Integrating deep and radiomics features in cancer bioimaging

A. Bizzego, N. Bussola
University of Trento &
Fondazione Bruno Kessler
Trento, Italy
andrea.bizzego@unitn.it
bussola@fbk.eu

D. Salvalai, M. Chierici, V. Maggio
G. Jurman†, C. Furlanello
Fondazione Bruno Kessler
Trento, Italy
{salvalai | chierici | vmaggio | jurman | furlan}@fbk.eu
† (Corresponding Author)

***Abstract*—Almost every clinical specialty will use artificial intelligence in the future. The first area of practical impact is expected to be the rapid and accurate interpretation of image streams such as radiology scans, histo-pathology slides, ophthalmic imaging, and any other bioimaging diagnostic systems, enriched by clinical phenotypes used as outcome labels or additional descriptors. In this study, we introduce a machine learning framework for automatic image interpretation that combines the current pattern recognition approach ("radiomics") with Deep Learning (DL). As a first application in cancer bioimaging, we apply the framework for prognosis of locoregional recurrence in head and neck squamous cell carcinoma (N=298) from Computed Tomography (CT) and Positron Emission Tomography (PET) imaging. The DL architecture is composed of two parallel cascades of Convolutional Neural Network (CNN) layers merging in a softmax classification layer. The network is first pretrained on head and neck tumor stage diagnosis, then fine-tuned on the prognostic task by internal transfer learning. In parallel, radiomics features (*e.g.*, shape of the tumor mass, texture and pixels intensity statistics) are derived by pre-defined feature extractors on the CT/PET pairs. We compare and mix deep learning and radiomics features into a unifying classification pipeline (RADLER), where model selection and evaluation are based on a data analysis plan developed in the MAQC initiative for reproducible biomarkers. On the multimodal CT/PET cancer dataset, the mixed deep learning/radiomics approach is more accurate than using only one feature type, or image mode. Further, RADLER significantly improves over published results on the same data.**

***Index Terms*—Radiomics, Deep Learning, Integration**

I. Introduction

Artificial Intelligence (AI) progress in medical image interpretation is rapidly gaining speed, with a wide range of applications [1]–[3]. Its translation to clinical practice is expected to accelerate due to faster regulatory approval procedures for medical algorithms [4]. As deep learning models (DL) aim to evolve status from exploratory to clinically effective solutions, interpretability remains a major stepping hindrance [4], [5]. In general terms, DL provides a class of machine learning methods that can model complex abstractions of patterns through multiple non-linear transformations estimated by data-driven training procedures. Convolutional Neural Networks (CNNs) are DL models widely successful in image recognition and classification. Their application in medical image analysis dates back to 1996, to discriminate tumor mass and normal tissue in mammography [6]. Since then, CNNs have provided results comparable to experts in the diagnosis of skin lesions [7], classification of colon polyps [8], [9], survival analysis of glioma [10], ophthalmology [11], histology [12], and other areas [1].

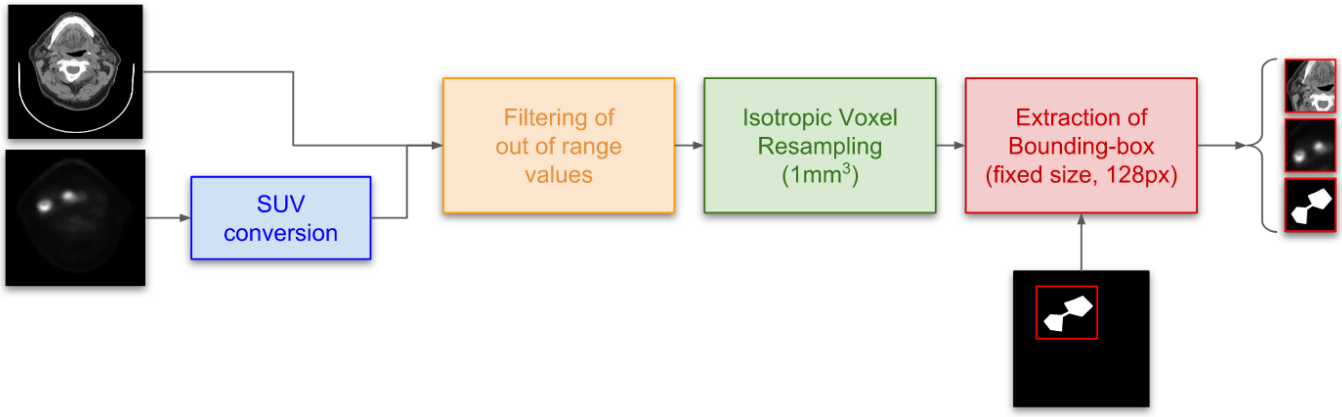
Medical imaging is indeed a key resource shaping the clinical trajectory of a patient. Based on these initial success stories, DL techniques are expected to represent a major breakthrough in diagnosis, treatment decision, prognosis and treatment evaluation. This breakthrough is expected to be pervasive and valid over the diverse medical imaging modalities, *i.e.*, anatomical (such as CT scan) or functional (*e.g.*, PET).

In apparent competition with DL, radiology is already walking fast on a critical innovation path enlightened by *radiomics*, the umbrella-term for pattern recognition methods composed by quantitative image feature extraction paired with statistical or standard machine learning classifiers. Radiomics is grounded on the underlying biological assumption that imaging features can capture distinct phenotype morphology [2], thus achieving both classification and clinical understanding in the machine learning process.

