PLOS | ONE

RESEARCH ARTICLE

Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks

Petros-Pavlos Ypsilantis[1], Musib Siddique[2], Hyon-Mok Sohn[2], Andrew Davies[2], Gary Cook[2], Vicky Goh[2], Giovanni Montana[1]*

1 Department of Biomedical Engineering, King's College London, London SE1 7EH, United Kingdom, **2** Department of Cancer Imaging, King's College London, London SE1 7EH, United Kingdom

* giovanni.montana@kcl.ac.uk

OPEN ACCESS

Citation: Ypsilantis P-P, Siddique M, Sohn H-M, Davies A, Cook G, Goh V, et al. (2015) Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PLoS ONE 10(9): e0137036. doi:10.1371/journal.pone.0137036

Editor: Ruby John Anto, Rajiv Gandhi Centre for Biotechnology, INDIA

Received: June 16, 2015

Accepted: August 11, 2015

Published: September 10, 2015

Data Availability Statement: Relevant data are available on Figshare: http://dx.doi.org/10.6084/m9.figshare.1513829.

Funding: Support was received from the NIHR Biomedical Research Centre of Guys & St Thomas' NHS Trust in partnership with Kings College London, and King's College London and UCL Comprehensive Cancer Imaging Centre funded by the CRUK and EPSRC in association with the MRC and DoH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Imaging of cancer with ^{18}F-fluorodeoxyglucose positron emission tomography (^{18}F-FDG PET) has become a standard component of diagnosis and staging in oncology, and is becoming more important as a quantitative monitor of individual response to therapy. In this article we investigate the challenging problem of predicting a patient's response to neoadjuvant chemotherapy from a single ^{18}F-FDG PET scan taken prior to treatment. We take a "radiomics" approach whereby a large amount of quantitative features is automatically extracted from pretherapy PET images in order to build a comprehensive quantification of the tumor phenotype. While the dominant methodology relies on hand-crafted texture features, we explore the potential of automatically learning low- to high-level features directly from PET scans. We report on a study that compares the performance of two competing radiomics strategies: an approach based on state-of-the-art statistical classifiers using over 100 quantitative imaging descriptors, including texture features as well as standardized uptake values, and a convolutional neural network, 3S-CNN, trained directly from PET scans by taking sets of adjacent intra-tumor slices. Our experimental results, based on a sample of 107 patients with esophageal cancer, provide initial evidence that convolutional neural networks have the potential to extract PET imaging representations that are highly predictive of response to therapy. On this dataset, 3S-CNN achieves an average 80.7% sensitivity and 81.6% specificity in predicting non-responders, and outperforms other competing predictive models.

Introduction

Neoadjuvant chemotherapy (NC) for cancer treatment is often given as a first step before the definitive surgery of a tumor, in order to facilitate surgical resection and improve the likelihood of a R0 resection [1], i.e. where there is a clear surgical margin on the pathological specimen. NC has been associated with improved survival after surgery for patients who respond to the therapy, and is considered the standard of care in some cancers [2, 3]. On the other hand, for

Competing Interests: The authors have declared that no competing interests exist.

patients who do not respond to NC, the prognosis after therapy is generally worse compared to primarily surgical approach alone [4]. When NC is not effective, it has also the disadvantage of exposing patients to unnecessary toxicity and may lead to adverse events while substantially increasing the associated health care costs and delaying definitive treatment. Identifying novel, non-invasive approaches for pretherapy prediction of NC response therefore holds the promise to stratify patients for preoperative therapy and has the potential to substantially improve the clinical outcome for certain patient populations or at least individualize treatment regimes.

Positron emission tomography (PET) is a nuclear medicine imaging technique based on the measurement of gamma rays resulting from positron emission using radiolabelled tracer molecules. These radiotracers allow biological processes to be measured and whole body images to be obtained which demonstrate sites of radiotracer accumulation. One of the most common radiotracers in use today is ^{18}F-fluorodeoxyglucose (^{18}F-FDG), a radiolabelled sugar (glucose analog) molecule. Imaging with ^{18}F-FDG PET is used to determine sites of abnormal glucose metabolism and can be used to localize and characterise many types of tumor non-invasively. There is extensive evidence in the literature indicating the importance of ^{18}F-FDG PET imaging in accurately characterizing disease, as well as determining stage and sites of recurrent disease in many cancer types [5]. For these indications, functional imaging with PET provides unique information which is not generally available from other standard medical imaging modalities such as CT and MRI.

Despite early indications that ^{18}F-FDG PET imaging may also be a viable approach for predicting NC response using pre-treatment imaging [6], only a handful of quantitative measurements or biomarkers carrying predictive power have been found to be clinically useful. Some evidence has been reported that the amount of FDG uptake on pretreatment scans, as measured by tumor metabolic concentrations known as Standardized Uptake Values (SUVs) may carry predictive power [7–9]. The rationale for this approach is that the elevated FDG uptake in malignant cells is hypothesized to be associated with biologically relevant features, such as perfusion, cell proliferation, tumor viability, aggressiveness, and hypoxia [10–12], which are predictive of resistance to chemotherapy. However, SUV measurements are significantly affected by the initial ^{18}F-FDG uptake kinetics and radiotracer distribution, which depend on the initial radiotracer injected activity as well as on the time elapsed between the tracer injection and the image acquisition. These factors can complicate the interpretation of SUV measurements due to their significant intra- and inter-observer variability [13]. For these reasons, and the fact that response prediction is not sufficiently accurate to use in the clinic, SUV measurements so far have been proved to be most useful in studies investigating the role of PET imaging to track ^{18}F-FDG uptake changes over the course of an existing treatment [14, 15] rather than in predicting response from a single PET scan prior to therapy.

A particularly promising research direction that could potentially overcome the above limitations consists of the high-throughput extraction of large amounts of imaging features that can be made in direct relationship with clinical endpoints of interest. Radiomics [16, 17], an emerging field of research concerned with this objective, can potentially have a large clinical impact, since imaging is routinely used in clinical practice world-wide. In PET imaging, there has been increasing interest in identifying imaging features that characterize the spatial distribution and heterogeneity of ^{18}F-FDG uptake patterns within a tumour by image analysis [18, 19]. This heterogeneity is hypothesized to originate in a number of physiological factors such as tumor metabolism, necrosis, cellularity, and angiogenesis, amongst others, and variability in these factors has been associated with more aggressive cancer behaviour, poorer response to treatment and worse prognosis [10–12]. The dominant methodologies for obtaining quantitative descriptors of spatial heterogeneity rely on texture analysis [20]. Such techniques encompass a broad range of mathematical descriptors that can be used to evaluate the spatial

variation of voxel intensities both within a single PET slice as well as between adjacent slices, thus providing measures of intra-lesional heterogeneity. In contrast to SUVs, these descriptors provide a more accurate spatial characterization of FDG uptake patterns, and can potentially capture more signal. Whilst recent studies have started to explore the benefits of texture analysis for predicting response to NC therapy [21–25], drawing definite conclusions is difficult as each study relies upon different definitions of texture and deploys different predictive models. Another limiting factor characterising existing studies has been the small sample sizes, typically ranging from 10 to 50 patients.

The objective of this work is two-fold. First, we set out to explore whether a machine learning algorithm would be able to infer a predictive representation of a cancer's metabolic profile, as captured by ^{18}F-FDG PET imaging, in a larger patient population. In very recent years, biologically-inspired convolutional neural networks (CNN) have shown the ability to learn hierarchically-organised, low- to high-level features from raw images [26, 27], and yield state-of-the-art performance in the classification of both natural and medical images [28–31]. To investigate this question, we propose a neural network to harness the predictive power of spatially-varying ^{18}F-FDG PET uptake patterns. The proposed architecture, 3S-CNN (three-slices convolutional neural network), produces features that are representative of metabolic activity in cancer, before therapy, and we expect that this method would ultimately distinguish responders to non-responders. Our second objective is to compare the performance of 3S-CNN with competing predictive algorithms where the quantitative tumour phenotypes are represented by over 100 "hand designed" texture features, capturing patterns in both two- and three-dimensions, as well as SUV summaries. Whereas previous studies have each reported on the performance of a very specific approach, generally combining a handful of selected texture features with a single statistical classifier, here we aim at a more comprehensive empirical characterization of a large battery of quantitative descriptors and predictive models. In this respect, our results set a comparative benchmark for future radiomics developments in this area.

Material and methods

Oesophageal cancer data

For this study we obtained a dataset consisting of $n = 107$ patients (83 males, 24 females) with newly diagnosed esophageal cancer at a tertiary referral centre, Guys and St Thomas NHS Trust (GSTT). The study was approved by the Westminster ethics committee, and all patient information was anonymized prior to analysis. The age at time of diagnosis ranged from 32 to 80 with an average of 62±25 years. All patients underwent pre-treatment whole-body ^{18}F-FDG PET/CT for staging. For each patient, the primary tumor was positively identified on axial ^{18}F-FDG PET images by an experienced nuclear medicine physician.

