Expert Systems With Applications 167 (2021) 114244

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Unsupervised supervoxel-based lung tumor segmentation across patient scans in hybrid PET/MRI

Stine Hansen [a,*], Samuel Kuttner [c,d], Michael Kampffmeyer [a], Tom-Vegard Markussen [e], Rune Sundset [c,d], Silje Kjærnes Øen [b], Live Eikenes [b], Robert Jenssen [a]

[a] Department of Physics and Technology, UiT The Arctic University of Norway, NO-9037 Tromsø, Norway
[b] Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
[c] PET Imaging Center, University Hospital of North Norway, NO-9038 Tromsø, Norway
[d] Department of Clinical Medicine, UiT The Arctic University of Norway, NO-9037 Tromsø, Norway
[e] University Hospital of North Norway, NO-9019 Tromsø, Norway

ARTICLE INFO

Keywords:
Clustering
Unsupervised learning
Medical image segmentation
Tumor segmentation
Hybrid PET/MRI

ABSTRACT

Tumor segmentation is a crucial but difficult task in treatment planning and follow-up of cancerous patients. The challenge of automating the tumor segmentation has recently received a lot of attention, but the potential of utilizing hybrid positron emission tomography (PET)/magnetic resonance imaging (MRI), a novel and promising imaging modality in oncology, is still under-explored. Recent approaches have either relied on manual user input and/or performed the segmentation patient-by-patient, whereas a fully unsupervised segmentation framework that exploits the available information from all patients is still lacking.

We present an unsupervised across-patients supervoxel-based clustering framework for lung tumor segmentation in hybrid PET/MRI. The method consists of two steps: First, each patient is represented by a set of PET/MRI supervoxel-features. Then the data points from all patients are transformed and clustered on a population level into tumor and non-tumor supervoxels. The proposed framework is tested on the scans of 18 non-small cell lung cancer patients with a total of 19 tumors and evaluated with respect to manual delineations provided by clinicians. Experiments study the performance of several commonly used clustering algorithms within the framework and provide analysis of (i) the effect of tumor size, (ii) the segmentation errors, (iii) the benefit of across-patient clustering, and (iv) the noise robustness.

The proposed framework detected 15 out of 19 tumors in an unsupervised manner. Moreover, performance increased considerably by segmenting across patients, with the mean dice score increasing from $\mathbf{0.169 \pm 0.295}$ (patient-by-patient) to $\mathbf{0.470 \pm 0.308}$ (across-patients). Results demonstrate that both spectral clustering and Manhattan hierarchical clustering have the potential to segment tumors in PET/MRI with a low number of missed tumors and a low number of false-positives, but that spectral clustering seems to be more robust to noise.

1. Introduction

Medical imaging is today an integrated part of diagnostics and treatment planning of cancer patients. In particular, hybrid positron emission tomography (PET)/computed tomography (CT) has become an established tool in tumor detection, characterization, staging, and monitoring (Flechsig, Mehndiratta, Haberkorn, Kratochwil, & Giesel, 2015; Ehman et al., 2017). A more recent advancement in hybrid radiologic imaging is the PET/magnetic resonance imaging (MRI) scanner, in which the anatomical information is obtained from MRI instead of CT. As opposed to CT, MRI does not involve harmful ionizing radiation and offers superior soft-tissue contrast with high spatial resolution, making PET/MRI a promising hybrid modality in, for instance, oncology. Nevertheless, the potential of hybrid PET/MRI is still being investigated and remains an open question (Ehman et al., 2017). This includes its potential in the important task of lung tumor segmentation,

* Corresponding author.
E-mail addresses: s.hansen@uit.no (S. Hansen), samuel.kuttner@uit.no (S. Kuttner), michael.c.kampffmeyer@uit.no (M. Kampffmeyer), tom-vegard.markussen@unn.no (T.-V. Markussen), rune.sundset@unn.no (R. Sundset), silje.k.oen@ntnu.no (S.K. Øen), live.eikenes@ntnu.no (L. Eikenes), robert.jenssen@uit.no (R. Jenssen).

https://doi.org/10.1016/j.eswa.2020.114244
Received 10 September 2020; Received in revised form 2 November 2020; Accepted 4 November 2020
Available online 29 November 2020

which is the focus of this paper.

