3D RPET-NET: DEVELOPMENT OF A 3D PET IMAGING CONVOLUTIONAL NEURAL NETWORK FOR RADIOMICS ANALYSIS AND OUTCOME PREDICTION

A. Amyar [1,2], S. Ruan [1], I. Gardin [1,2], C. Chatelain [1], P. Decazes [1,2], R. Modzelewski [1,2]

[1] LITIS - EA4108 - University of Rouen and INSA of Rouen, Rouen, France
[2] Nuclear Medicine Department, Henri Becquerel Center, Rouen, France

ABSTRACT

Radiomics is now widely used to improve the prediction of treatment response and patient prognosis in oncology. In this work, we propose an end-to-end prediction model based on a 3D convolutional neural network (CNN), called 3D RPET-NET, that extracts 3D image features through four layers. Our model was evaluated for its ability to predict the response to radio-chemotherapy in 97 patients with esophageal cancer from positron emission tomography (PET) images. The accuracy of the model was compared to five other methods proposed in the literature for PET images, based on 2D CNN and random forest algorithms. The role of the volume of interest on the accuracy of 3D RPET-NET was also evaluated using isotropic margins of 1, 2, 3 and 4 cm around the tumor volume. After segmentation of the lesion using a fixed threshold value of 40% of the maximum standard uptake value, the best accuracy of 3D RPET-NET reached 72% and outperformed the other methods tested. We also showed that using an isotropic margin of 2 cm around the tumor volume improved the performances of 3D RPET-NET to reach an accuracy of 75%.

***Index Terms*— Positron Emission Tomography, Machine Learning, Deep learning, Esophageal cancer**

1. INTRODUCTION

Predicting patient response to radio-chemotherapy (RCT) is a very promising field of research in personalized medicine. PET imaging with ^{18}F-FDG, which is a radioactive glucose analog, has mainly been used in radiomics analysis, but other radio-tracers have also been tested [1]. However, the roles of traditional imaging biomarkers such as SUVmax and metabolic tumor volume (MTV) have not been well established in esophageal cancer for therapy response [2]. Other biomarkers such as handcrafted texture features have been proposed [3] that are associated with standard statistics. Recently, radiomics biomarkers with complex statistical classifiers [4] have been proposed in the literature. The concept of radiomics is defined as the extraction of dozens of quantitative features from the image that could be incorporated in predictive models for patient management [5]. Many reports suggest that radiomic features extracted from baseline images can contribute to improving patient prognosis and prediction of treatment response in oncology [6]. Images can be obtained from computed tomography (CT), magnetic resonance imaging (MRI) [7] and positron emission tomography (PET). The visualization of glucose metabolism of tumor cells and other radiotracers provides additional information to that obtained from anatomical imaging (CT or MRI). These so-called radiomic features are assumed to highlight some informative tissue characteristics, such as heterogeneity in glucose metabolic activity, necrosis, etc. Numerous image features have been proposed in the literature [8] based on the shape and size of the lesion, 1st order statistics, textural features, filter and model-based features, potentially leading to hundreds of image characteristics.

Several authors have used machine learning (ML) methods to build models for predicting treatment response or patient survival based on radiomic features, such as random forests (RFs) and support vector machines (SVMs) with or without a feature selection strategy [4] [9]. The main drawback of these approaches is that they require an initial extraction of radiomic features using hand crafted methods, which usually results in a large number of features. However, handcrafted features are affected by some parameters [10] such as noise, reconstruction, etc. and significantly by the contouring methods used.

CNNs have proven to be very powerful tools in computer vision for classifying images from different domains. CNN architectures for medical imaging have been introduced and usually containing fewer convolutional layers because of the small datasets[11]. Recently, a new paradigm in PET radiomic analysis has been proposed based on CNNs for predicting response to therapy [12]. It has been shown that deep learning architectures can outperform traditional ML methods in classification tasks.

CNNs were not fully studied in radiomics, especially in PET imaging. Some papers have investigated baseline PET analysis based on 2 Dimensional (2D) CNN architectures [12] [13], but to our knowledge, there are no studies using 3D-CNN. These first two applications dealt with the prediction of the response to neoadjuvant chemotherapy in esophageal cancer [12] and the classification of mediastinal lymph node

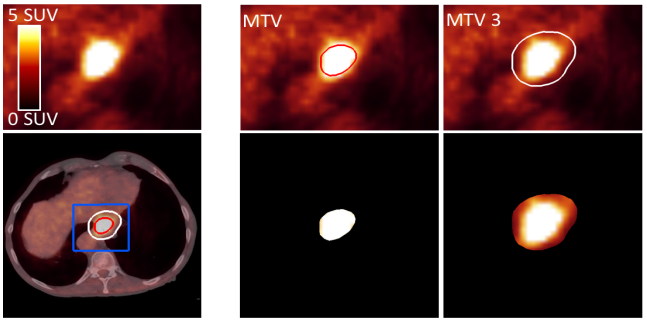

Fig. 1. Columns from left to right: Fused PET/CT slice, zoomed on the esophageal tumor seen on FDG-PET only. MTV (40% SUVmax thresholding) in red and MTV included in the cuboid. MTV3 (MTV + 3 cm isotropic margin) in white and MTV3 included in the cuboid.

metastasis of non-small cell lung cancer (NSCLC) [13].