Bespoke software was developed for the automatic delineation of the tumor ROIs. This was achieved by applying a 40% slice-wise maximum intensity threshold to exclude voxels with less than 40% of the activity in the voxel of maximum intensity within the same axial slice. Using this approach we were able to accurately delineate the regions of high ^{18}F-FDG uptake across all tumors in the study. Each pixel corresponded to a voxel size of 4.7×4.7 *mm* with 3.27 *mm* slice thickness, and the size of the slices in the dataset varies from 13×16 to 93×79 pixels. Also, as can be observed in Fig 1, there is a wide inter individual variability in the number of axial slices which were extracted from the 3D tumor volumes. The number of axial slices per tumor varied from 4 to 32. 86 tumors were adenocarcinoma, 20 were squamous cell carcinoma, and one was undefined. Nearly half of the patients had a moderately differentiated tumour (58). 70 patients had a T3 primary lesion, 80 had N1 lymph node metastasis, and 1 patient had distant metastasis. All patients were treated with neoadjuvant chemotherapy, and 38 responded

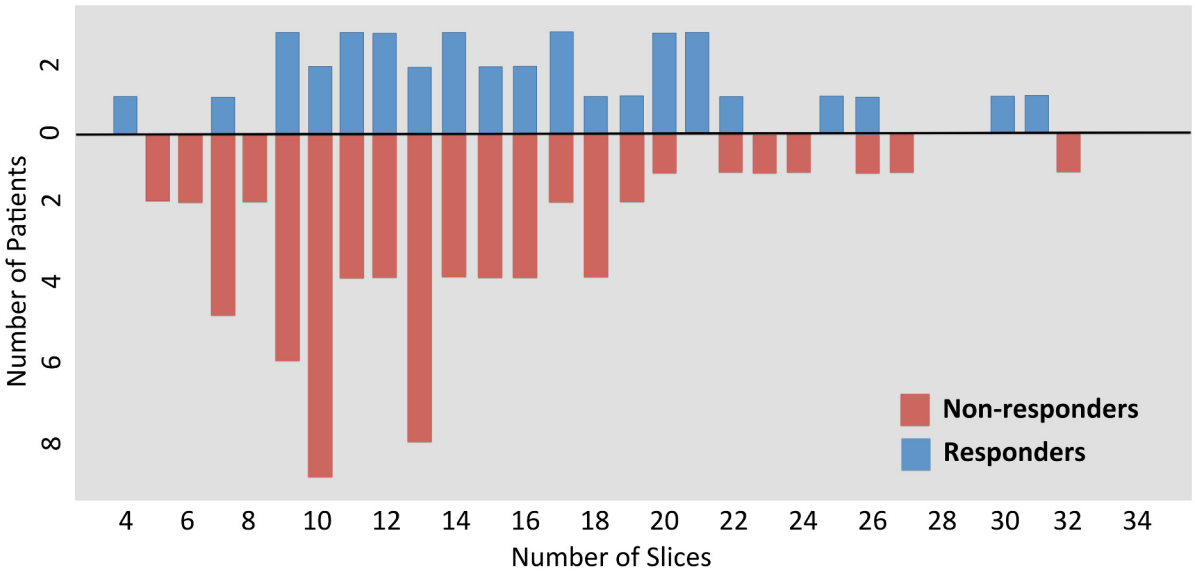

Fig 1. Distributions of axial ^{18}F-FDG PET intra slices extracted from the 3D tumor volume of non-responders and responders.

doi:10.1371/journal.pone.0137036.g001

to treatment. The response was assessed using a pathological Mandard tumor regression grade [32]. For the purpose of this study, we grouped the patients into two distinct pathologic groups: the non-responders group, which includes subjects whose tumor showed no regressive changes (Grades 4 and 5), and the responders group, which includes all those cases in which regressive changes were noted (Grades 1, 2 and 3) as originally proposed in [32]. Among the responders, 26 (68.4%) had TNM clinical stage III; among the non-responders, 20 (29.0%) had clinical TNM stage II. The overall survival (OS) period was defined as the time in days between the PET scan and the date of death. Responders to therapy had a median OS of 972.5 days compared with 714 days for non-responders. Fig 2 illustrates the overall survival rates, which were found to be significantly different by a Kaplan-Meier analysis (p-value = 0.00045).

Texture analysis

Texture analysis refers to a variety of mathematical methods used to provide information about the spatial arrangement of voxel intensities within a volume neighborhood containing the tumor [21, 25]. In our study, we used texture analysis to characterize the 3D uptake heterogeneity in a tumor that is captured by the PET scans. We employed two broad classes of texture feature extraction techniques, statistical- and model-based. The statistical approach consisted of quantifying some aspects of the spatial distribution of voxel values by taking into account local features at each point in the image, and extracting a set of statistics from the distributions of these features.

The statistics-based approach relies on first-, second- and higher-order statistics. First-order statistics were calculated from the original voxel intensity values without taking into consideration the relationship of each voxel with its neighbors. This class encompassed measures of central tendency (including mean, median, mode, percentiles, quartiles), variability (including range, interquartile range, variance, standard deviation, coefficient of variation, skewness, and kurtosis), first order energy, and entropy. Second order statistics consist of co-occurence measurements between two pixels calculated using both Gray-Level Co-occurrence Matrices

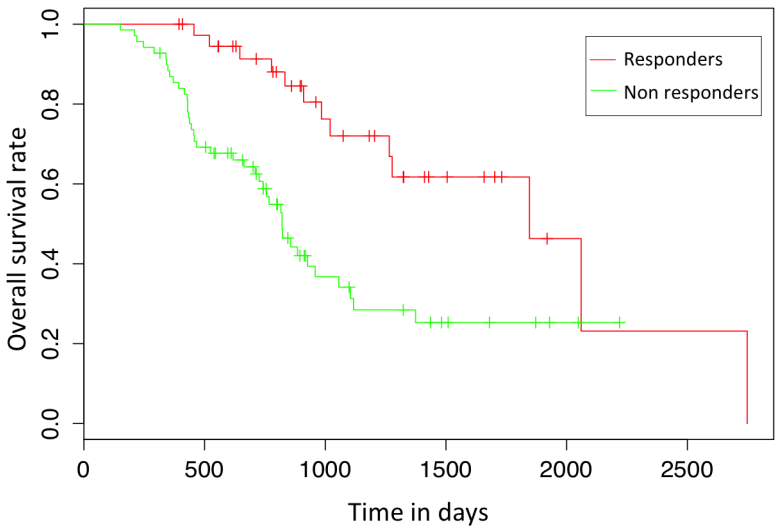

Fig 2. Kaplan-Meier plot showing the survival rates of responders and non-responders.

doi:10.1371/journal.pone.0137036.g002

(GLCMs) [33] and Gray-Level Difference Matrices (GLDMs) [34]. GLCMs determine how many times a voxel with a given intensity occurs jointly with another voxel having a different intensity, whereas GLDMs are based on absolute differences between pairs of voxel intensities. Higher order statistics capture properties of three or more voxel values occurring at specific locations relative to each other, and are extracted from Gray-Level Run Length Matrices (GLRLMs) [35], Gray-Level Size Zone matrices (GLSZMs) [36] and Neighborhood Gray-Tone Difference Matrices (NGTDMs) [37]. Both GLRLMs and GLSZMs analyze texture in a specific gray-level run and zone, respectively. Gray-level run is the length of consecutive voxels having the same intensity in a preset direction in the image, whereas the zone is a cluster of consecutive voxels having the same intensity. A GLRL matrix is a two-dimensional matrix in which each element $p(i, j|\alpha)$ gives the total number of occurrences of runs of length j at gray level i in a specific direction α. Following the same principle, the entries of a GLSZ matrix are the number of clusters of size s of gray-level i, where the size of a cluster is defined as the number of consecutive pixels with gray-level i. Finally, NGTDM are column matrices that describe the differences between each voxel and its neighboring voxels in adjacent image planes, and are thought to closely resemble the human experience of the image [37]. In NGTDM, the i^{th} entry is a summation of the differences between all pixels with gray-tone i and the average value of their surrounding neighbors.