Lung cancer is the most frequently diagnosed cancer type in the world, with a predicted number of 2.1 million new incidences in 2018 (Bray et al., 2018). An important, but inherently difficult, part of the treatment planning and follow-up of these cancerous patients is the process of isolating the tumor volume in medical images (Sauwen et al., 2016). Today, this tumor segmentation is commonly performed manually in a slice-by-slice manner. However, this work is tedious and susceptible to subjective interpretation (Caldwell et al., 2001; Hurkmans et al., 2001). A great amount of effort has therefore been put into the investigation of automatic tumor segmentation (Foster, Bagci, Mansoor, Xu, & Mollura, 2014; Gordillo, Montseny, & Sobrevilla, 2013).

The majority of existing methods for automatic medical image segmentation are based on *supervised* models that require fully annotated data sets to learn a classification of voxels into tumor and non-tumor voxels (De Bruijne, 2016). Such data sets are time-consuming to generate because segmentations have to be manually delineated for a large set of images. *Unsupervised* segmentation methods, on the other hand, have the benefit of not requiring annotations and is typically relying on voxel-wise clustering based on similarity within the data.

Only a few studies have considered tumor segmentation in hybrid PET/MRI (Bagci et al., 2013; Xu, Bagci, Udupa, & Mollura, 2015; Leibfarth et al., 2015; Sbei, ElBedoui, Barhoumi, Maksud, & Maktouf, 2017; Sbei, ElBedoui, Barhoumi, & Maktouf, 2020) (see related work section for details). In this paper, we aim to contribute to the recent line of work in order to further investigate the potential of PET/MRI for *unsupervised* lung tumor segmentation.

Unlike previous approaches to unsupervised tumor segmentation in hybrid PET/MRI, which perform segmentation in a patient-by-patient manner (Bagci et al., 2013; Sbei et al., 2020), we take advantage of the information in *all* available patient scans. In patient-by-patient segmentation approaches where the segmentation is based on single image pairs (PET and MRI from one patient), the number of voxels representing the tumor might be insufficient for the clustering algorithms to recognize them as a separate cluster. By instead clustering *across* patients in a population-level manner, we open up to taking advantage of the information in all patients when finding patterns to base the segmentation on. Voxel-wise clustering across all patients is, however, not computationally feasible as the total number of voxels becomes too high. To overcome this barrier, we take inspiration from a recent innovative approach to the problem of clustering tumor subvolumes (Wu et al., 2016; Even et al., 2017), by employing supervoxels rather than working directly on voxel level.

In our work, we thus examine a two-stage clustering approach for automatic lung tumor segmentation, where we first do a patient-level over-segmentation into homogeneous supervoxels, before we group the supervoxels *across all patients* and do a population-level clustering into "tumor" and "non-tumor" supervoxels. Since the problem at hand is complex and requires a systematic analysis of the proposed two-stage approach, we provide a comparison and analysis of several different clustering procedures to achieve this task. We further evaluate the advantage of utilizing across-patients information, the method's robustness to noise, the effect of tumor size and the types of segmentation errors.

The key contributions of this paper are:

1. A novel unsupervised lung tumor segmentation framework that can utilize information across patients in PET/MRI images.
2. An analysis of several commonly used clustering approaches within the proposed framework.
3. An analysis of the segmentation mistakes that the different clustering algorithms make and how tumor size affects the performance.
4. An analysis of the benefit of across-patients clustering compared to patient-by-patient clustering.
5. An analysis of the proposed method's sensitivity to image noise.