In [12], Ypsilantis *et al.* proposed to learn a hierarchical representation directly from PET images in 107 patients with esophageal cancer using two CNN architectures. The first one, called 1S-CNN, corresponds to an architecture where the input is one slice. The process is repeated on each slice where the tumor is present. The spatial dependency between slices is not exploited in this architecture. For this reason, a second architecture was proposed where the input of the CNN is composed of 3 adjacent slices, called 3S-CNN. For each exam containing *m* slices, each set of three spatially adjacent slices is taken as input, leading to a total of *m-2* possible combinations. This 3S-CNN better exploits the spatial relationship between slices but is limited to 3 slices. For both architectures, a post processing step is required to predict the response based on a majority vote process. This study has shown the superiority of these two deep learning methods compared to other ML methods, such as RF, SVM, gradient boosting, and logistic regression.

In [13], Wang *et al.* used a centered axial slice and two others that were separated by 4 mm in two image modalities (PET and CT) to obtain a limited number of six slices for each tumor to make a prediction. They compared the performances of their CNN and four other methods including RF, SVM, adaptive boosting, and artificial neural network. The methods were evaluated to discriminate against benign and malignant lymph nodes (1397) in 168 patients. The study showed that there were no significant differences between the CNN and the best classical ML method for classifying mediastinal lymph node metastasis of NSCLC from PET/CT images. Nevertheless, Wang *et al.* concluded that CNNs are more convenient to use because the method does not require an initial feature extraction.

Radiotherapy planning is based on CT by delineating the gross tumor volume (GTV). This GTV can also be segmented using other image modalities, such as MR and PET images. Segmentation of the tumor in PET imaging is usually performed using a fixed threshold value of 40% of the maximum standard uptake value (SUVmax) [1], leading to the biological or metabolic target volume (BTV or MTV). Then, the radiation oncologist adds several margins that account for the non-visible tumor infiltration (CTV: clinical tumor volume) as well as uncertainties in positioning and treatment to obtain the PTV (planning target volume) [14]. The peritumoral part of the tumour is therefore a volume that is not neglected in the treatment. By analogy, taking into account the intratumoral and peritumoral regions in radiomics analysis is likely a strategy that can improve the results. At present, a few studies have tested this hypothesis in other modalities [15] [16] but never with PET imaging.

Our goal was to develop a new 3D-CNN architecture, called 3D RPET-NET, to predict the response to treatment by learning from FDG-PET images of the tumor. Considering our small dataset, a four-layer 3D-CNN was proposed. Our study used a database of baseline FDG PET images of 97 patients treated by radio-chemotherapy (RCT) for esophageal cancer. The optimal hyperparameters of 3D RPET-NET were found and the influence of the learning volume (intratumoral volume with different peritumoral volumes) was investigated. The performances of the model were compared to 1S-CNN and 3S-CNN [12], as well as to three RF methods [4] considered as state-of-the-art radiomics classifiers.

2. METHOD

2.1. Data

In this study, 97 patients with one lesion that was histologically proven to be locally advanced esophageal cancer and eligible for RCT were included. All procedures performed in this study were conducted according to the principles expressed in the Declaration of Helsinki. The study was approved as a retrospective study by the Henri Becquerel Center Institutional Review Board (number 1506B). All patient information was de-identified and anonymized prior to analysis.

Patients underwent FDG PET with a CT before treatment (baseline PET), at the initial stage. They were treated by RCT, corresponding to an uninterrupted radiation therapy in the form of external radiation delivered by a 2-field technique of 2 Gy per fraction per day, 5 sessions per week, for a total of 50 Gy, as well as chemotherapy including platinum and 5-fluorouracil.