Model-based approaches used mathematical models such as fractal analysis to represent texture information. Fractal analysis is a form of geometric pattern recognition that evaluates the self-similarity and roughness of a surface at different levels [38, 39]. Such evaluation in the context of PET imaging can quantify the ^{18}F-FDG uptake heterogeneity of a tumor volume [40, 41]. A fractal is defined as a set for which Hausdorff-Besicovich dimension is strictly greater than the topological dimension [42]. The fractal dimension (FD) is the defining property in the study of texture analysis. The fractal dimension of each voxel of the ^{18}F-FDG uptake is calculated inside a moving window centred on the voxels by using a differential box-counting method [43]. A summary of the texture matrices and the extracted texture features we have calculated is provided in Table 1. In total we considered 85 texture features, on top of which we then added 18 statistical summaries of SUV measurements (including minimum, maximum,

Table 1. Summary of second and high order texture features extracted from texture analysis.

Texture Matrices	Texture Features
Gray Level Co-occurence	Energy, Autocorrelation,
	Cluster Prominence, Cluster shade,
	Contrast, Correlation, Difference Entropy,
	Difference Variance, Dissimilarity,
	Entropy, Homogeneity, Difference Moment
	Information Measure Cor.1/Cor.2,
	Sum Average, Sum Entropy, Sum Variance
	Inverse Difference Moment Normalized,
	Inv. Difference. Normalized, Max. Probability,
Gray Level Run Length	Short Run Emphasis, Long Run Emphasis,
	Short Run Low/High Gray Level Intensity,
	Long Run High/Low Gray Level Intensity,
	Run Length Non-uniformity,Run Percentage,
	Intensity Variability, Run Length Variability
	High/Low Gray Level Run Emphasis,
Gray Level Size Zone	Short/Long Zone Emphasis, Zone Percentage,
	Short Zone Low/High Emphasis,
	Long Zone High/Low Emphasis,
	Intensity Non-uniformity,
	Zone Length Non-uniformity,
	Low/High Intensity Zone Emphasis,
	Intensity Variability, Size zone Variability
Gray Level Difference	Mean, Entropy, Variance, Contrast
Fractal Based Features	Mean, Standard Deviation,
	Hurst Exponent, Lacunarity
Neigh. Gray Tone Difference	Coarseness, Contrast, Busyness,
	Texture Strength, Complexity

doi:10.1371/journal.pone.0137036.t001

mean). By including all the above features, for each tumor volume i, we end up extracting a quantitative descriptor of the ROI consisting of a 103-dimensonal vector denoted $\mathbf{t}_i = (t_{i,1}, \cdots, t_{i,103})$, for $i = 1, 2, \ldots, n$.

The individual feature vectors $\mathbf{t}_i$ obtained for the pretherapy PET scans were used as predictors to train and test state-of-the art machine learning techniques for the prediction of NC therapy response. The objective was to minimize the classification error on unseen test data. We considered four different statistical classifiers: logistic regression (LR) [44], gradient boosting (GB) [45], random forests (RF) [46], and support vector machines (SVM) [47]. Logistic regression is a common linear method for multi-variable modeling of binary outcomes. Both GB and RF embrace the notion of ensemble learning, whereby an entire collection of learning algorithms is deployed in order to obtain superior predictive performance. More explicitly, GB algorithm builds an ensemble of regression trees in a stage-wise fashion, where each one is trained with respect to the error of the whole ensemble learnt so far. On the other hand, the RF algorithm builds an ensemble of de-correlated classification trees, where each one is trained on a random subsample of the training dataset and then combines the trained classification trees by averaging their probabilistic prediction. Finally, the kernel-based SVM algorithm

discriminates between responders and non-responders using hyperplanes that maximize the margin between the two classes in a non-lineal feature space. The key idea of kernels is to project the input explanatory variables of our dataset into high dimension hyperplanes where the discrimination between responders and non responders is improved.

Convolutional neural networks

A Convolutional Neural Network (CNN) is a special feed-forward neural network for learning a hierarchical representation of imaging data [26, 27] and then using these representations for imaging recognition tasks. In traditional classifiers like LR, SVM, GB and RF, there is a need for preprocessing the images and to extract texture features relevant to a specific task. The limitation of these classifiers originates from the fact that the performance is highly dependent on the design of the texture features, thus requiring prior knowledge for a specific task and expertise in hand-engineering the necessary features. By contrast, CNN operates directly on raw images and attempts to automatically extracts highly expressive imaging features relevant to a specific task at hand.

Compared to standard neural networks, the individual neurons in a CNN are tiled in such a way that they respond to overlapping portions of the input image. The main architectural components of CNN are the convolutional and subsampling layers. The neurons of the convolutional layer receive information from only a subset of the inputs, called receptive field. As a result, each neuron learns to detect features from a local region of the input image. This allows us to capture the local substructure and preserve the topology of the input image. In addition to local connectivity, a convolutional layer also imposes groups of neurons, whose receptive fields are located in different places of the input image, to share exactly the same weight values. The outputs of these groups of neurons are called feature maps. The technique of the weight sharing reduces the number of free parameters, thus increasing the generalization ability of the network [48]. A convolutional layer is composed of several feature maps, so that a rich variety of imaging features can be extracted. The convolutional layers are then followed by subsampling layers whose purpose is to reduce the dimensionality of the convolutional responses by selecting superior invariant imaging features. In an attempt to achieve a distributed and more abstract representation of the input image, multiple convolutional-subsampling layers are stacked on top of one another, thus delivering a deep architecture of multiple non-linear transformations. Each layer generates a representation of the image based on the feature-detecting role of the neurons. By stacking layers of feature-detecting neurons, a CNN is able to infer highly expressive representations carrying predictive power for imaging recognition tasks [28].

In our application the object to be classified is a ROI representing the tumor, which has a three-dimensional shape. Using ROIs as direct input of the CNN is infeasible due to the fact that every tumor has a different shape and size. To address this issue, we initially embedded all individual ROIs into 3D cuboids of standard width and length, and height varying according to the number of slices of each ROI. Specifically, each 2D intra-tumor slice was embedded into a larger and squared background of standard size 100 × 100 pixels, which was sufficiently large to include all the observed tumors (see Fig 3). For a given tumor i having m_i slices, we denote each enlarged slice as $\mathbf{x}_{i,j}$ where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$. We denote the entire standardized volume containing the tumor for patient i as $\mathbf{x}_i$. Our assumption is that a neural network architecture able to capture patterns of FDG uptake that occur within each 2D slice as well as across multiple adjacent slices may detect salient imaging features that are important for predicting chemotherapy response. Under this assumption, we propose an architecture that initially fuses the spatial information across adjacent intra slices. For a given standardized volume $\mathbf{x}_i$ containing m_i slices, we build all possible sets of three adjacent slices, which we denote

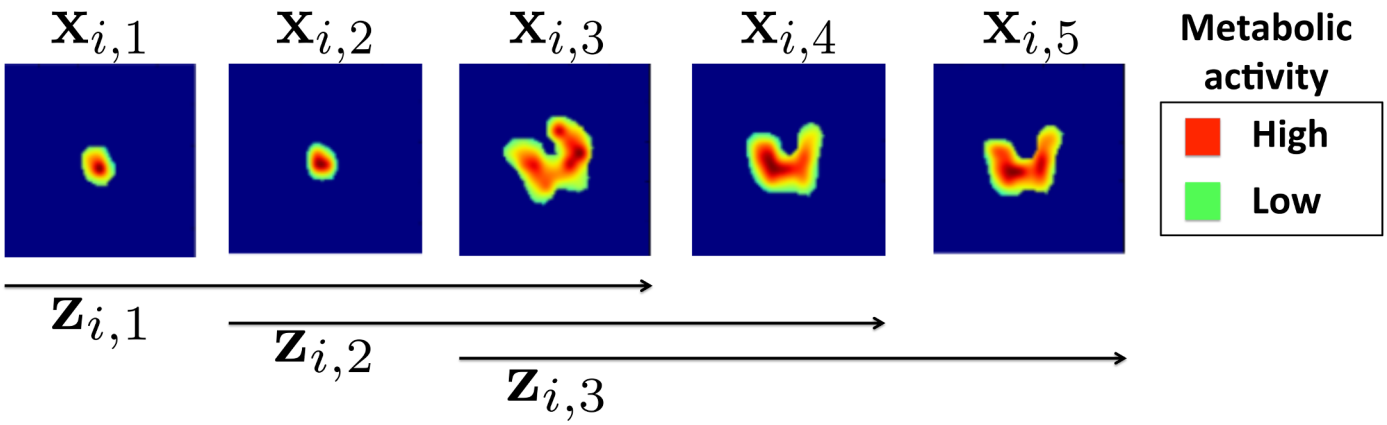

Fig 3. ^{18}F-FDG PET ROIs of a specific tumor *i* after segmentation embedded into larger square background of standard size of 100 × 100 pixels. Each enlarged slice is denoted by $\mathbf{x}_{i,j}$ and each set of three spatially adjacent enlarged slides is denoted by $\mathbf{z}_{i,k}$, where *j* and *k* represent the slices and triplets of the specific tumor *i*. In this example only 3 triplets, from the 5 available slices can be formed, so *k* = 1,2,3.

doi:10.1371/journal.pone.0137036.g003

as $\mathbf{z}_{i,k} = \{\mathbf{x}_{i,k}, \mathbf{x}_{i,k+1}, \mathbf{x}_{i,k+2}\}$ where $k = 1, \ldots, m_i - 2$. This process is illustrated in Fig 3. Each triplet $\mathbf{z}_{i,k}$ was then treated as a three-channel input for the CNN. Associated with each triplet $\mathbf{z}_{i,k}$ there is a corresponding binary label, $y_{i,\,k}$, indicating whether the patient has responded ($y_{i,\,k} = 1$) or not responded ($y_{i,\,k} = 0$) to therapy.