In the following, Section 2 provides a brief overview of the related work. Section 3 introduces the data set used as part of this study and Section 4 presents the proposed framework for lung tumor segmentation. In Section 5, the experimental results and an analysis of the segmentation mistakes, the effect of tumor size, the benefit of clustering across-patients, and a noise analysis are provided. Finally, Section 6 and Section 7 discuss outlook, limitations, and provide conclusions.

2. Related work

Today there exists a large range of methods for tumor segmentation in established modalities such as PET, CT, MRI and hybrid PET/CT (Foster et al., 2014; Moghbel, Mashohor, Mahmud, & Saripan, 2018; Wadhwa, Bhardwaj, & Verma, 2019; Ju et al., 2015), while the use of hybrid PET/MRI is less explored. In order to provide the necessary context to place this paper's contributions in the field, this section will highlight previous work within hybrid PET/MRI tumor segmentation. As this is a relatively new modality, previous studies are limited to only a handful of articles. To the authors' knowledge, the first study on tumor segmentation in hybrid PET/MRI was the study by Bagci et al. (2013), in which a random walk based co-segmentation approach with automatic foreground/background seed selection was developed. By unifying the graph representation of each modality in a single product lattice, they reformulated the random walk method to jointly delineate objects in different image modalities. A few years later, in the study by Xu et al., 2015, a tumor segmentation approach based on fuzzy connectedness with a visibility weighting scheme was proposed as a faster alternative achieving similar performance to Bagci et al. (2013). However, as opposed to Bagci et al. (2013), which performed segmentation unsupervised, the approach by Xu et al. (2015) required user-specified weights for each modality. Sbei et al. (2017) further developed the fuzzy connectedness approach and combined it with the graph cut method to address problems with leakage through weak boundaries. In recent work, Sbei et al. (2020) made additional modifications to the method by improving the automatic seed generation step and automatically generating intermediate images with reduced heterogeneity, which the segmentation is based on.

Common to all these approaches is that the segmentation is performed patient-by-patient. That is, only the information in the PET/MRI from one patient is considered at a time. A quite different approach was developed by Leibfarth et al. (2015), where tumor probability maps were derived for both PET and MRI images using heuristic probability mapping functions relating probability values and intensities on voxel level. Then the tumor delineation was derived using the threshold level set segmentation algorithm on the combined probability map defined by the weighted sum of the single maps. In this approach, the parameters were optimized by considering multiple patients in a supervised leave-one-out manner.

In our proposed framework, we depart from previous work and perform a PET/MRI tumor segmentation that is both *unsupervised* and exploits the information in *all available patient scans* by performing an across-patients clustering.

3. Dataset

For the current work, we used 18 PET/MRI acquisitions from a previous lung cancer study (Kuttner et al., 2020). The study was approved by the Norwegian Regional Committees for Medical and Health Research Ethics (REC reference 2017/915), and all patients signed written informed consent. The benefit of using these exams is that all scans contain one or multiple tumors diagnosed as either adenocarcinoma or squamous cell carcinoma, which are the two most common types of non-small cell lung cancer (Raponi et al., 2006).

Prior to PET/MRI, each patient was injected with 4 MBq/kg 18F-fluorodeoxyglucose (FDG). Approximately two hours post-injection, a 10-min, one-bed position PET acquisition of the mediastinum was

performed in a Siemens Biograph mMR (software version VB20P) (Siemens Healthineers, Erlangen, Germany) using a free-breathing and arms-down scan protocol. Simultaneous with PET, a T2-weighted TIRM MRI sequence was acquired. Furthermore, a standard DIXON-based MR sequence was used for attenuation correction of the PET images.