This emphasis on interpretability is a key factor in oncology, where molecular expressions of cancer subtypes may manifest as tissue architecture and nuclear morphological alterations [13]; hence automatic evaluation of disease aggressiveness and patient subtyping can be derived to inform the therapeutic decision. Radiomics features include descriptors of intensity distribution, spatial relationships, texture heterogeneity patterns, descriptors of shape and morphology, and volumetric quantification [14], [15]. Radiomics features can be extracted by tools such as the cancer imaging phenomics toolkit (*CaPTk*) for radiographic images [16], *histomicsTK* for histological Whole Slide Images, or *pyRadiomics* [17].

The role of features in DL is remarkably different: by construction, data are non-linearly mapped throughout the transformation connecting the input and output spaces of a neural network. At each layer, data are projected in a synthetic feature space defined by the training process; such latent features can be investigated in association to the outcome

Fig. 1. Workflow of CT and PET volumes preprocessing pipeline. SUV: Standardized Uptake Values



labels. Although hard to define in biological or morphological terms, these learned features can outperform the hand-crafted ones [18]–[21].

However, DL models typically need a much larger amount of data for training for optimal results than statistical machine learning models; thus these models are often bootstrapped With the transfer learning approach, *i.e.* borrowing weights of models trained on different domains, and possibly retraining only a sector of interest of the network with the data from the novel task [22]. This trick is extensively used in non-medical domains, based on the availability of large-scale data and pre-trained architectures [23], [24]. Recently, these resources are becoming available in cancer research. For example, the *DeepLesion* dataset, containing over $32,000$ annotated lesions in CT scans [25], and *The Cancer Imaging Archive* (TCIA), which provides medical images of different modalities (MRI, CT, etc.) [26].

The success of transfer learning schemas is clearly contributing to approaching DL models as powerful extractors of useful feature sets (i.e. *deep features*). However, linking deep features to meaningful clinical properties interpretable by physicians remains a key challenge [3]. Statistical machine learning approaches are also still widely used in radiomics [27], [28]. This state of the art has naturally led to the idea of a hybrid combination of hand-crafted radiomics (HCR) and deep-learning radiomics (DLR) in an integrated system system [4], [29], [30]. These systems can provide objective characterizations of tumor and a more effective decision support environment, activating expertise in interpretation by clinicians, biologists and bioinformaticians [31].

Notably, the fusion between the two radiomics feature types operates either at decision level or at feature level. With the first approach, models built on HCR and DLR features are developed separately and a final decision module combines their outputs [19], [32]. With the second approach, the integration of HCR and DLR features operates at early level in a multimodal learning framework (e.g. by concatenation), usually with better classification performance [23], [33]–[37].

In this work, we propose RADLER, an automatic pipeline for the integration of DLR and HCR features for medical images analysis, in a first application on multimodal PET/CT scans. To support reproducibility, models are trained with a Data Analysis Plan (DAP) that includes repeate cdross-validation, model selection and feature ranking techniques. To validate the framework, an application is shown on a dataset of two-modality 3D CT/PET scans for prognosis of locoregional recurrence (LR) in head and neck squamous cell carcinoma (N=298), previously solved with a HCR approach and a logistic regression model [38]. The multimodal network architecture is derived from a multi-stream multi-scale architecture for lung cancer screening [39]. The network is first pretrained on head and neck tumor stage (T-stage) diagnosis, then fine-tuned on the prognostic task (internal transfer learning). The RADLER model integrates in this case up to four feature types (CT-HCR, CT-DLR, PET-HCR, PET-DLR) improving over the published results on the same data [38]. Moreover the mixed deep learning/radiomics approach is more accurate than using only one feature type, or image mode.

II. Materials and Methods

Head-Neck-PET-CT Dataset

The Head-Neck-PET-CT (HN) dataset [1] has been originally introduced in [38], and further used in [40]. It includes medical images and clinical data of 298 patients with head and neck squamous cell carcinoma.

For each patient, the HN dataset provides CT and PET scans and Gross Tumor Volume (GTV) mask, preprocessed according to the pipeline in Fig. 1. Several clinical variables are included, in particular the Locoregional Recurrence (LR) within the follow-up period (median: 43 months; range: 6-112 months), and T-stage at diagnosis. Data are gathered from four different hospitals, each one representing a single cohort. Notably, each hospital has its own image acquisition equipment and acquisition settings, which is a cause of heterogeneity in image characteristics, in particular resolution of the PET

[1]publicly available at https://wiki.cancerimagingarchive.net/display/Public/Head-Neck-PET-CT

	no-LR	LR	total
Train	164	27	191
Test	88	15	103

TABLE I

HN DATA: CLASS DISTRIBUTION OF THE LR PROGNOSTIC TASK IN TRAIN AND TEST SETS (N=294).

images. Moreover, The HN dataset is highly unbalanced for the LR prognosis, with 15.8% of recurrence (Table I).

For the sake of comparison, we split the HN dataset into training and test sets with the same partition as in the original study [38]. In particular, two hospital cohorts are used to train the model (*training set*, $N_{tr} = 191$), and two cohorts are used for testing (*test set*, $N_{te} = 103$), with proportioned class stratification (see Table I); the design is chosen to consider possible batch effects in the subcohorts due to the hospital of provenance.