The PET/CT data were acquired on a Biograph® Sensation 16 Hi-Rez device (Siemens Medical Solutions, IL, USA). This device does not provide point spread function (PSF) modeling or time-of-flight (TOF) technology. Patients were required to fast for at least 6 hours before imaging. A total of 5 MBq/kg of FDG was injected after 20 min of rest. Sixty minutes later ($\pm$10 min), 6 to 8 bed positions per patient

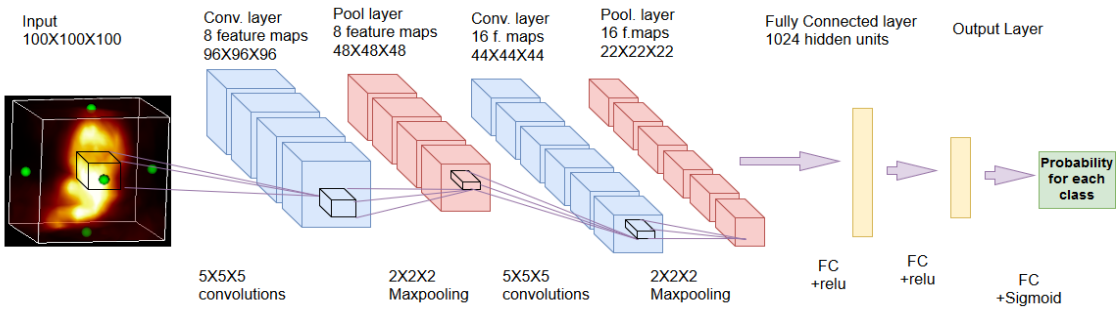

Fig. 2. 3D RPET-NET architecture composed by two 3D convolutional layers followed by 3D pooling layers and two dense layers.

were acquired using a whole-body protocol (3 min per bed position). The PET images were reconstructed using Fourier rebinding (FORE) and attenuation-weighted ordered subset expectation maximization algorithms (AW-OSEM with 4 iterations and 8 subsets). The images were corrected for random coincidences, scatter, and attenuation. Finally, the FDG-PET images were smoothed with a Gaussian filter (full width at half maximum (FWHM) = 5 mm). The reconstructed image voxel size was $4.06 \times 4.06 \times 2.0$ mm^3.

For the determination of treatment response, the response assessment included clinical examination, CT, FDG-PET, and esophagoscopy with biopsies performed 1 month after the end of treatment. Patients were classified as showing a clinically complete response (CR, 56 patients) to RCT if no residual tumor was detected on the endoscopy (negative biopsies) and if no locoregional or distant disease were identified on CT or PET evaluation. Patients were classified as showing a non-complete response (NCR, 41 patients) if a residual tumor or locoregional or distant disease was detected or if death occurred.

2.2. Image preprocessing

Tumor images were spatially normalized by re-sampling all the dataset to an isotropic resolution of $2 \times 2 \times 2$ mm^3 using the k-nearest neighbor interpolation algorithm.

The metabolic tumor volume (MTV) was segmented by a physician who manually defined a cuboid volume around the lesion and used a fixed threshold value of 40% of the maximum standard uptake value (SUVmax) in the cuboid. To study the influence of the volume of interest on the performances of 3D RPET-NET, several isotropic margins of 1, 2, 3 and 4 cm around MTV were also applied, leading to defining MTV1 to MTV4. In Fig. 1, an example of a PET/CT slice with two volumes of interest (MTV and MTV3) is shown.

Tumor gray level intensities were normalized to absolute SUV level between [0 30] and translated between [0 1] to be used in CNN architectures. The volumes of interest were included into a 3D empty cuboid of standard width, length and height of 100^3 voxels to learn tumoral radiomic features (see Fig. 1 and input part in Fig. 2).

2.3. 3D RPET-NET architecture

We have developed a new CNN architecture based on two 3D convolutional layers and two fully connected layers, as shown in Fig. 2. Each convolutional layer, denoted $C^{(m)}$, consists of $F^{(m)}$ feature maps, where m is the layer number (1 or 2). For the first layer, $C^{(1)}$, each feature map is obtained by convolving the volume of interest with a weight matrix $W_i^{(1)}$ to which a bias term $b_i^{(1)}$ is added, where i is the feature map number. Then, the output is processed by a non linear function f(x) called the activation function, where x is the input to a neuron, such as:

$$c_i^{(1)} = f(b_i^{(1)} + W_i^{(1)} * x) \qquad \text{with } i = 1, ..., F^{(1)}. \quad (1)$$

Each element of a feature map, $c_i^{(1)}$, is obtained by convolving the input x with a 3D kernel. A large receptive field tends to better preseve the relationship between slices and the local 3D tumor information than a small one ((5 x 5 x 5) *vs.* (3 x 3 x 3)). The $F^{(1)}$ weight matrices (one matrix per feature map) are learned by observing different positions of the input, leading to the extraction of the description of features. Thus, the weight parameters are shared for all tumor input sites, so that the layer has an equivariance property and is invariant to the input tumor transformations (such as translation and rotation). It also results in a sparse weight, which means that the kernel can detect small, but meaningful features, as shown in Fig. 3. For instance, it can be seen that some kernels are learning the tumor shape (e.g, [(1,1),(1,3),(2,4),(2,6)..etc.]), while others tend to focus on features within the tumor (e.g, [(1,2),(1,5),(2,2),(2,3)..etc.]).