PET images were reconstructed using the ordered-subset expectation–maximization (OSEM) algorithm with three iterations, 21 subsets, and 4 mm Gaussian smoothing. For each PET image, the measured tissue radioactivity concentration [kBq/mL] was normalized against patient body weight and injected dose to obtain the standardized uptake value (SUV) [g/mL]. The gross tumor volume was delineated in the T2 images for all patients based on morphology. PET images were used as an aid to differentiate pathology from anatomy or atelectasis, or from large hilar vessels. Delineations were performed by a thorax radiologist ($> 10y$ experience) using Varian Eclipse Treatment Planning System version 10.0.42. In the PET images, FDG-avid lesions were segmented using a 41% SUVmax threshold. The union of the PET and MRI masks is considered the ground truth mask.

As a pre-processing step, the images were re-sampled to the same isotropic voxel resolution of $2 \times 2 \times 2$ mm^3 using cubic interpolation, resulting in an image size of $114 \times 152 \times 93$ voxels. More information about the data is summarized in Table 1. Fig. 1 shows two PET/MRI pair examples with corresponding ground truth masks.

4. Framework for lung tumor segmentation

Our proposed across-patients supervoxel-based clustering framework segments lung tumors in hybrid PET/MRI. Supervoxels are computed for each patient and features extracted from these are grouped. In order to improve the segmentation performance, the features are transformed using a Box-Cox transformation. Finally, the transformed features are clustered into a foreground (tumor) and a background class. Fig. 2 shows a schematic overview of the lung tumor segmentation approach. The details of the individual stages are discussed in the following.

4.1. Co-registration

We transform the PET image volume from each PET/MRI pair by a B-spline transformation model to align with the MRI scan. The registration is performed unsupervised using the Elastix software (Klein, Staring, Murphy, Viergever, & Pluim, 2009) in Python by running the SimpleElastix toolbox (Marstal, Berendsen, Staring, & Klein, 2016). Elastix is an openly available and frequently used software package for intensity-based medical image registration where the registration problem is formulated as an optimization problem and solved iteratively (Viergever et al., 2016). In this work, the cost function consists of a similarity measure, defined by the mutual information between the two images, and a regularisation term penalizing the displacement magnitude. The cost is optimized in an iterative manner using adaptive stochastic gradient descent in a three-stage pyramidal multi-resolution approach.

Table 1
Detailed information about the dataset and generated supervoxels.

Patients	Total number	18
	Mean age (at exam) [yrs]	72.1
Gender	Male	12
	Female	6
Tumors	Total number	19
Pathology	Adenocarcinoma	12
	Squamous cell carcinoma	7
Resolution	PET [mm]	$[2 \times 2 \times 2]$
	MRI [mm]	$[1 \times 1 \times 5]$
Ground truth	Median TV (union) [mm^3]	9712
	Median TV (MRI) [mm^3]	6664
	Median TV (PET) [mm^3]	8672
Supervoxels	Median no supervoxels	1495
	Minimum no supervoxels	1480
	Maximum no supervoxels	1511
	Minimum size [mm^3]	1600
	Maximum size [mm^3]	40672

TV = tumor volume.

4.2. Supervoxel generation

The idea of supervoxels is to group similar voxels into basic regions that are more meaningful than individual voxels. In this way, supervoxels capture redundancy in the image and provide local image features.

We apply the simple linear iterative clustering (SLIC) algorithm (Achanta et al., 2010), which, in the field of medical image analysis, has extensively been used as a pre-processing step to reduce the computational cost and the effects of noise and imperfect co-registration (Even et al., 2017; Lucchi, Smith, Achanta, Knott, & Fua, 2011; Roth, Farag, Lu, Turkbey, & Summers, 2015; Soltaninejad et al., 2017).