The data set includes the secondary diagnostic label tumour stage (T-stage: score in a 1-4 scale), which was considered for the internal transfer learning strategy. Patients missing the T-stage were not considered in training the diagnostic model, thus developed on a subset of 269 patients, partitioned into 60/40% train/test sets (see Table II).

Image Processing Workflow

As summarized in the diagram in Fig. 1, CT and PET images are first preprocessed to obtain a standardized input information. In particular, PET images are converted to Standardized Uptake Values (SUVs), applying the protocol proposed by the Quantitative Imaging Biomarkers Alliance (QIBA), which also considers vendor-dependent parameters. It is worth mentioning that the conversion of PET images to SUV format is still an open question [41], [42]. The GTV pattern is imported by creating a binary mask with the same size of the CT and PET images ("GTV mask"). For both CT and PET modalities, the preprocessing pipeline includes: thresholding on the pixel values; isotropic voxel resampling; and extraction of the 3D volume containing the GTV.

The intensity in CT images is associated to tissue density and it is measured with the Hounsfield scale (HS). Specific HS value ranges are defined for each type of tissue, which allows the direct comparison of images from different vendors. However, artifacts or acquisition errors might affect the image and give pixel values outside the physiological range. We filter these artifacts by thresholding the pixel values of CT images between HS=-1050 (air density score) and HS=3050 (bone density score). Similarly, the SUVs in PET images are thresholded in the range between 0 and 50 to avoid artifacts due to errors in sensors readings.

Further, isotropic voxel resampling was performed on CT and PET images, as well as on the GTV mask, based on cubic interpolation of each image on a 3D grid with 1 mm^3 voxels, in order to have an homogeneous standard spatial information.

The last module in the image preprocessing pipeline extracts a subvolume of the image which contains the GTV. This reduction enables to compute the radiomics features only from the voxels, also reducing the size of the 3D image portion to analyze with DL on the Graphical Processing Unit (GPU) memory. The drawback of this operation is the loss of contextual information near the GTW, thus the normalized size of the subvolume was set to 128 mm^3, a reasonable trade-off between the size of the GTVs in the dataset and the amount of context included. The volume of interest was centered in the center of mass of the GTV, also used to center the subvolumes of the CT, PET and GTV mask images.

In summary, the output of the image processing workflow (see Fig. 1) is composed by three 128 mm^3 images, one for each modality and for the GTV mask.

	T1	T2	T3	T4	total
Train	29	64	36	34	163
Test	10	45	28	23	106

TABLE II

HN DATA: CLASS DISTRIBUTION OF THE DIAGNOSTIC T-STAGE TASK IN TRAIN AND TEST SETS (N=269).

III. THE RADLER INTEGRATIVE RADIOMICS DAP

We have developed the RADLER radiomics pipeline as a general framework for predictive models that can integrate Deep Learning and predefined features. The framework is also designed to manage multimodal imaging datasets, as in the case of study of LR prognosis on the HN cancer dataset. The main steps in the RADLER pipeline after the preprocessing phase are described below as exemplified on the LR HN task (see Fig. 3.

Radiomics Feature Extraction and Integration

Three sets of radiomics features are considered:

i) *HCR.* A total of 3,249 radiomics features are extracted for each patient, replicating [38]. The feature extraction is based on the `pyradiomics` framework [29]. The HCR features are chosen to describe three main image properties: shape (13 features, based on the GTV contours), intensity (18 features, based on the voxel intensities), and texture (1,600 features, based on four Gray-Level Matrices). Following [38], 40 types of texture features were considered, each one computed on 40 sets of parameters that define pixel spacing, quantization method and number of gray levels;

ii) *DLR.* A total number of 512 deep features are extracted (256 from PET images and 256 from CT images) as a byproduct of a multimodal neural network. The network is trained on CT and PET simultaneously [39], with two identical and parallel convolutional branches merged in a fully connected layer (see Fig. III for the details on the architecture). An internal transfer learning procedure is applied by first training the whole network on the T-stage dataset, then predicting LR by fine-tuning, *i.e.* retraining only the linear blocks (final blue box in Fig. III). Fixed hyper-parameters are used to regulate the training process

Fig. 2. UMAP embedding of the radiomics deep features extracted from the PET images (PET-DLR). Each point represents a patient, colour coded for T-stage. On the right panel, a qualitative trajectory of cancer severity is overlaid on patient clusters of increasing T-stage.