The first convolutional layer of the CNN, denoted $\mathbf{U}^{(1)}$, consists of $R^{(1)}$ feature maps. Each feature map is obtained by convolving all slices within a triplet $\mathbf{z}_{i,k}$ with a weight matrix $\mathbf{k}_{p,j}^{(1)}$, to which we then add a bias term $b_j^{(1)}$. The output is then passed through an hyperbolic tangent function $f(\cdot)$, i.e.

$$\mathbf{u}_j^{(1)} = f\left(\sum_{p=0}^{2} \mathbf{k}_{p,j}^{(1)} * \mathbf{x}_{i,k+p} + b_j^{(1)}\right) \quad j = 1, \ldots, R^{(1)}.$$

Each element of a feature map $\mathbf{u}_j^{(1)}$ in the first convolutional layer enclose information from a local 3D tumor uptake region. The $R^{(1)}$ weight matrices, one for each feature map, are learned in order to build a library of low-level features which are extracted by inspecting various locations of the input triplet. Within each PET slice, the same weight matrix is convolved with the entire slice. This results in the weight being shared by many overlapping squared sub-windows of the slice, and also in sparse connections between the input units and the hidden units in the first layer.

Once each feature has been learned, its exact location within the triplet becomes less important, as long as its approximate position relative to other features is preserved. The convolutional layer is then followed by a subsampling layer which reduces the dimensionality of each feature map. This is achieved by retaining only the maximum value within each non-overlapping sub-region of size (2 × 2) for each feature map. This max-pooling operation is carried out in order to down-sample each feature map by a factor of 2 along each direction and improve generalization performance by selecting invariant features [49]. The max-pooling layer has the same number of output and input feature maps and does not require any additional parameters.

In order to extract higher-level features from the low-level features obtained in the initial layers, additional convolutional layers are added, which are always followed by a pooling layer.

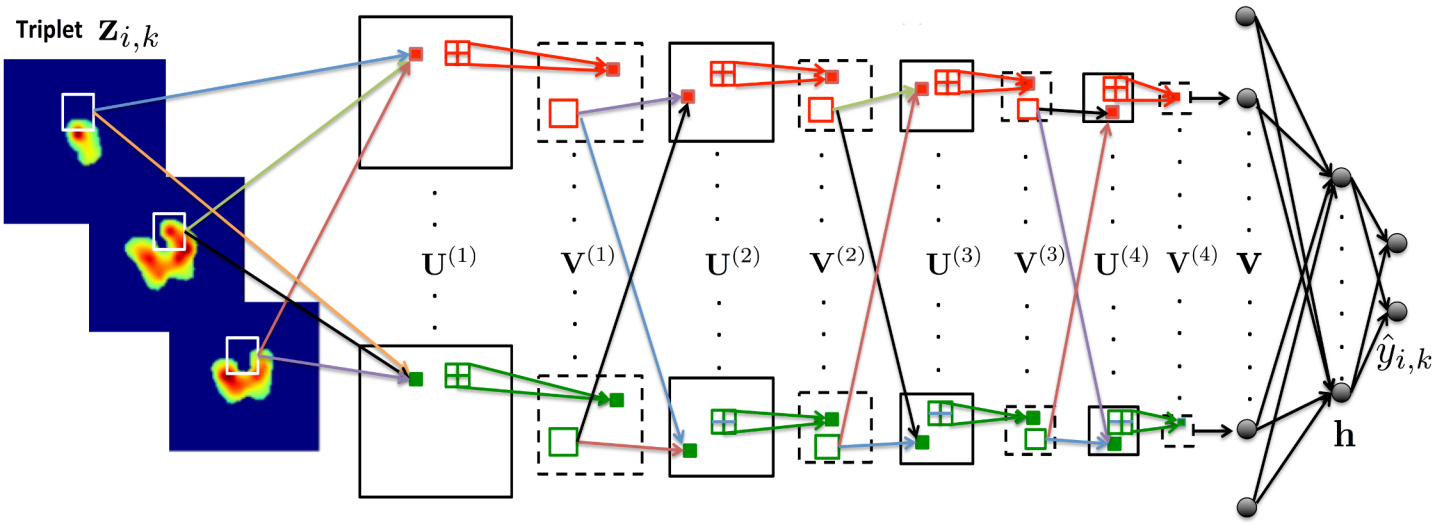

Fig 4. CNN architecture for fusion of 3 adjacent ^{18}F-FDG PET intra slices into a vector v. The CNN architecture is composed from 4 convolutional and 4 max-pooling layers denoted by $\{\mathbf{U}^{(i)}\}_{i=1}^{4}$ and $\{\mathbf{V}^{(i)}\}_{i=1}^{4}$. In the first convolutional layer $\mathbf{U}^{(1)}$, different coloured arrows represent the usage of different learnable weight matrices for convolving each PET slice in the triplet. Colored dotted rectangles in the feature maps represent elements of the feature maps that enclose local spatial information of the previous layer in the architecture. In the Max-pooling layers 2 × 2 element windows represent non-overlapped grids from which we choose the maximum element to downsample the feature maps.

doi:10.1371/journal.pone.0137036.g004

Each additional convolutional layer spans all the pooled feature maps obtained at the previous layer (see Fig 4). For instance, each feature map in the second layer is obtained as

$$\mathbf{u}_j^{(2)} = f\left(\sum_{p=1}^{R^{(1)}} \mathbf{k}_{p,j}^{(2)} * \mathbf{v}_p^{(1)} + b_j^{(2)}\right) \quad j = 1, \ldots, R^{(2)}.$$

The complete architecture contains four convolutional/max-pooling layers (see Fig 4). The resulting set of max-pooled feature maps $\mathbf{v}_j^{(4)}, j = 1, 2, \ldots, R^{(4)}$ enclose the entire spatial local information as well as the rich hierarchical representation of the input triplet $\mathbf{z}_{i,k}$. Each feature map $\mathbf{v}_j^{(4)}$ is then flattened out and all the elements are collected into a single vector $\mathbf{v}$ of dimension R. These units provide the input for a fully connected (FC) hidden layer, $\mathbf{h}$ consisting of H units. The activation of the j^{th} unit of the FC layer is given by

$$h_j = f\left(\sum_{k=1}^{R} M_{kj} v_k + b_j\right), \quad j = 1, \ldots, H.$$

All weights are collected into a matrix $\mathbf{M}$. The probability that each $\mathbf{z}_{i,k}$ is assigned to class 1 (responder) is given by the soft-max function

$$p(\hat{y}_{i,k} = 1 | \mathbf{h}; \theta_1, \theta_2) = \frac{\exp\{\theta_1 \mathbf{h}\}}{\sum_{j=1}^{2} \exp\{\theta_j \mathbf{h}\}},$$

where the vectors θ_1 and θ_2 are the columns of the softmax matrix $\Theta_{H\times 2}$ According to this rule, a triplet $\mathbf{z}_{i,k}$ is assigned to class 1 when $p(\hat{y}_{i,k} = 1) > 0.5$. In case of ties, we take the prediction as being wrong.

In order to predict whether an unseen tumor volume $\mathbf{x}_i$ respond or not to the therapy, we use a majority vote rule based on the estimated prediction probabilities for all triplets extracted from the tumor. We predict that $\mathbf{x}_i$ is a responder when

$$\frac{1}{m_i - 2}\sum_{k=1}^{m_i-2}\mathrm{I}(\hat{y}_{i,k} = 1) > 0.5,$$

where $\mathrm{I}(\cdot)$ is an indicator function that is 1 when $\hat{y}_{i,k} = 1$ and otherwise is zero.