The SLIC algorithm is based on k-means clustering with k cluster centers initialized on a regular grid with intervals of $S = \sqrt[3]{N/k}$, where N is the number of voxels in the image volume. However, two particular modifications differentiate the SLIC algorithm from standard k-means clustering: (1) Whereas standard k-means searches the entire image, the search space in SLIC is limited to a region proportional to the initial supervoxel size N/k. (2) The distance measure D is a weighted combination of the intensity proximity and the spatial proximity in order to control the size and compactness of the supervoxels (Achanta et al., 2012):

$$D = \sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2}, \tag{1}$$

where m is a constant controlling the compactness of the supervoxels, and the distances d_c and d_s, for the case of a three-dimensional grayscale image with voxel intensities l and spatial cooridnates (x,y,z), are given by

$$d_c = \sqrt{(l_j - l_i)^2}, \tag{2}$$

and

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}. \tag{3}$$

The supervoxel generation is based on the same approach as in Even et al., 2017. All image volumes are z-normalized (subtract mean and divide by standard deviation) and for each image pair $\{I_i^{MRI} I_i^{PET}\}_{i=1}^{N_p}$, an average image volume I_i^z is computed and used to extract supervoxels according to the SLIC algorithm. An initial number of $k = 1500$ supervoxels per patient reduces the total number of data points from 28.7 millions (voxels) to less than 27,000 (supervoxels). The left part of Fig. 3 shows an example slice for one patient.

4.3. Feature extraction

To make the analysis clean, we extract two basic intensity features for each supervoxel, i, and define the feature vector

$$\boldsymbol{x}_i = \left[x_i^{MRI}, x_i^{PET}\right], \tag{4}$$

where x_i^{MRI} is the median intensity within the volume of supervoxel i in the MRI image and x_i^{PET} is the median intensity within the volume of supervoxel i in the PET image (Fig. 3, right). By extracting the median intensities, the effects from outlier voxels are suppressed.

Fig. 4 shows a scatter plot of the extracted feature vectors from all patients, with x^{MRI} on the x-axis and x^{PET} on the y-axis. The colors indicate the fraction of tumor voxels (according to the ground truth