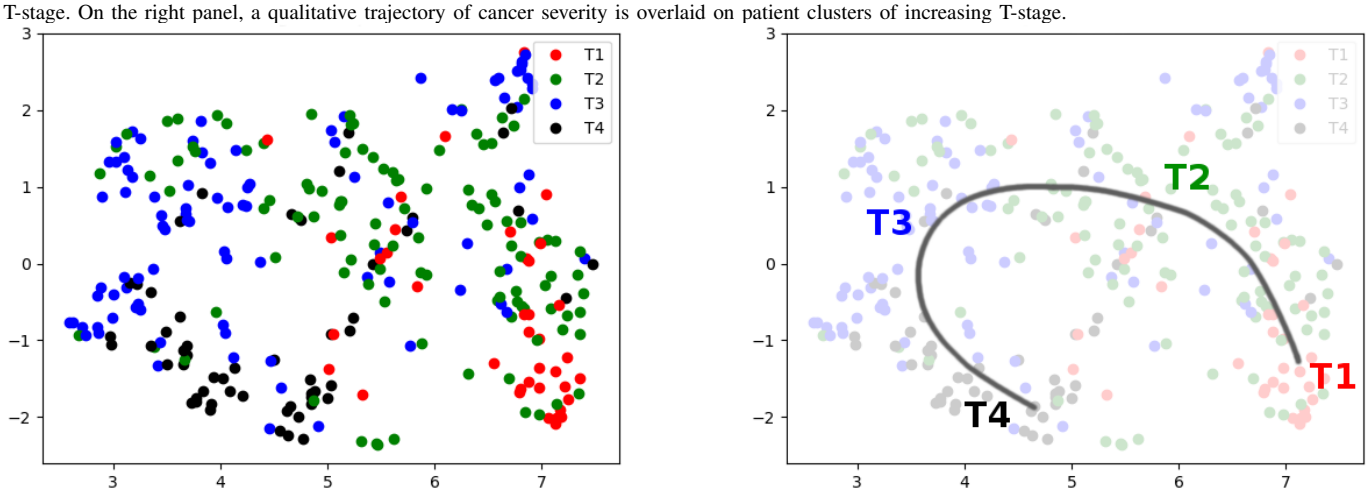
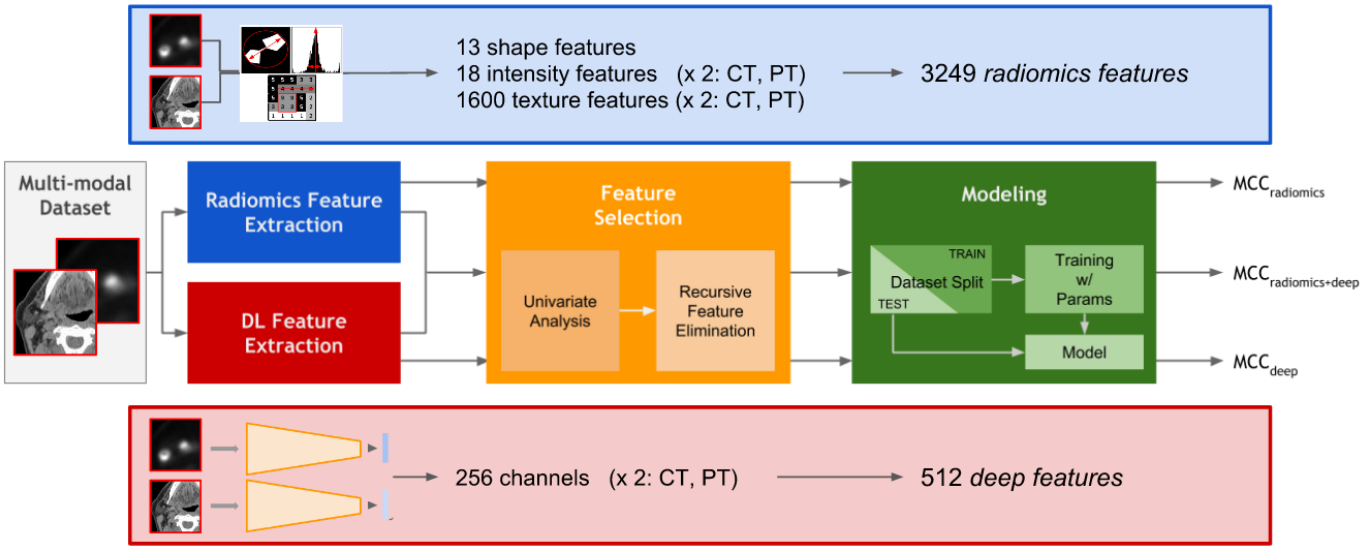


Fig. 3. The RADLER pipeline on CT/PET data: predictive models from the integration of radiomics and deep features.



with Adam [43] optimizer (batch size: 32, epochs: 500, learning rate 10^{-3}). Data augmentation procedures were used to improve the performance and reduce overfitting: i.e., minimal rotations, translations and Gaussian noise. The transformed images were resized to cubes of $64 \times 64 \times 64$ to better fit the GPU memory size.

iii) *HCR + DLR.* The two types of features are concatenated into an integrative dataset. A more accurate model is expected from two types that should capture different and complementary information from the input images.

Feature selection and Ranking

The feature selection section in RADLER leverages a combination of three methods from `scikit-learn` [44]. Features are standardized after imputation of missing values (*Nan* and *inf*) by mean feature values. The procedure is composed of three main steps:

- Removal of correlated features (UNCORR). Since the same types of radiomics features are extracted several times with different sets of parameters (e.g. voxel size for interpolation), the HCR feature set includes highly correlated features. Thus, the Pearson's correlation matrix is computed, and high correlated features ($\rho > 0.95$) are removed;
- Univariate analysis (UA). An association score (ANOVA F-test) is computed between each feature and the target. Features are ranked based on the association score, keeping the top $1,000$ features;
- The remaining features are ranked based on their predictive power within a Recursive Feature Elimination (RFE) procedure and ordered by decreasing importance.