Then, the output of this first convolutional layer is followed by a 3D pooling layer, to reduce the dimensionality of feature maps. The max-pooling operator is used as a stage detector to report the maximum value within each cuboid of size (2 x 2 x 2) for all feature maps. The purpose of this operation is to down sample the feature maps by a factor of 2 along each direction (width, high, length) and to better generalize learning by selecting approximately invariant features. This

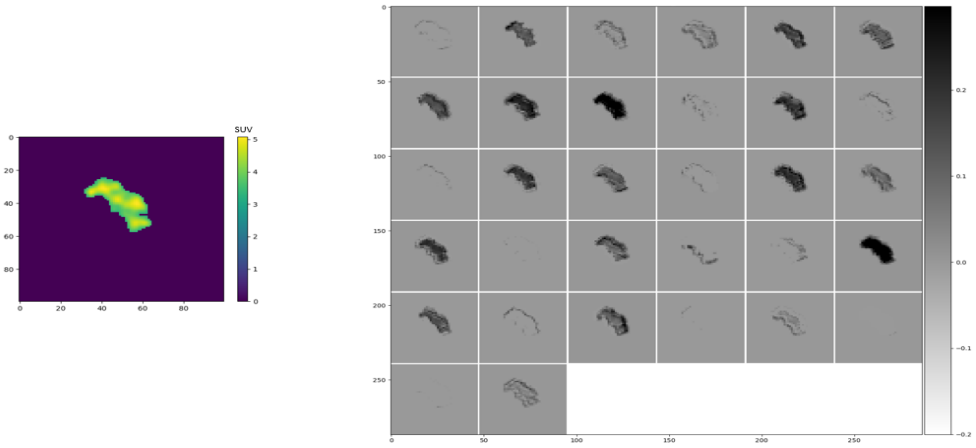

Fig. 3. Visualization of a 2D slice of a segmented tumor and the resulting 32 feature maps in the second convolutional layer of the 1S-CNN architecture.

invariance to local translation is very important in radiomics because tumors do not have a particular direction. The resulting feature maps are denoted $P^{(m)}$.

To extract high-level features from the low-level ones obtained in the initial layer, a second convolutional layer is added, followed by a pooling layer. This convolutional layer learns from the pooled feature maps of the first layer (see Fig. 2).

The parameters of the CNN consist of all the convolutional weights W, and the weight matrix Wh, denoted by θ. They are learned by minimizing the binary cross-entropy function:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)] \quad (2)$$

which is a special case of the multinomial cross-entropy loss function for m = 2 :

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} log(\hat{y}_{ij}) \quad (3)$$

where n is the number of patients, y is the tumor label (binary, 1 if the patient responded to treatment, 0 otherwise) and $\hat{y}_{ij} \in (0,1): \sum_j \hat{y}_{ij}=1 \ \forall i,j$ is the prediction response of a patient.

In our experiments, the adaptive gradient algorithm optimizer (AdaDelta) was used with mini batches. At each update of weights using the AdaDelta algorithm, only one mini batch of training data was used, which is changed for each gradient calculation. Our CNN also incorporated L2 normalization of the weights and a dropout regularization of 50% to prevent the model from overfitting.

To find the best 3D RPET-NET, called 3D RPET-NETBest, the best hyperparameters were found. The hyperparameters optimized include the number of 3D feature maps (we tested from 8 to 64 feature maps), the number of neurons (128, 256, 512, 1024, 2048 and 4096), as well as different receptive field sizes ($3 \times 3 \times 3$, $5 \times 5 \times 5$) and different sizes of mini-batches (2, 4, 8 and 16). Several (4) expressions of f(x), the activation function, were also evaluated (relu, elu, selu and tanh). Several numbers of 3D convolutional layers and 3D pooling layers (2 to 5) and fully connected layers (2, 3, 4 and 5) were evaluated.

2.4. Implementation

The implementation of 3D RPET-NET was conducted using the Keras library which is built on top of Theano and Tensorflow. We took advantage of graphical processing units (GPUs) to accelerate the algorithm. The CNNs training was performed on an NVIDIA Tesla 80 with 12 GB of memory.

3. EXPERIMENTATION

Three experiments were performed to evaluate 3D RPET-NET.