The parameters of the CNN consists of all the convolutional weights $\mathbf{k}_{j,R_l}^{(l)}$, the weight matrix $\mathbf{M}$ and the soft-max parameter matrix $\Theta_{H\times 2}$. These unknown parameters, denoted by $\mathbf{W}$, are learned by minimizing the negative log-likelihood function,

$$\ell(\mathbf{W}) = -\sum_{i=1}^{n}\log\left(p(\hat{y}_{i,k} = y_{i,k}|\mathbf{z}_{i,k}, \mathbf{W})\right). \quad (1)$$

In our experiments we used a stochastic gradient-descent algorithm with mini batches (MSGD). MSGD is a variant of the gradient descent algorithm commonly used to train neural networks on large datasets [50]. At each update of the weights in the SGD algorithm, instead of considering all the training data to compute the gradient of the loss function ℓ, only one small batch of training data at a time is used. We also take advantage from the parallelization of the MSGD algorithm in order to accelerate the training of our CNN on GPU cards. Our code is based on Theano, a Python library that compiles symbolical expressions into C/CUDA code for deployment on both CPUs and GPUs [51].

Comparative analysis

We carried out a comparative analysis of different machine learning algorithms for NC therapy response prediction. As well as the 3S-CNN model, which takes sets of intra-tumor triplets as input, we also implemented a simpler 1S-CNN architecture that treats each individual slice $\mathbf{x}_{i,k}$ as an independent sample, and eventually make a decision based on a majority vote rule, exactly as in 3S-CNN. This simpler architecture is added here to study the potential advantages deriving from exploiting inter-slice patterns that capture 3D information as in 3S-CNN. The performance of all comparable algorithms were obtained by averaging the outcome of three independent experiments. Each experiment was conducted using a different combination of training and test sets. In each case, 96 patients were assigned to the training set, and the remaining 11 were utilized for testing.

For the 3S-CNN architecture, each training set consisted of all triplets of adjacent ^{18}F-FDG PET slices extracted from the tumors, and each triplet was treated as a training example. Furthermore, in order to create more training examples and prevent our CNN model from overfitting, the training set was artificially augmented by rotating each triplet by $\kappa\cdot 60°$, κ = 1,2,3,4,5. Overall, we created a balanced training dataset of 5316 FDG-PET triplets for both responders and non-responders to therapy. Training of the 1S-CNN architecture was done in a similar way, whereby each slice within a tumour contributed a training example. For fair comparisons between CNN models, the same augmentation strategy was always used. For the purpose of tuning and optimising all the hyperparameters, which include the number of layers, weight matrices size, and number of feature maps in each layer and learning rate, 30% of the training dataset was used as validation set.

The predictive models trained on texture and SUV features are denoted LR, GB, RF, and SVM when using the original feature vectors, and LR-PCA, GB-PCA, RF-PCA and SVM-PCA when using the ten largest principal components extracted from the feature vectors. PCA was

used to reduce the dimensionality of the input by a factor of 10 whilst retaining as much information as possible. For each model we deployed a grid search using 10-fold cross validation to choose the set of hyperparameters. In 10-fold cross-validation, the original training sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimation. For the SVM classifier we tested linear, polynomial and Gaussian kernels, and here report only on the best SVM performance, which was obtained by the polynomial kernel.

The forward stage-wise fashion of the GB allows us to automatically assess the contribution of each variable in the construction of a robust classification rule [52]. In the GB algorithm at each node of each regression tree a specific variable is used to partition the sample of patients associated with that node into subregions. The particular variable chosen is the one that gives maximal estimated improvement in squared error risk over that for a constant fit over the entire sample of patients. In each regression tree, the squared relative importance of this variable is the sum of such squared improvements over all the internal nodes for which it was chosen as the splitting variable. This importance measure can be generalized to the forward stage-wise expansion of regression trees of the GB algorithm by simply averaged over the trees which were induced in the ensemble.

We also examined the performance of a classifier based only on SUVmax measurements. The SUVmax summaries were thresholded by performing a receiver operating characteristic (ROC) analysis. The optimal threshold was identified by means of a grid search using values within the range of the SUVmax measurements extracted from the training dataset. From the ROC curve we chose the threshold associated to the maximum sum of true positive and true negative rates. Then we classified each of the remaining 11 patients as responders if the corresponding SUVmax measurement was below the threshold.

Finally, we explored a potential association between response to treatment and TNM staging and grading, as these two parameters are commonly adopted in clinical practice. TNM stages were divided into two groups, stage II and stage III, and the strength of their association was not found to be statistically significant (p-value = 0.73) using a Pearson's χ^2 test. Analogously, a potential association between the grading system and response to therapy was tested by first lumping together well and moderately differentiated tumours into one group, and using poorly differentiated tumors as second group. Again, there was no evidence of a statistically significant association (p-value = 0.41).

Experimental results

The performance metrics relative to all the predictive models are summarized in Table 2. Specificity represents the proportion of actual respondents (positives) which are correctly identified as such, and the sensitivity represents the proportion of non-respondents (negatives) which are correctly identified as such. In terms of average accuracy, the 3S-CNN algorithm outperforms all other models, and its performance is followed by a GB algorithm trained on hand-crafted features. Excluding LR, all the classifiers trained on texture features perform better when the feature vector is replaced by principal components. Finally, we note that apart from the 3S-CNN algorithm, all other algorithms were outperformed by the SUVmax median threshold.

Fig 5 reports the top 10 features and their corresponding score extracted from the GB algorithm. In this figure the feature with largest importance has been given a score of 100%, and all the other features have been scaled accordingly. From this feature ranking analysis it emerges

Table 2. Classification results: each figure is the average of three independent experiments using different training and test datasets.

Method	Sensitivity	Specificity	Accuracy
3S-CNN	80.7±11.5	81.6±9.2	73.4±5.3
1S-CNN	77.9±12.9	58.3±4.2	66.4±5.9
GB	70.5±6.0	63.8±6.1	66.7±5.2
GB with PCA	68.1±7.9	46.8±16.2	66.8±6.0
RF	61.0±8.6	36.4±18.4	57.3±7.8
RF with PCA	65.8±7.5	52.0±28.9	65.7±5.6
SVM	66.9±8.5	38.4±19.2	55.9±8.1
SVM with PCA	67.4±10.3	50.9±5.0	60.5±8.0
Logistic Reg.	60.4±6.2	38.3±7.3	51.4±3.0
Logistic Reg. with PCA	58.9±4.9	38.9±12.5	48.4±8.0
SUV max with threshold	33.0±33.0	35.2±10.2	41.0±4.5
SUVmax median threshold	81.5±1.5	53.0±13.0	67.7±4.2

doi:10.1371/journal.pone.0137036.t002

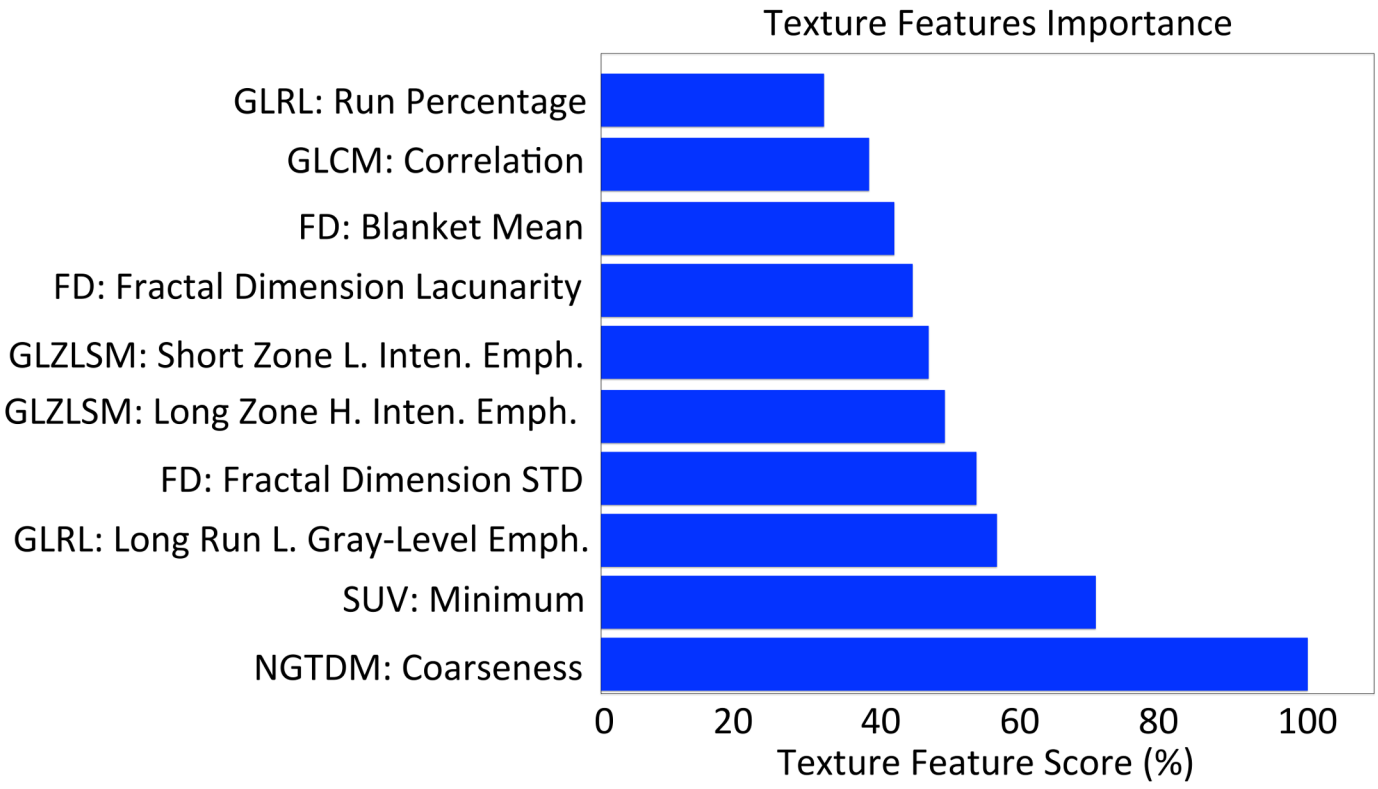

Fig 5. Ten most important texture features for prediction of the chemotherapy response using the GB algorithm. Since these measures are relative, we assign the largest importance a value of 100% and then scale the others accordingly.