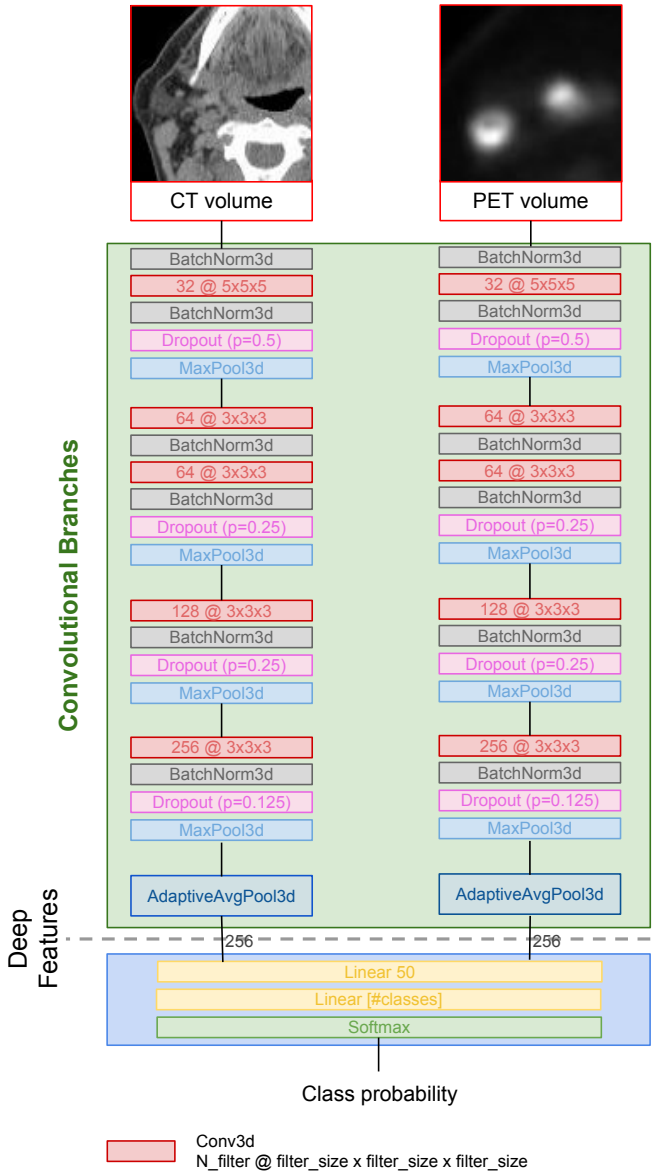
Feature selection and ranking are performed for each feature set type. Notably, no feature from the deep feature sets is removed by the UNCORR step; thus the features automatically

Feature Set	# samples	Number of features		
		Initial	After UNCORR	After UA
HCR	295	3249	968	968
DLR	294	512	512	512
HCR + DLR	293	1480	1427	1000

TABLE III
SUMMARY OF THE THREE FEATURE SETS

created by the network are highly uncorrelated, *i.e.* the information content is maximized. Table III summarizes the three feature sets and the results of the feature selection section.

Fig. 4. Multimodal network architecture for CT/PET scans. The network inputs are pairs of volumes of size $64 \times 64 \times 64$, one for each channel (CT and PET). The total number of output features from the convolutional branches is 512. The final output of the network is the probability of LR occurrence.



Classification within a Data Analysis Protocol framework

A linear Support Vector Machine (LSVM) model is trained on the three feature sets within a Data Analysis Protocol (DAP) framework. The DAP was derived from a bioinformatic machine learning procedure developed by the MAQC consortium to grant reproducibility of predictive biomarkers from microarrays and next-generation sequencing platforms, thus in a massive data context [45]–[47]. The dataset is split beforehand into training and test sets (two cohorts for each split, see Table I): the training set is used to develop the model and the test set is used only to asses the predictive performance. The Matthews Correlation Coefficient (MCC) [48]–[50] is used as the evaluation metric.

A grid of parameters is created from the values of the LSVM regularization parameter $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ and the increasing number of features $nf \in \{0.1\%, 0.2\%, 0.5\%, 1\%, 2\%, 5\%, 10\%, 20\%, 50\%, 100\%\}$. For each parameter point, the training set is randomly split into 5 folds, which are cyclically used to train and validate the model. The optimal parameters are selected by maximizing the predictive results on the validation set. This procedure is repeated 10 times, thus obtaining 50 predictive scores for each parameter point, which are averaged and used to select the best parameter set. Finally, the optimal predictive model is trained on the whole training set using the best parameters and evaluated on the test set.

IV. RESULTS

In order to obtain the Deep Learning network for the LR task, the architecture was first trained to classify the T-stage, with $\mathrm{MCC} = 0.863$ on the training set (one-shot) and $\mathrm{MCC} = 0.279$ on the test set. As this network is only used to initialize the parameters to predict the LR, this result was not validated within the DAP procedure. To qualitatively investigate the embedding resulting in the T-stage network, we considered the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction method [51]. The UMAP projection for the T-stage data of the Deep Features extracted from the PET images is displayed in Fig. 2 (left); the deep learning model transforms the input images into a representation of the T-stage severity, which can be qualitatively represented as a trajectory in the projection plan (see Fig. 2, right).