Experiment 1: The first experiment consisted of tuning the optimal hyperparameters to find 3D RPET-NETBest based on MTV. Optimizing the hyperparameters was performed entirely on the training dataset.

Experiment 2: The second experiment consisted of comparing our architecture with 2 other CNN methods proposed in the literature : 1S-CNN and 3S-CNN [12]. The same tuning process of 3D RPET-NET was performed to find the best 1S-CNN and 3S-CNN hyperparameters. This experiment was performed on test data.

The results were also compared to 3 RF-based methods: one without any feature selection strategy, called RF, and 2 other RF methods proposed in the literature based on a feature selection strategy, called GARF (genetic algorithm based on

	Method	VOI	Acc	Sens	Spec	AUC
Experiment 1						
	3D RPET-NETBest	**MTV**	**0.83±0.04**	**0.91±0.06**	**0.73±0.16**	**0.81±0.06**
	3D RPET-NET1	MTV	0.80±0.06	0.93±0.05	0.61±0.15	0.77 ±0.06
	3D RPET-NET2	MTV	0.76±0.04	0.87±0.12	0.62±0.19	0.75±0.05
Experiment 2						
	3D RPET-NET	**MTV**	**0.72±0.08**	**0.79±0.17**	**0.62±0.21**	**0.70±0.04**
	1S-CNN	MTV	0.69±0.06	0.79±0.15	0.57±0.24	0.65±0.08
	3S-CNN	MTV	0.67±0.08	0.73±0.19	0.60±0.20	0.67±0.08
	GARF	MTV	0.68±0.08	0.80±0.11	0.46±0.09	0.62±0.04
	FIC	MTV	0.65±0.07	0.78±0.21	0.46±0.38	0.61 ±0.16
	RF	MTV	0.65±0.04	0.65±0.18	0.53 ±0.18	0.59±0.04
Experiment 3						
	3D RPET-NET	MTV1	0.73±0.04	0.76±0.07	0.69±0.1	0.72±0.04
	GARF	MTV1	0.70±0.08	0.74±0.07	0.54±0.07	0.62±0.02
	FIC	MT1V	0.62±0.10	0.58±0.18	0.64±0.12	0.59 ±0.04
	RF	MTV1	0.62±0.09	0.62±0.08	0.61 ±0.07	0.59±0.03
	3D RPET-NET	**MTV2**	**0.75±0.03**	**0.76±0.45**	**0.74±0.15**	**0.74±0.02**
	GARF	MTV2	0.71±0.09	0.73±0.11	0.54±0.09	0.63±0.04
	FIC	MTV2	0.58±0.01	0.58±0.25	0.57±0.18	0.54 ±0.07
	RF	MTV2	0.62±0.11	0.56±0.20	0.65 ±0.12	0.59±0.05
	3D RPET-NET	MTV3	0.72±0.09	0.71±0.09	0.74±0.14	0.72±0.09
	GARF	MTV3	0.66±0.07	0.68±0.19	0.57±0.12	0.63±0.04
	FIC	MTV3	0.61±0.11	0.63±0.17	0.58±0.16	0.59 ±0.04
	RF	MTV3	0.62±0.14	0.66±0.17	0.55 ±0.20	0.59±0.04
	3D RPET-NET	MTV4	0.63±0.09	0.77±0.10	0.46±0.21	0.61±0.11
	GARF	MTV4	0.65±0.09	0.73±0.14	0.52±0.16	0.62±0.02
	FIC	MTV4	0.59±0.08	0.54±0.14	0.63±0.08	0.56 ±0.04
	RF	MTV4	0.60±0.13	0.66±0.12	0.56±0.05	0.58±0.04

Table 1. Classification results: Each result corresponds to the average of five independent experiments and the standard deviation, using the training dataset (Experiment 1) or the test dataset (Experiment 2 and 3).