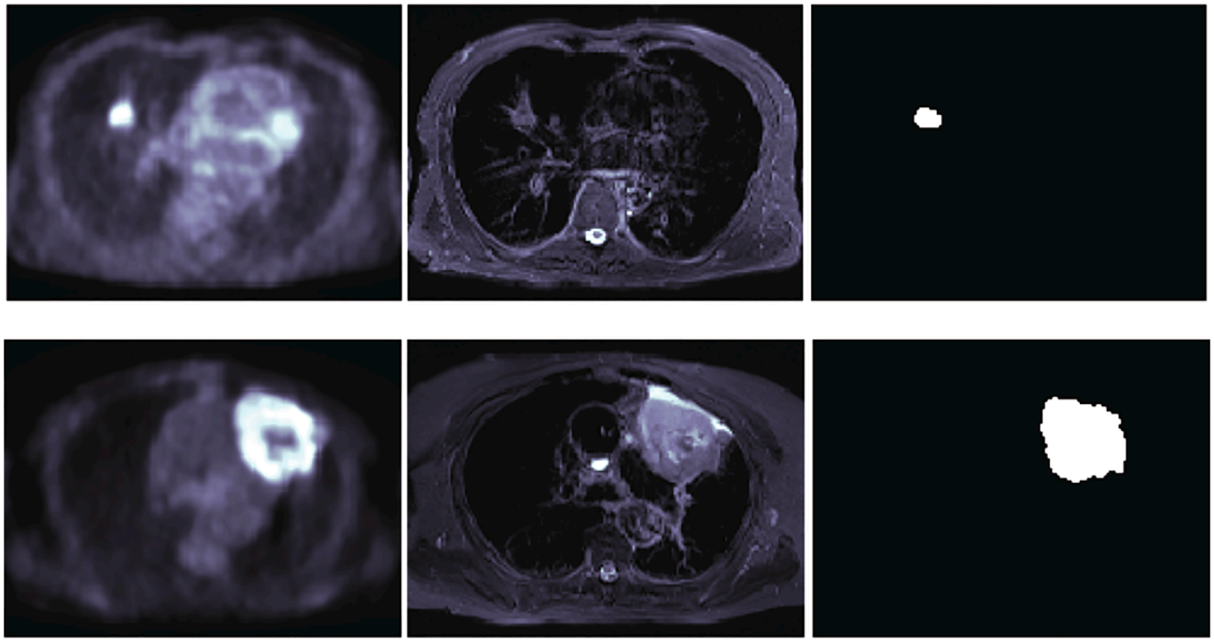

Fig. 1. Two PET/MRI pair examples. Left: PET image slice. Middle: Corresponding MRI image slice. Right: Corresponding ground truth mask indicating the tumor.

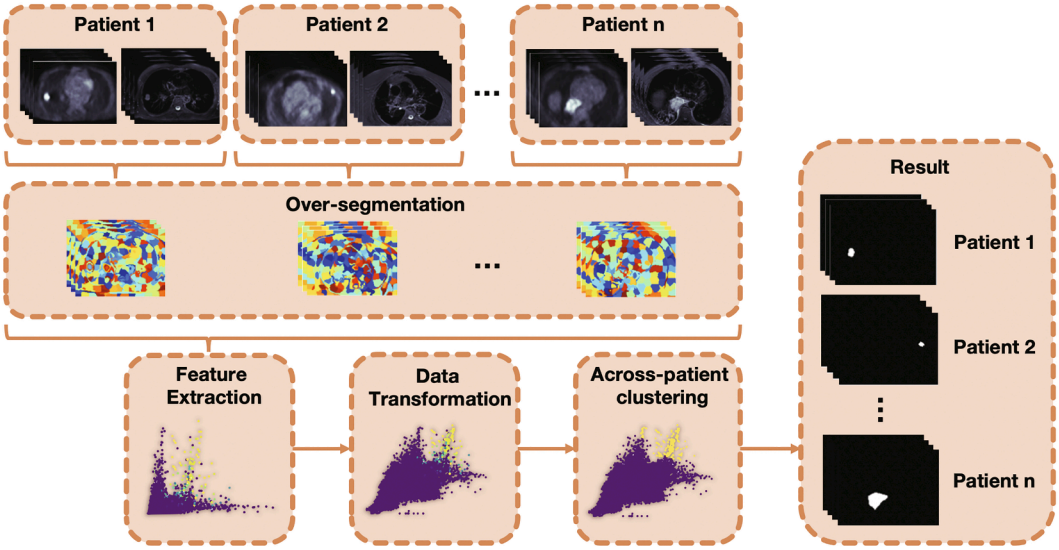

Fig. 2. Schematic of the across-patients supervoxel-based clustering for lung tumor segmentation in hybrid PET/MRI. Based on each co-registered PET/MRI image pair, an over-segmentation is performed to generate supervoxels. From each supervoxel in every patient, two basic intensity features are extracted from the PET/MRI, resulting in a two-dimensional feature space. This feature space is transformed to improve the following clustering into two clusters. Finally, the clustering labels are mapped back to pixel space, giving the resulting segmentation masks. Note that the colors in the scatter plots *before* the clustering indicate the supervoxels' tumor fractions according to the ground truth and is only used for illustration purposes.

labels) within the supervoxels, where yellow translates to pure tumor supervoxel and purple corresponds to pure background supervoxel. This plot illustrates that both modalities contribute with important information in the segmentation task: thresholding any of the two marginal distributions will lead to significant mixing of tumor and non-tumor supervoxels.