We transferred the weights of the convolutional branches to a new network and trained the linear layers in the final block to predict LR. These branches are used to generate the deep features (DLR feature set). The trained network obtains $\mathrm{MCC}_{\mathrm{train}} = 0.367$ and $\mathrm{MCC}_{\mathrm{test}} = 0.245$, with a small overfitting.

We compared the performance of the LSVM model trained within the DAP on the different feature sets (see Table IV) and against the reference study [38] (see Table V).

Notably, on the test dataset we improve the original results for all feature sets and metrics, except for the sensitivity achieved using the deep features (0.53 vs 0.56: Table V). In particular, the best results on the test dataset are achieved with

	Sensitivity	Specificity	Accuracy	MCC	# HCR features	# DLR features
HCR	0.553 (0.487 - 0.623)	0.932 (0.916 - 0.946)	0.878 (0.861 - 0.893)	0.498 (0.435 - 0.558)	48 CT:19, PET:28, GTV:1	
DLR	0.387 (0.344 - 0.428)	0.866 (0.850 - 0.882)	0.799 (0.784 - 0.813)	0.244 (0.202 - 0.288)		51 CT:22, PET:29
HCR+DLR	0.805 (0.749 - 0.853)	0.991 (0.988 - 0.995)	0.965 (0.956 - 0.972)	0.848 (0.806 - 0.884)	261 CT:138, PET:118, GTV:5	239 CT:108, PET:131

TABLE IV

CROSS-VALIDATION PERFORMANCE FOR THE DIFFERENT FEATURE SETS ON THE TRAINING SET, USING THE DAP PROCEDURE. METRICS ARE EXPRESSED AS MEDIAN VALUES WITH 95% BOOTSTRAPPED CONFIDENCE INTERVALS. THE LAST TWO COLUMNS REPORT THE NUMBER OF FEATURES PER CLASS AND MODALITY.

	Sensitivity	Specificity	Accuracy	MCC
Vallières CT-PET	0.56	0.67	0.65	Not reported
HCR	0.94	0.95	0.95	0.832
DLR	0.53	0.97	0.9	0.57
HCR+DLR	0.67	0.91	0.94	0.748

TABLE V

PERFORMANCE FOR THE DIFFERENT FEATURE SETS ON THE TEST SET, COMPARED TO REFERENCE RESULTS ("VALLIÈRES CT-PET").

the radiomics features (Sensitivity: 0.94; Specificity: 0.95; Accuracy: 0.95).

However, an underfitting effect is observed for both the HCR and DLR feature sets, with higher accuracy on the test than on the training set. The overall best performance is achieved with the HCR+DLR feature set, i.e. $\mathrm{MCC}_{\mathrm{DAP}} = 0.848$.

Considering the DLR set, we note that both the linear layers of the DL network and the LSVM model have the same sets of features as input, and so their performance can be compared. Notably, the performance of the linear layers ($\mathrm{MCC} = 0.245$) on the test set are within the confidence intervals of the LSVM model ($\mathrm{MCC} = 0.244$).

We also investigated whether considering both CT and PET is effectively useful, *i.e.* if the two modalities contribute with complementary information (see Table VI). Please note that the single modality feature sets (CT and PET) are obtained from the corresponding multimodality features sets after the feature selection and ranking step, by selecting the features computed from the respective images.

Results reported in Table VI show that the underfitting issues, *i.e.* $\mathrm{MCC}_{\mathrm{test}} \gg \mathrm{MCC}_{\mathrm{train}}$, mostly affect the PET-only modality, thus confirming the intrinsic difficulties of quantitative interpretation of PET images. This is also an open problem also in the clinical practice, in particular for head and neck pathologies [52], [53]. In fact, despite the applied conversion to SUV, technical differences between PET scanners, and non-linear effects associated to GTV segmentation as well as other patient-dependent parameters are not taken into account [41], [42], [54]. Further, we observe that the best MCC is achieved by the HCR+DLR feature set, also resolving the mentioned underfitting issues.

V. DISCUSSION

The RADLER framework introduced in this study aims at the integration of deep and radiomics features for medical image analysis and classification. Its first application in a prognostic task of locoregional recurrence (LR task) of head and neck cancer improves with respect to the state of art [38], both in terms of sensitivity (0.94 vs 0.56) and specificity (0.95 vs 0.67). Moreover, the DAP included in the framework is used to evaluate variability due to resampling and control for selection bias in the model selection phase. As assessed by the DAP, the feature set integrating radiomics and deep features is more effective in predicting LR than only one of the feature types.

The RADLER framework is demonstrated with a DL architecture for CT/PET; in detail, a 3D multimodal CNN is adapted from a 2D solution originally aimed at classifying lung nodules from CT imaging [39]. Secondly, we adopted an internal transfer learning approach, starting from the diagnostic classification of tumour stage. This domain adaptation approach proved useful in dealing with class unbalance and a relatively low number of samples, while achieving good predictive performance, as shown on the HN dataset, with high class unbalance and low number of samples.

This design makes the RADLER pipeline and its DL network potentially effective to model other clinical tasks in which different image modalities (*e.g.*, MRI) and anatomical regions (*e.g.*, lung, brain) are considered. The pipeline still requires the manual annotation of the GTV; however, this task could be tackled by automatic segmentation models, thus moving towards a fully automated pipeline. The development of a multimodal network in the RADLER framework was driven as a first step by the integration of PET and CT images in a clinical context. Further work is needed to confirm the robustness of the approach on different cohorts and hospitals.