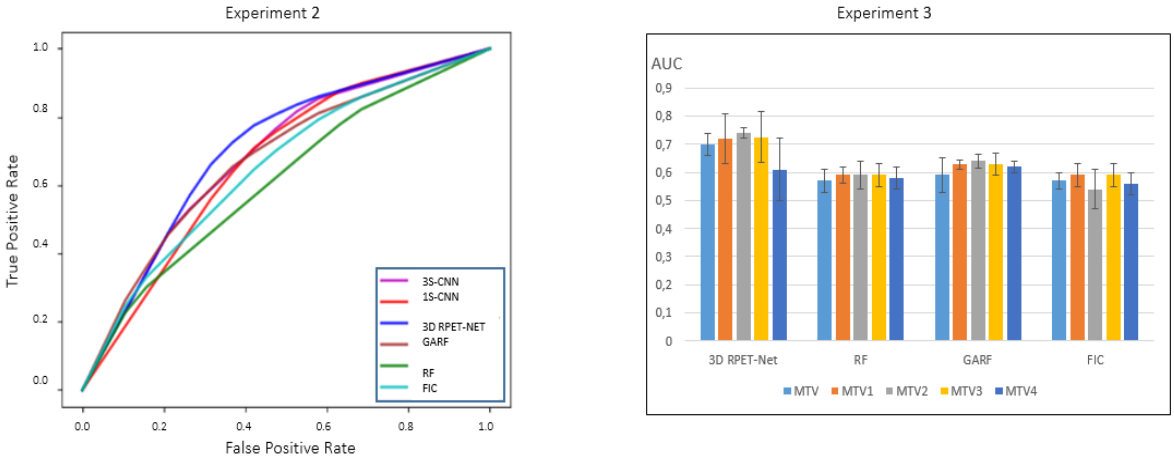

Fig. 4. a. On the left: ROC curve comparing the 6 classifiers (RF, GARF, FIC, 1S-CNN, 3S-CNN and 3D RPET-NET) with the best parameters on MTV. b. Right: Comparison of the four classifiers on different VOIs (MTVs). Error bars correspond to standard deviation.

random forest) and FIC (features important coefficient) methods. For the details of the methods refer to [4]. Briefly, 45 image features were extracted from PET images corresponding to first-order statistics (18), one feature of the lesion form, and textural features (26). Five hundred decision trees were built leading to the creation of the random forest classifiers.

Experiment 3: The third experiment consisted of assessing the influence of the volume of interest on the performances of 3D RPET-NETBest, RF, GARF and FIC using MTV, MTV1, MTV2, MTV3 and MTV4.

4. VALIDATION METHODOLOGY

For method validation, cross-validation (CV) was performed. We split the data into 2 groups to train and test the machine learning methods for each fold. One group was used for training the models (77 patients) and one group for testing (20 patients). Furthermore, for the CNN, the training samples were split into a dataset of 2 groups, a train set (55 patients) and a validation set (20 patients), and a grid search was conducted to derive the optimal hyperparameters based on the validation set. For a fair comparison, different machine learning methods were trained and tested with the same fold, i.e, trained with the same training sets and tested with the same test sets. To keep the same ratio between the two classes CR and NCR, for each fold, the training set contained 44 CR patients and 33 NCR patients, and the testing set contained 12 CR and 8 NCR.

The performances of the methods were evaluated for each cross-validation, including sensitivity (Sens), specificity (Spec), accuracy (Acc), and area under the receiver operating characteristic (ROC) curve (AUC). For each curve, the definition of the thresholds was determined using the method proposed by Fawcett [17], and the optimal cut-off point was defined using Youden's index.

A comparison between different methods was mainly performed based on the AUC values. Due to the 5-fold CV, 5 groups of performance values were calculated for each method; therefore, paired hypothesis tests of 5 samples were performed. The p values were calculated using Student t-test. To correct for multiple comparisons, we additionally adjusted p-values by the false-discovery-rate (FDR) procedure according to Benjamini-Hochberg [18]. The null hypotheses were rejected at the level of $p < 0.05$ after correction.

5. RESULTS

The main results from the 3 experiments (accuracy, sensitivity, specificity, AUC of ROC curves) are shown in table 1.

Experiment 1: As shown in Fig.2, the best accuracy Acc=0.72 and AUC=0.70 were achieved by two 3D convolutions layers and two 3D pooling layers, followed by two fully connected layers with the following hyperparameters for the first 3D convolutional layer: 8 3D feature maps with a filter size of $5 \times 5 \times 5$ and a relu activation function. This operation is followed by 3D Max-pooling of size $2 \times 2 \times 2$. The second 3D convolutional layer corresponds to 16 3D feature maps of $5 \times 5 \times 5$ convolutions, followed again by a $2 \times 2 \times 2$ 3D pooling layer. Then, the last two layers are composed of fully connected layers of 1024 hidden neurons and finally 2 neurons for both classes.

In Experiment 1, the results of two other models show interesting performances, with no significant difference from 3D RPET-NETBest. 3D RPET-NETBest and 3D RPET-NET1 differ by the activation function (relu *vs.* elu). 3D RPET-NETBest and 3D RPET-NET2 differ by the activation function (relu *vs.* elu) and the kernel size ((5 x 5 x 5) *vs.* (3 x 3 x 3)).

Experiment 2: In Table 1, the best results found with 1S-CNN, 3S-CNN, RF, GARF and FIC are shown. TheROC curves of Experiment 2 are presented in Fig. 4.a.