doi:10.1371/journal.pone.0137036.g005

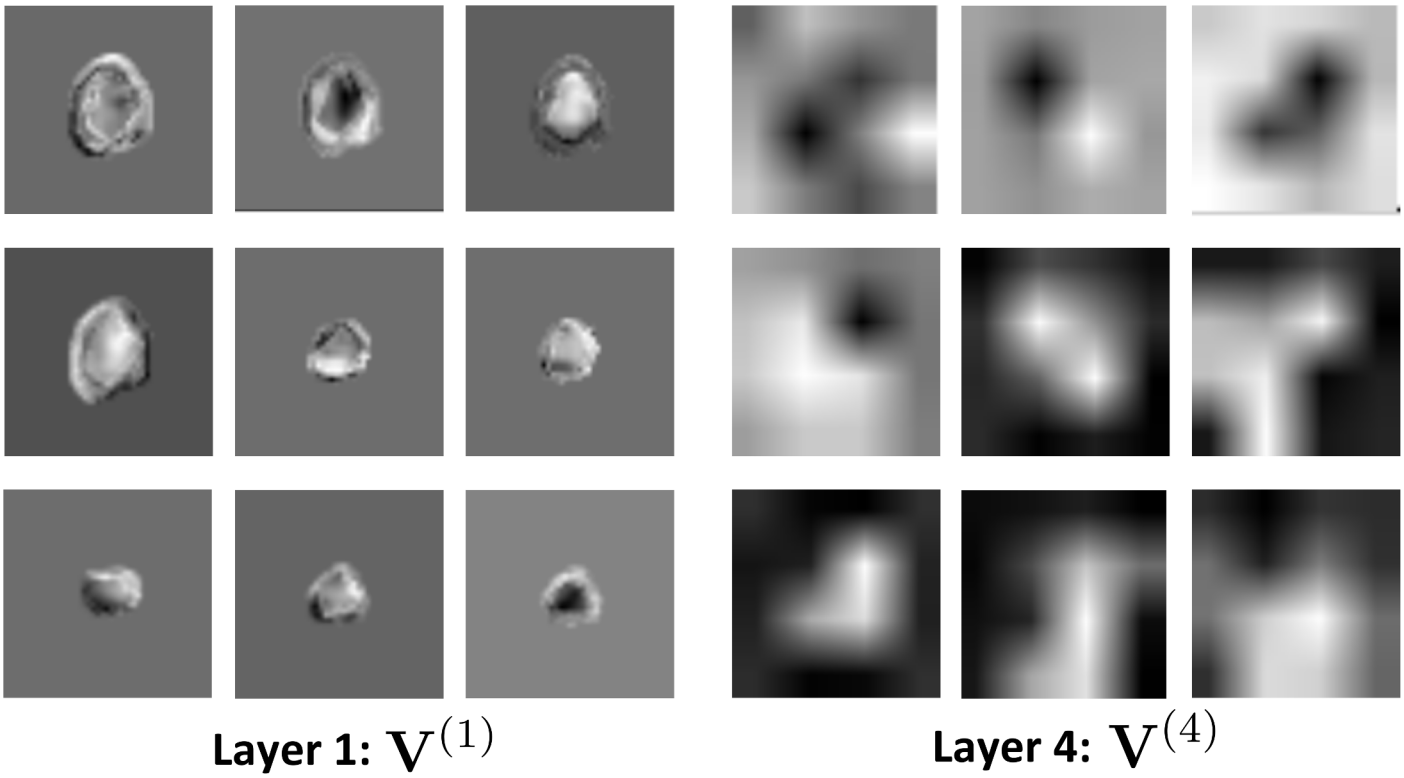

Fig 6. Examples of feature maps in the first and last max-pooling layers $\mathbf{V}^{(1)}$, $\mathbf{V}^{(4)}$ of the CNN architecture. The feature maps illustrate how a specific triplet is represented in the first and last max-pooling layers.

doi:10.1371/journal.pone.0137036.g006

that coarseness, which has been linked to granularity within an image, is the most important feature for response prediction. Coarseness describes local tumor texture based on differences between each voxel and the neighboring voxels in adjacent axial ^{18}F-FDG PET images.

In Fig 6, we illustrate feature maps from the first and last max-pooling layers $\mathbf{V}^{(1)}$, $\mathbf{V}^{(4)}$ of the CNN architecture. These feature maps demonstrate how a specific triplet is represented in the first and last max-pooling layers. The feature maps in the first layer appear to have fused the information from the three adjacent slices of the triplets. In the last layer, the 3S-CNN architecture represents the triplets by remarkably clear and well-defined geometrical patterns with the same level of metabolic activity.

Discussion

The experimental results in Table 2 provide evidence that 3S-CNN outperforms the predictive algorithms trained on a large set of pre-determined imaging features. We believe that the features produced by this approach encompass several of the standard texture features, with the advantage of being completely automatic. The superior performance achieved by the CNN algorithm is due to the exploitation of 3D ^{18}F-FDG uptake information captured by the PET scans and the fact that the architecture learns imaging features that are directly relevant to the clinical endpoint. To our knowledge, the potential predictive power of deep neural networks that only use the raw data as input, and build internal representations of the PET images, has never been assessed for the prediction of chemotherapy response.

In the literature, the decrease in mean or maximum metabolic activity measured by SUV parameters within the first two weeks of neoadjuvant therapy is often considered to be the best available predictor of histopathologic tumor response, however the sensitivities and specificities are still below 67 − 70% (95% confidence intervals ranging from 62% to 76%) [53, 54]. Beyond the prognostic role of FDG uptake changes over the course of the treatment, the role of SUV parameters from only baseline ^{18}F-FDG PET images has been investigated in various studies [7–9], but results conflict as to whether SUV parameters carry any predictive power to assess response to therapy. Specifically, two studies reported that patients with high initial SUVmax values associated with higher probability of response to chemoradiation while one study reported that SUV measurements were not significant factors of the response.

Two studies have explored the capacity of textural features extracted from only pretherapy PET images to differentiate patients with respect to response to therapy. Tixier et al. [21] have analyzed the association between 38 textural features extracted from pretherapy ^{18}F-FDG PET images of 41 patients with esophageal cancer and response to therapy using the Kruskal-Wallis non-parametric test. The sensitivity and specificity reported here varied from 46% to 92% and 45% to 91%, respectively. The predictive capacity of SUV parameters and textural features extracted from GLCMs and NGTDMs has been investigated by Cheng et al. [55] in a cohort of 70 patients with esophageal cancer. This study reported AUC (area under the curve) values of 0.662 for SUV entropy and 0.663 for uniformity. Finally, several studies have investigated the role of textural features and SUV parameters from pretherapy ^{18}F-FDG PET images in predicting response to therapy in breast, lung, cervix and head and neck cancers, reporting AUC values between 0.7 and 1.0 [22–25]. In particular, Cook et al. [25] carried out a Kaplan-Meier analysis to analyze the association between textural features and survival outcomes such as overall survival (OS), progression free survival (PFS) and local PFS. They reported sensitivity and specificity to have varied between 59% to 94% and 42% to 63%, respectively. The main limitation in these studies is the use of low sample sizes ranging from 9 to 53. Thus, the role of these metrics and the clinical relevance remains to be further validated.