This work aimed mainly at investigating a new framework for the integration of radiomics and deep learning; limited effort was focused on tuning the DL model, and we restricted the types of radiomics features to those proposed in the reference paper [38]. A similar combination approach of deep and radiomic features has been applied on a subset of the HN dataset to predict distant metastasis by applying CNNs to CT scans only [40]. In particular, we expect that better

	CT-only		PET-only	
Feature set	MCC train	MCC test	MCC train	MCC test
HRC	0.342 (0.284 - 0.401)	0.328	0.163 (0.106 - 0.222)	0.409
DLR	0.168 (0.113 - 0.223)	0.193	0.158 (0.105 - 0.213)	0.545
HCR+DLR	0.552 (0.5 - 0.602)	0.36	0.351 (0.289 - 0.406)	0.365

TABLE VI
PERFORMANCES ON SINGLE MODALITY FEATURE SETS.

accuracy can be achieved by adopting specific deep learning architectures or considering more complex methods to extract radiomics features, for instance applying Wavelet filters [17].

ACKNOWLEDGMENTS

The authors thank the WebValley2018 Students Team for initial development of the radiomics environment. The project has been motivated by a discussion on CT/PET modeling with M. Farsad and A. Fracchetti. Part of this work has been supported by the Microsoft Azure Research Award "Deep Learning for Precision Medicine", assigned to CF.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. Arindra Adiyoso Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] H.-H. Tseng, L. Wei, S. Cui, Y. Luo, R. K. Ten Haken, and I. El Naqa, "Machine Learning and Imaging Informatics in Oncology," *Oncology*, vol. Nov 23, pp. 1–19, 2018.

[3] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, p. 44, 2019.

[4] R. B. Parikh, Z. Obermeyer, and A. S. Navathe, "Regulation of predictive analytics in medicine," *Science*, vol. 363, no. 6429, pp. 810–812, 2019.

[5] Editorial, "Towards trustable machine learning," *Nature Biomedical Engineering*, vol. 2, pp. 709–710, 2018.

[6] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 598–610, 1996.

[7] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[8] Y. Komeda, H. Handa, T. Watanabe, T. Nomura, M. Kitahashi, T. Sakurai, A. Okamoto, T. Minami, M. Kono, T. Arizumi, M. Takenaka, S. Hagiwara, S. Matsui, N. Nishida, H. Kashida, and M. Kudo, "Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience," *Oncology*, vol. 93, no. Suppl. 1, pp. 30–34, 2017.

[9] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep Learning for Classification of Colorectal Polyps on Whole-Slide Images," *Journal of Pathology Informatics*, vol. 8, p. 30, 2017.

[10] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. D. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[11] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[12] A. Bizzego, N. Bussola, M. Chierici, M. Cristoforetti, M. Francescatto, V. Maggio, G. Jurman, and C. Furlanello, "Evaluating reproducibility of AI algorithms in digital pathology with DAPPER," 2019, PLOS Computational Biology, in press.

[13] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, and A. Madabhushi, "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 2089–2099, 2013.

[14] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. Aerts, "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[15] F. Orlhac, C. Nioche, M. Soussan, and I. Buvat, "Understanding Changes in Tumor Textural Indices in PET: a Comparison Between Visual Assessment and Index Values in Simulated and Patient Data," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 58, no. 3, pp. 387–392, 2017.

[16] C. Davatzikos, S. Rathore, S. Bakas, S. Pati, M. Bergman, R. Kalarot, P. Sridharan, A. Gastounioti, N. Jahani, E. Cohen, H. Akbari, B. Tunc, J. Doshi, D. Parker, M. Hsieh, A. Sotiras, H. Li, Y. Ou, R. K. Doot, M. Bilello, Y. Fan, R. T. Shinohara, P. Yushkevich, R. Verma, and D. Kontos, "Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome," *Journal of Medical Imaging (Bellingham)*, vol. 5, no. 1, p. 011018, 2018.

[17] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[18] W. Sun, B. Zheng, and W. Qian, "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis," *Computers in Biology and Medicine*, vol. 89, pp. 530–539, 2017.

[19] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, "Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma," *Scientific Reports*, vol. 7, no. 1, p. 5467, 2017.

[20] D. Kontos and R. M. Summers, "Radiomics and Deep Learning," *Journal of Medical Imaging*, vol. 4, no. 4, p. 041301, 2018.

[21] H. Arimura, M. Soufi, H. Kamezawa, K. Ninomiya, and M. Yamada, "Radiomics with artificial intelligence for precision medicine in radiation therapy," *Journal of Radiation Research*, vol. 60, no. 1, pp. 150–157, 2019.

[22] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[23] R. Paul, S. H. Hawkins, Y. Balagurunathan, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma," *Tomography*, vol. 2, no. 4, pp. 388–395, 2016.

[24] M. Hatt, F. Tixier, D. Visvikis, and C. Cheze Le Rest, "Radiomics in PET/CT: More Than Meets the Eye?" *Journal of Nuclear Medicine*, vol. 58, no. 3, pp. 365–366, 2017.

[25] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging (Bellingham)*, vol. 5, no. 3, p. 036501, 2018.

[26] C. Vendt, "The Cancer Imaging Archive (TCIA): Maintaining and

Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–57, 2013.