The best results were found with 3D PET-NETBest. 1S-CNN, seems to have lower performances (Acc=0.67$\pm$0.06, AUC=0.67$\pm$0.06), but the 1S-CNN ROC curve was not statistically significantly different from 3D RPET-NETBest (p=0.53) and 3S-CNN (p=0.48) ROC curves. For the RF classifiers, the best results were found with the GARF algorithm. The GARF ROC curve was not statistically significantly different from 1S-CNN (p=0.10) and 3S-CNN (p=0.058) ROC curves, while the 3D RPET-NETBest ROC curve had better results than the GARF ROC curve (p=0.028).

Experiment 3: The results of Experiment 3 are given in Table 1 and the comparisons of different AUC in Fig. 4.b. When studying the influence of the volume of interest, the best performances of 3D RPET-NETBest were obtained with MTV2 (Acc=0.75 and AUC=0.74). The performances of the 3D RPET-NETBest tend to increase from no margin to a magin of 2 cm, and then decrease with higher margins (MTV3 and MTV4). Only 3D RPET-NETBest performances on MTV2 were statistically significantly better than those on MTV4 (p=0.04). The same trend is observed with RF classifiers.

6. DISCUSSION

We have developed an end-to-end 3D convolutional neural network (3D PET-NET) based on PET images. We have also evaluated 5 other methods from the literature [12] [4]. For each CNN, the search for the best architecture was achieved using a validation procedure to tune hyperparameters, such as the number of feature maps and the size of filters. For comparison, a standard radiomic analysis was conducted using 3 random forest classifiers: without feature selection (RF), with a selection strategy based on a genetic algorithm (GARF) and based on the features importance coefficients (FIC).

Apart from the numerous advantages of CNNs (avoiding handcrafted feature design and feature selection, state-of-the-art performance on almost any computer vision task, end-to-

end trainable models), it is now well known that convolutional architectures build high level representations of the input signals. They typically extract low level features such as textures of edge detectors in the low layers and accumulate these informations to form higher level features in the last layers. Low level features are generally rather generic and can be exploited through transfer learning [19]. Higher level features are more domain-specific and depend upon the application. A neural network is often considered as a black box, but CNN layers provide interpretability through the feature maps that highlight the activation of each kernel within the input signal. Therefore, we think that CNN features are likely to be related to classical hancrafted radiomic features (see Fig. 3).

PET imaging suffers from low resolution and high noise, leading to challenges in PET radiomics [10]. However, neural networks provide a robust mechanism to avoid encoding the noise in the data such as 'early-Stopping' and 'dropout' which provide better generalization [20].

Unlike Ypsilantis et al. in [12] who claimed that the use of a 3D ROI as direct input of the CNN is infeasible because every tumor has a different shape and size, we show that englobing the tumor into a 3D cuboid of standard width, length and height allows the benefit of the spatial relationship between slices using a large 3D receptive field to be realized. Our assumption is that a neural network architecture able to capture patterns of FDG uptake that occur within the whole lesion may detect imaging features that are more relevant to predict treatment response than each slice individually or 3 adjacent slices. Under this assumption, we propose an architecture that initially fuses the spatial information across intra-slices images. 3D RPET-NETBest is composed of only 2 convolutional layers. A higher number of convolutional layers were tested, without conclusive results. The small number of patients in our database (without artificial data augmentation) is a limiting factor not only for the development of a deeper network but also for radiomic analysis in general. Indeed, the current trend is in favour of the use of a network with an increasing number of convolutional layers (very deep neural network). This is only possible on large image databases (e.g., ImageNet [21], containing now more than 14 million images, 30 high level categories and 20K subcategories) that are not currently available in medical imaging. It is possible to artificially increase the number of data. However since, learning takes place on a tumor inside a black box, this solution leads to overfitting.

To ensure a fair comparison between the different methods, the database was divided into 3 groups of 57 patients for the training, 20 for the validation and 20 for the test before any operation. Every CNN and RF classifier used the same folds to obtain an exact comparison between methodologies.

There are several segmentation methods available for PET imaging. Many automatic frameworks have been proposed during the last decade [22], but few of them are used/available in clinical routine. The simple threshold is still mainly used but with different values depending on pathologies [23]. A segmentation of the MTV can be accurately performed with a 40% threshold value because esophageal cancer can be considered as a massive non moving tumor [24] and it has been proven that this segmentation is highly correlated with a manual segmentation [5].