4.4. Data transformation

As the original form of the data is not necessarily more suitable for analysis than any function of the data, transformations often play an important role in exploratory data analysis (Stoto & Emerson, 1983). We apply the Box-Cox transformation (Box & Cox, 1964), a long-established power transformation. This transform is a widely used pre-processing step in various fields of applications (Hossain, 2011; Liu, Yin, Wang, & Wang, 2013; Rayens & Srinivasan, 1991; Boroojeni et al., 2017), including tumor detection in MRI (Vos, Barentsz, Karssemeijer, & Huisman, 2012) and classification of lung nodules in CT (Shah et al., 2005). The transformed data $y_i^{(\lambda)}$ is given by

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \text{if} \lambda \neq 0, \\ \ln y_i, & \text{if} \lambda = 0, \end{cases} \tag{5}$$

where the parameter λ is found by maximizing the log-likelihood under the assumption that the *transformed* data is Gaussian, thereby

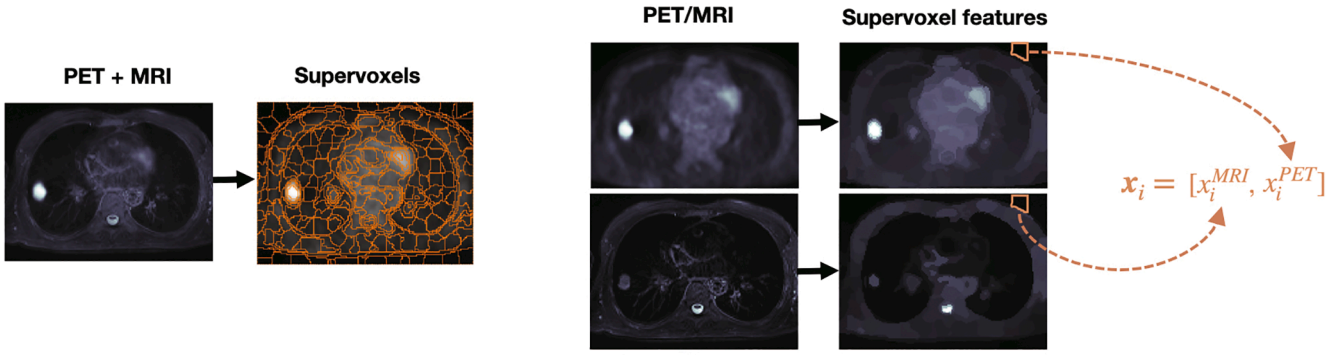

Fig. 3. Illustration of supervoxel generation and feature extraction. Left: From the average image volume I^z, computed based on the z-normalized PET and MRI volumes, we compute the supervoxels. In this specific image slice, the 17,000 voxels are aggregated into 315 supervoxels. Right: Within each supervoxel, we compute the median PET and the median MRI intensity and extract these as supervoxel feature vectors.

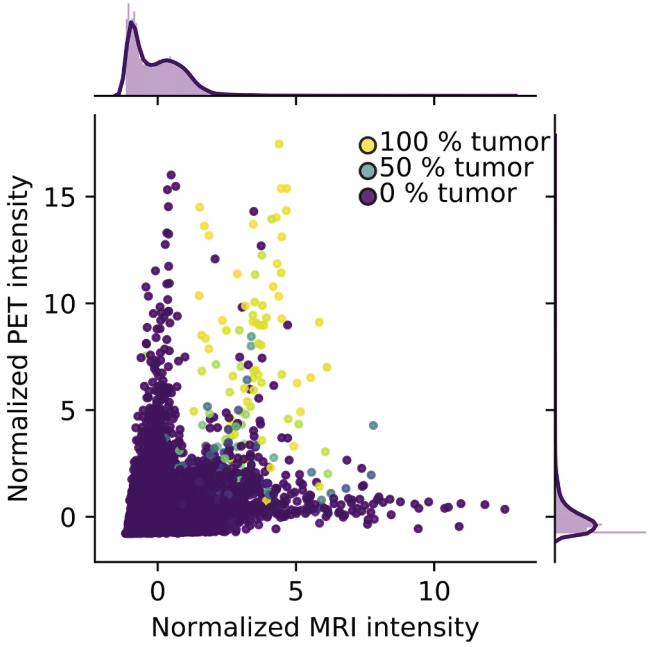

Fig. 4. Scatter plot showing the z-normalized feature space. The x-axis represents the MRI feature whereas the y-axis represents the PET feature.