From the results in Table 2, it is particularly interesting to notice the difference in performance between 1S-CNN and 3S-CNN. These results enhance the belief that local 3D information of the ^{18}F-FDG uptake can be beneficial for the chemotherapy response prediction. Excluding LR, all the classifiers trained on texture features perform better when the feature vector is replaced by principal components–this is expected since several features contain redundant information. Also, the threshold SUVmax median outperformed all the algorithms except the 3S-CNN, revealing that the SUV carry predictive power. According to the rankings in Fig 5, the texture feature coarseness derived from NGTDMs is the parameter that best differentiates responders and non-responders. Coarseness describes local tumor texture based on differences between each voxel and the neighboring voxels in adjacent axial ^{18}F-FDG PET images. This result is consistent with previously reported evidence that high coarseness values are associated with a greater risk of local tumor progression in non-small lung cancer [25]. Moreover, previous findings have also suggested that coarseness is a texture feature that may discriminate well between responders to chemoradiotherapy from non-responders in oesophageal cancer [21]. Remarkably, many texture features appearing in the top 10 ranking were extracted using a variety of different methods, including fractal analysis, statistical based texture matrices (including GLRLM, GLCM, GLZLSM, NGTDM) and the SUV parameter. This message stresses again the importance of including a very large ensemble of texture features in a radiomics approach. Finally, from the performance of SUVmax median threshold (see Table 2) and the importance ranking of the SUVmin (see Fig 5), our study supports the conflicting evidence that the SUV parameters can discriminate the behavior of a tumor to treatment before therapy.

The number of tumor volumes that is available for this study may lead to overfitting and consequently to degradation of an algorithm's generalization ability on unseen test examples. The predictive algorithms we included in the comparison, such as GB, RF and SVM, encompass several mechanisms to prevent overfitting [52]. For both 3S-CNN and 1S-CNN several additional attempts were made to further reduce overfitting. For instance, we replaced the hyperbolic tangent non-linearities with rectified linear unit (ReLu) non-linearities. Compared to hyperbolic tangent non-linearities, ReLu accelerate the convergence of the MSGD algorithm for the training of the CNN and it is less prone to the gradient vanishing problem [56]. Also, we deployed the technique of Dropout in the FC layer of the CNN. Dropout prevents the neurons from co-adaptation, thus reducing the overfitting of the training dataset [57]. Despite the known advantages of these techniques, they did not significantly improve the generalization performance in our case.

Substantial improvements would be expected by increasing the number of training PET images for which we have clinical information. In future work will also aim at developing a multi-modality algorithm in order to take advantage from both PET and CT images, since PET images ignore the anatomical information and do not present well-defined tumor boundaries because of their relatively poor spatial resolution. We believe that a combination of anatomical and corresponding FDG uptake information will further improve the quality of extracted imaging features and lead to significant improvement in the neoadjuvant chemotherapy response prediction [58]. These model predictions could offer the potential to stratify patients for preoperative therapy before surgery in clinical trials.

Conclusions

Esophageal cancer is associated with high mortality and it is of vital importance to be detected and treated in early stage. In advanced stages, preoperative chemotherapy or radiotherapy can play an essential role in the improvement of survival for patients who respond to the treatment. By contrast, for patients who do not respond to preoperative treatment there is a need for different treatment tactics in order to increase the probability of tumor control. Therefore, the ability to noninvasively predict treatment response before therapy is of great interest and could allow oncologists to personalize future cancer treatments in the clinic.

In the present study we have proposed two different methodologies to predict neoadjuvant chemotherapy response based on pretherapy ^{18}F-FDG PET images. In the first methodology, 3S-CNN were employed to hierarchically learn FDG uptake patterns that are associated with response to neoadjuvant chemotherapy by Mandard. In the second methodology, a wide variety of "hand-engineered" features were derived from the same images and then used as predictor variables in machine learning algorithms. 3S-CNN algorithm outperformed all machine learning algorithms trained on "hand-engineered" ^{18}F-FDG PET imaging features. In conjuction with the variety of the textural features ranked by GB algorithm, our preliminary results indicates that synthesizing features that extensively exploit the heterogeneity of the FDG uptake information with respect to chemotherapy response prediction might offer the potential to capture all the relevant information in the ^{18}F-FDG PET images. However, further testing using larger datasets is required to validate the predictive power of 3S-CNN for clinical decision-making.

Acknowledgments

The authors acknowledge support from the NIHR Biomedical Research Centre of Guys & St Thomas' NHS Trust in partnership with Kings College London, and King's College London and UCL Comprehensive Cancer Imaging Centre funded by the CRUK and EPSRC in association with the MRC and DoH (England).

Author Contributions

Conceived and designed the experiments: PY GM GC VG. Performed the experiments: PY MS HS. Analyzed the data: PY. Wrote the paper: PY GC VG GM. Contributed to data collection: AD.

References

1. Hofstetter W, Swisher SG, Correa AM, Hess K, Putnam JB, Ajani JA, et al. Treatment outcomes of resected esophageal cancer. Annals of Surgery. 2002; 236(3):376–384. doi: 10.1097/00000658-200209000-00014 PMID: 12192324
2. Ychou M, Boige V, Pignon JP. Perioperative chemotherapy compared with surgery alone for resectable gastroesophageal adenocarcinoma: an FNCLCC and FFCD multicenter phase III trial. Clinical Oncology. 2011; 29(13):1715–1721. doi: 10.1200/JCO.2010.33.0597
3. van Hagel P, Hulshof MC, van Lanschot JB, Steyerberg EW, van Berge Henegouwen MI, Wijnhoven BLP, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. The New England Journal of Medicine. 2012; 366(22):2074–2084. doi: 10.1056/NEJMoa1112088
4. Kelsen DP, Ginsberg R, Pajak TF, Sheahan DG, Gunderson L, Mortimer J, et al. Chemotherapy followed by surgery compared with surgery alone for localized esophageal cancer. The New England Journal of Medicine. 1998; 366:1979–1984. doi: 10.1056/NEJM199812313392704
5. Rohren EM, Turkington TG, Coleman RE. Clinical applications of PET in oncology. Radiology. 2004; 231(2):305–332. doi: 10.1148/radiol.2312021185 PMID: 15044750
6. Juweid ME, Cheson BD. Positron-emission tomography and assessment of cancer therapy. The New England Journal of Medicine. 2006; 354:496–507. doi: 10.1056/NEJMra050276 PMID: 16452561
7. Hatt M, Visvikis D, Pradier O, le Rest CC. Baseline 18F-FDG PET image-derived parameters for therapy response prediction in oesophageal cancer. European Journal Nuclear Medicine and Molecular Imaging. 2011; 38:1595–1606. doi: 10.1007/s00259-011-1755-7
8. Rizk NP, Tang L, Adusumilli PS, Bains MS, Akhurst TJ, Ilson D, et al. Predictive value of initial PET SUVmax in patients with locally advanced esophageal and gastroesophageal junction adenocarcinoma. Journal of Thoracic Oncology. 2009; 4(7):875–879. doi: 10.1097/JTO.0b013e3181a8cebf PMID: 19487968
9. Javeri H, Xiao L, Rohren E, Komaki R, Hofstetter W, Lee JH, et al. Influence of the baseline 18FDG positron emission tomography results on survival and pathologic response in patients with gastroesophageal cancer undergoing chemoradiation. Cancer. 2009; 115(3):624–630. doi: 10.1002/cncr.24056 PMID: 19130466
10. Vesselle H, Schmidt RA, Pugsley JM, Li M, Kohlmyer S, Valliueres E, et al. Lung cancer proliferation correlates with 18FDG uptake by positron emission tomography. Clinical Cancer Research. 2000; 6:3837–3844. PMID: 11051227
11. Rajendran JG, Schwartz DL, OSullivan J, Peterson LM, Ng P, Scharnhorst J, et al. Tumour hypoxia imaging with 18F fluoromisonidazole positron emission tomography in head and neck cancer. Clinical Cancer Research. 2006; 12:5435–5441. doi: 10.1158/1078-0432.CCR-05-1773 PMID: 17000677
12. Kunkel M, Reichert TE, Benz P, Lehr HA, Jeong JH, Wieand S, et al. Overexpression of Glut-1 and increased glucose metabolism in tumours are associated with a poor prognosis in patients with oral squamous cell carcinoma Cancer. Cancer. 2003; 97:1015–1024. doi: 10.1002/cncr.11159 PMID: 12569601
13. Kalff V, Duong C, Drummond EG, Matthews JP, Hicks RJ. Findings on 18F-FDG PET scans after neoadjuvant chemoradiation provides prognostic stratification in patients with locally advanced rectal carcinoma subsequently treated by radical surgery. Journal of Nuclear Medicine. 2006; 47(1):14–22. PMID: 16391182
14. Brun E, Kjellen E, Tennvall J, Ohlsson T, Sandell A, Perfekt R, et al. FDG PET studies during treatment: prediction of therapy outcome in head and neck squamous cell carcinoma. Head and Neck. 2002; 24 (2):127–135. doi: 10.1002/hed.10037.abs PMID: 11891942
15. Weber WA, Ott K, Becker K, Dittler HJ, Helmberger H, Avril NE, et al. Prediction of response to preoperative chemotherapy in adenocarcinomas of the esophagogastric junction by metabolic imaging. Journal of Clinical Oncology. 2001; 19(12):3058–3065. PMID: 11408502
16. Cook GJR, Siddique M, Taylor BP, Yip C, Chicklore S, Goh V. Radiomics in PET: principles and applications. Clinical and Translational Imaging. 2014; 2(3):269–276. doi: 10.1007/s40336-014-0064-0
17. Lampin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012; 48(4):441–446. doi: 10.1016/j.ejca.2011.11.036