We have shown that isotropic dilation of MTV tends to increase the performances of RPET-NET 3D. When the margin around the MTV is too large (>2 cm) the network performances decrease. When the MTV is increased by a margin which is too large, the volume of interest can include parts of metabolically active organs that are likely to interfere with the CNN analysis. Our results suggest that between 3 cm and 4 cm of the peritumoral volume, the relevent information to predict treatment response decreases, is responsible for a drop in the model's performance. Adding a peritumoral volume to the radiomic analysis has already been tested in MRI [25] but never in PET imaging. These initial results must be confirmed on other types of cancer. Moreover, the influences of the initial volume of interest and the segmentation methods require further study.

7. CONCLUSION

The analysis of PET tumor images with a3D CNN architecture (3D-RPET-NET) shows very promising results in the prediction of treatment response in esophageal cancer. 3D-RPET-NET outperformed 2D CNN architectures, as well as the traditional radiomics approach (handcrafted feature extraction with RF classifiers) . Moreover, since the CNN does not take hand-crafted features as input, it eliminated the need for feature selection, making the entire process much more convenient and less prone to user bias. In addition, we have shown that the best volume to be used for PET radiomic prediction is the metabolic tumor volume with an isotopic margin of 2 cm. This peritumoral region seems to contain information that is potentially relevant to building better prediction algorithms since currently approaches are based only on the quantification of the intratumoral region alone.

These results need to be confirmed on a larger database. The integration of clinical data in the model is an interesting and challenging perspective for such architectures that could improved the performances of the classifier.

8. REFERENCES

[1] L Lu et al., "Robustness of radiomic features in [11 c] choline and [18 f] fdg pet/ct imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization," *Molecular Imaging and Biology*, vol. 18, no. 6, pp. 935–945, 2016.

[2] R Kwee et al., "Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with

use of 18f fdg pet: a systematic review," *Radiology*, vol. 254, no. 3, pp. 707–717, 2010.

[3] F Tixier et al., "Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer," *Journal of Nuclear Medicine*, vol. 52, no. 3, pp. 369, 2011.

[4] P Desbordes et al., "Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier," *Computerized Medical Imaging and Graphics*, vol. 60, pp. 42–49, 2017.

[5] P Lambin et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[6] M Avanzo et al., "Beyond imaging: The promise of radiomics," *Physica Medica: European Journal of Medical Physics*, vol. 38, pp. 122–139, 2017.

[7] T Nishioka et al., "Image fusion between 18fdg-pet and mri/ct for radiotherapy planning of oropharyngeal and nasopharyngeal carcinomas," *International Journal of Radiation Oncology Biology Physics*, vol. 53, no. 4, pp. 1051–1057, 2002.

[8] M Sollini et al., "Pet radiomics in nsclc: state of the art and a proposal for harmonization of methodology," *Scientific reports*, vol. 7, no. 1, pp. 358, 2017.

[9] S Leger et al., "A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling," *Scientific reports*, vol. 7, no. 1, pp. 13206, 2017.

[10] M Hatt et al., "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.

[11] M Frid-Adar et al., "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *arXiv preprint arXiv:1803.01229*, 2018.

[12] P Ypsilantis et al., "Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks," *PLoS One*, vol. 10, no. 9, pp. e0137036, sep 2015.

[13] H Wang et al., "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images," *EJNMMI Res.*, vol. 7, no. 1, pp. 11, 2017.

[14] B Dubray et al., "Nuclear imaging and target volumes for radiotherapy," *MEDECINE NUCLEAIRE-IMAGERIE FONCTIONNELLE ET METABOLIQUE*, vol. 37, no. 5, pp. 198–202, 2013.

[15] N Braman et al., "Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri," *Breast Cancer Research*, vol. 19, no. 1, pp. 57, 2017.

[16] Y Zhou et al., "A radiomics approach with cnn for shear-wave elastography breast tumor classification," *IEEE Transactions on Biomedical Engineering*, 2018.

[17] T Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[18] Y Benjamini et al., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[19] S Belharbi et al., "Spotting l3 slice in ct scans using deep convolutional network and transfer learning," *Computers in biology and medicine*, vol. 87, pp. 95–103, 2017.

[20] N Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] O Russakovsky et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] B Foster et al., "A review on segmentation of positron emission tomography images," *Computers in biology and medicine*, vol. 50, pp. 76–96, 2014.

[23] A Dewalle-Vignion et al., "Evaluation of pet volume segmentation methods: comparisons with expert manual delineations," *Nuclear medicine communications*, vol. 33, no. 1, pp. 34–42, 2012.

[24] W Kawakamiet al., "The use of positron emission tomography/computed tomography imaging in radiation therapy: a phantom study for setting internal target volume of biological target volume," *Radiation Oncology*, vol. 10, no. 1, pp. 1, 2015.

[25] N Bramanet al., "Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI.," *Breast cancer research : BCR*, vol. 19, no. 1, pp. 57, 2017.