[27] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. W. Yeom, M. Iv, Y. Ou, J. Kalpathy-Cramer, S. Napel, R. Gillies, O. Gevaert, and R. Gatenby, "Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches," *American Journal of Neuroradiology*, vol. 39, no. 2, pp. 208–216, 2018.

[28] G. S. Colafati, I. P. Voicu, C. Carducci, E. Miele, A. Carai, S. Di Loreto, A. Marrazzo, A. Cacchione, V. Cecinati, A. Tornesello, and A. Mastronuzzi, "MRI features as a helpful tool to predict the molecular subgroups of medulloblastoma: state of the art," *Therapeutic Advances in Neurological Disorders*, vol. 11, pp. 1–14, 2018.

[29] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics Ssystem to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[30] Z. Bodalal, S. Trebeschi, and R. Beets-Tan, "Radiomics: a critical step towards integrated healthcare," *Insights into Imaging*, vol. 9, no. 6, p. 911, 2018.

[31] A. Vial, D. Stirling, M. Field, M. Ros, C. Ritz, M. Carolan, L. Holloway, and A. Miller, "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review," *Translational Cancer Research*, vol. 7, no. 3, pp. 803–816, 2018.

[32] B. Huynh, H. Li, and M. L. Giger, "Digital Mammographic Tumor Classification Using Transfer Learning from Deep Convolutional Neural Networks," *Journal of Medical Imaging (Bellingham)*, vol. 2, no. 3, p. 034501, 2016.

[33] L. Fu, J. Ma, Y. Ren, Y. S. Han, and J. Zhao, "Automatic Detection of Lung Nodules: False Positive Reduction Using Convolutional Neural Networks and Handcrafted Features," in *Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis*. SPIE, 2017, p. 101340A.

[34] S. Chen, J. Qin, X. Ji, B. Lei, T. Wang, D. Ni, and J.-Z. Cheng, "Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 3, pp. 802–814, 2017.

[35] Z. Ning, J. Luo, Y. Li, S. Han, Q. Feng, Y. Xu, W. Chen, C. Tao, and Y. Zhang, "Pattern Classification for Gastrointestinal Stromal Tumors by Integration of Radiomics and Deep Convolutional Features," *IEEE Journal of Biomedical and Health Informatics*, vol. May 29, p. [Epub ahead of print], 2018.

[36] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities," *arXiv*, vol. 1808.07954, pp. 1–31, 2018.

[37] J.-E. Bibault, P. Giraud, M. Housset, C. Durdux, J. Taieb, A. Berger, R. Coriat, S. Chaussade, B. Dousset, B. Nordlinger, and A. Burgun, "Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer," *Scientific Report*, vol. 8, no. 1, p. 12611, 2018.

[38] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. W. L. Aerts, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, p. 10117, 2017.

[39] F. Ciompi, K. Chung, S. J. Van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. W. Wille, A. Marchianò, U. Pastorino, M. Prokop, and B. van Ginneken, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Scientific Reports*, vol. 7, p. 46479, 2017.

[40] A. Diamant, A. Chatterjee, M. Vallires, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Scientific Report*, vol. 9, no. 1, p. 2764, 2019.

[41] T. Beyer, J. Czernin, and L. S. Freudenberg, "Variations in clinical PET/CT operations: results of an international survey of active PET/CT users," *Journal of Nuclear Medicine*, vol. 52, no. 2, pp. 303–310, 2011.

[42] F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat, "A postreconstruction harmonization method for multicenter radiomic studies in PET," *Journal of Nuclear Medicine*, vol. 59, no. 8, pp. 1321–1328, 2018.

[43] D. Kingma and J. B. Adam, "Adam: A Method for Stochastic Optimization," in *Proc. 3rd International Conference on Learning Representations (ICLR 2015)*. arXiv:1412.6980, 2014, pp. 1–15.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[45] The MAQC Consortium, "The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnology*, vol. 28, no. 8, pp. 827–838, 2010.

[46] The SEQC/MAQC-III Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium," *Nature Biotechnology*, vol. 32, pp. 903–914, 2014.

[47] L. Shi, R. Kusko, R. D. Wolfinger, B. Haibe-Kains, M. Fischer, S.-A. Sansone, C. E. Mason, C. Furlanello, W. D. Jones, B. Ning, and W. Tong, "The international MAQC Society launches to enhance reproducibility of high-throughput technologies," *Nature Biotechnology*, vol. 35, no. 12, pp. 1127–1128, 2017.

[48] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[49] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[50] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of MCC and CEN error measures in multi-class prediction," *PLOS ONE*, vol. 7, no. 8, p. e41882, 2012.

[51] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[52] B. S. Purohit, A. Ailianou, N. Dulguerov, C. D. Becker, O. Ratib, and M. Becker, "FDG-PET/CT pitfalls in oncological head and neck imaging," *Insights into Imaging*, vol. 5, no. 5, pp. 585–602, 2014.

[53] M. S. Hofman and R. J. Hicks, "How We Read Oncologic FDG PET/CT," *Cancer Imaging*, vol. 16, no. 1, p. 35, 2016.

[54] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Le Rest, and D. Visvikis, "Characterization of PET/CT images using texture analysis: the past, the present any future?" *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 44, no. 1, pp. 151–165, 2017.