18. Hicks RJ, Manus MPM, Matthews JP, Hogg A, Binns D, Rischin D, et al. Early FDG-PET imaging after radical radiotherapy for non-small-cell lung cancer: Inflammatory changes in normal tissues correlate with tumoor response and do not confound therapeutic response evaluation. International Journal of Radiation Oncology, Biology, and Physics. 2004; 60(2):412–418. doi: 10.1016/j.ijrobp.2004.03.036

19. Miller TR, Pinkus E, Dehdashti F, Grigsby PW. Improved prognostic value of 18F-FDG PET using a simple visual analysis of tumor characteristics in patients with cervical cancer. Journal of Nuclear Medicine. 2003; 44(2):192–197. PMID: 12571208

20. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clinical Radiology. 2004; 59(12):1061–1069. doi: 10.1016/j.crad.2004.07.008 PMID: 15556588

21. Tixier F, Rest CCL, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumoral heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concominant radiochemotherapy in esophageal cancer. Journal of Nuclear Medicine. 2011; 52:369–378. doi: 10.2967/jnumed.110.082404 PMID: 21321270

22. Willaime J, Turkheimer F, Kenny L, Aboagye E. Image descriptors of intra-tumor proliferative heterogeneity predict chemotherapy response in breast tumors. Journal of Nuclear Medicine Meeting Abstracts. 2012; 53:387.

23. el Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment. Pattern Recognition. 2009; 42(6):1162–1171. doi: 10.1016/j.patcog.2008.08.011 PMID: 20161266

24. Ha S, Lee HY, Kim SE. Prediction of response to neoadjuvant chemotherapy in patients with breast cancer using texture analysis of 18F-FDG PET/CT. Journal of Nuclear Medicine. 2014; 55:623.

25. Cook GJR, Yip C, Siddique M, Goh V, Chicklore S, Roy A, et al. Are Pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? Journal of Nuclear Medicine. 2013; 54:19–26. doi: 10.2967/jnumed.112.107375 PMID: 23204495

26. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics. 1980; 36(4):193–202. doi: 10.1007/BF00344251 PMID: 7370364

27. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11):2278–2324. doi: 10.1109/5.726791

28. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges C, Bottou L, Weinberger K, editors. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012. p. 1097–1105.

29. Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI. vol. 2; 2013. p. 411–418.

30. Ciresan DC, Meier U, Schmidhuber J. Multi-column Deep Neural Networks for Image Classification. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012. p. 3642–3649.

31. Xu Y, Mo T, Feng Q, Zhong P, Lai M, Chang EI. Deep Learning of Feature Representation with Multiple Instance Learning for Medical Image Analysis. In: Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2014. p. 1626–1630.

32. Mandard AM, Dalibard F, Mandard JC, Manray J, Henry-Amar M, Petiot JF, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma: Clinicopathologic correlations. Cancer. 1994; 73(11):2680–2686. doi: 10.1002/1097-0142(19940601)73:11%3C2680::AID-CNCR2820731105%3E3.0.CO;2-C PMID: 8194005

33. Haralick RM, Shanmungam K, Dinstein I. Textural Features of Image Classification. IEEE Transactions on Systems Man and Cybernetics. 1973; 3(6):610–621. doi: 10.1109/TSMC.1973.4309314

34. Conners RW, Harlow CA. A theoretical comparison of texture algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1980; 2(3):204–222. doi: 10.1109/TPAMI.1980.4767008 PMID: 21868894

35. Galloway MM. Texture analysis using gray level run lengths. Computer Graphics and Image Processing. 1975; 4(2):172–179. doi: 10.1016/S0146-664X(75)80008-6

36. Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. In: Int. Conf, Pattern Recognition and Information Processing (ICPRIP); 2009. p. 140–145.

37. Amadasun M, King R. Textural Features Correspondind to Textural Properties. IEEE Transactions on Systems, Man, and Cybernetics. 1989; 19(5):1264–1274. doi: 10.1109/21.44046

38. Chen CH, Pau LF, Wang PSP. The Handbook of Pattern Recognition and Computer Vision. World Scientific Publishing Pte. Ltd; 2011.

39. Pentland AP. Fractal-based description of natural scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984; 6(6):661–674. doi: 10.1109/TPAMI.1984.4767591 PMID: 22499648

40. Miwa K, Inubushi M, Wagatsuma K. FDG uptake heterogeneity evaluated by fractal analysis improves the differential diagnosis of pulmonary nodules. European Journal of Radiology. 2014; 83(4):715–719. doi: 10.1016/j.ejrad.2013.12.020 PMID: 24418285

41. Kalliokoski KK, Kuusela TA, Nuutila P, Tolvanen T, Oikonen V, Terras M, et al. Perfusion heterogeneity in human skeletal muscle: fractal analysis of PET data. European Journal of Nuclear Medicine. 2001; 28(4):450–456. doi: 10.1007/s002590000458 PMID: 11357494

42. Mandelbrot BB. Fractal Geometry of Nature. W. H. Freeman and Company; 1982.

43. Sarkar N, Chaudhuri B. An efficient differential box-counting approach to compute fractal dimension of images. IEEE Transactions on Systems, Man and Cybernetics. 1994; 24(1):115–120. doi: 10.1109/21.259692

44. Peng CYJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. Journal of Educational Research. 2002; 96(1):3–14. doi: 10.1080/00220670209598786

45. Friedman J. Greedy Function Approximation: A gradient Boosting Machine. Annals of Statistics. 2001; 29(5):1189–1232. doi: 10.1214/aos/1013203450

46. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. doi: 10.1023/A:1010933404324

47. Cortes C, Vapnik V. Support vector networks. Machine Learning. 1995; 20(3):273–297. doi: 10.1023/A:1022627411411

48. LeCun Y. Generalization and Network Design Strategies. Department of Computer Science, University of Toronto; 1989.

49. Ranzato MA, Huang FJ, Boureau YL, LeCun Y. Unsupervised learning of invariant feature hierarchies with application to object recognition. In: Computer Vision and Pattern Recognition (CVPR); 1995. p. 1–8.

50. Bousquet O, Bottou L. The tradeoffs of large scale learning. In: Platt JC, Koller D, Singer Y, Roweis S, editors. Advances in Neural Information Processing Systems. vol. 20. NIPS Foundation; 2008. p. 161–168.

51. Bergstra J, Breuleux O, Bastien F, Lamblin L, Pascanu R, Desjardins G, et al. Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference; 2010. p. 1–7.

52. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inferences, and prediction. Springer; 2001.

53. Schollaert P, Crott R, Bertrand C, D'Hondt L, Borght TV, Krug B. A systematic review of the predictive value of (18)FDG-PET in esophageal and esophagogastric junction cancer after neoadjuvant chemoradiation on the survival outcome stratification. Journal of Gastrointestinal Surgery. 2014; 18(5):894–905. doi: 10.1007/s11605-014-2488-2 PMID: 24638928

54. Zhu W, Xing L, Yue J, Sun X, Sun X, Zhao H, et al. Prognostic significance of SUV on PET/CT in patients with localised oesophagogastric junction cancer receiving neoadjuvant chemotherapy/chemoradiation:a systematic review and meta-analysis. British Journal of Radiology. 2012; 85(1017):694–701. doi: 10.1259/bjr/29946900

55. Cheng NM, Fang YH, Chang JT, Huang CG, Tsan DL, Ng SH, et al. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. Journal of Nuclear Medicine. 2013; 54(10):1703–1709. doi: 10.2967/jnumed.112.119289 PMID: 24042030

56. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: Geoffrey GJ, Dunson DB, editors. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, AISTATS-11. vol. 15. Journal of Machine Learning Research—Workshop and Conference Proceedings; 2011. p. 315–323.

57. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Deep Sparse Rectifier Neural Networks. Journal of Machine Learning Research. 2014; 15:1929–1958.

58. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, el Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiotherapy and Oncology. 2012; 102(2):239–245. doi: 10.1016/j.radonc.2011.10.014 PMID: 22098794