cancer," *Amer. J. Neuroradiol.*, vol. 31, no. 4, pp. 598–604, Apr. 2010.

[39] D. Du *et al.*, "Machine learning methods for optimal radiomics-based differentiation between recurrence and inflammation: Application to nasopharyngeal carcinoma post-therapy PET/CT images," *Mol. Imag. Biol.*, to be published.

[40] P. Jagalingam and A. V. Hegde, "A review of quality metrics for fused image," in *Proc. Int. Conf. Water Resources, Coastal Ocean Eng.*, 2015, vol. 4, pp. 133–142.

[41] B. Muz, P. de la Puente, F. Azab, and A. K. Azab, "The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy," *Hypoxia*, vol. 3, pp. 83–92, 2015.

[42] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden, "False discovery rates in PET and CT studies with texture features: A systematic review," *PLOS One*, vol. 10, no. 5, May 2015.