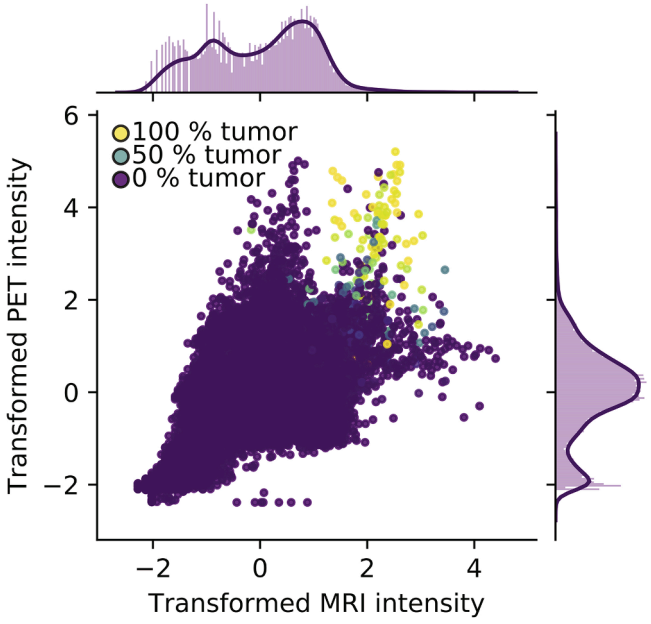

Fig. 5. Scatter plot showing the z-normalized Box-Cox transformed feature space. The x-axis represents the transformed MRI feature whereas the y-axis represents the transformed PET feature.

encouraging the transformed data to be more Gaussian.

Since the SUV values in the original PET images lie in the range [0, 12] whereas the MRI intensities lie in the range [0,1000], we additionally apply a z-normalization to the transformed data. Fig. 5 shows the normalized feature plot after Box-Cox transformation.

4.5. Clustering

Cluster analysis is the study of discovering natural groupings in unlabeled data, such that samples within the same cluster are *similar* and samples in different clusters are *dissimilar*. There exist thousands of clustering algorithms in the literature, and different clustering algorithms (and their parameter settings) often result in different groupings. However, no general "best clustering algorithm" can be named (Jain, 2010). In some way or another, each algorithm enforces a structure on the data and depending on the fit between the model and the data, the resulting clusters will be "good" or "bad" (Jain, 2010).

In this paper, we consider some of the most well-known clustering algorithms in the literature in order to perform lung tumor segmentation in an unsupervised manner. We examine k-means clustering, spectral clustering, and hierarchical clustering. For the benefit of the reader not familiar with these algorithms, we provide a short overview in the following.

4.5.1. K-means clustering

K-means is, due to its simplicity and computational efficiency, one of the most used clustering algorithms in the literature (Jain, 2010). The algorithm partitions the data into k disjoint clusters in a two-step iterative optimization of the cost function, given by (Bishop, 2006):

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2, \tag{6}$$

where N is the number of data points (supervoxels), $\mathbf{x}_n$ is the feature vector of the nth data point, $r_{nk} \in \{0, 1\}$ is the cluster assignment of the nth data point to cluster k, and $\boldsymbol{\mu}_k$ is the cluster representative of the kth cluster, given by the mean of the feature vectors assigned to that cluster.

The algorithm is initialized by choosing a set of initial cluster representatives. Then, each iteration consists of two steps which are repeated until convergence:

1. Assign $\mathbf{x}_n, n = 1, \ldots, N$ to the closest cluster, defined by its cluster representative $\boldsymbol{\mu}_k, k = 1, \ldots, K$.
2. Update cluster representatives $\boldsymbol{\mu}_k, k = 1, \ldots, K$ as the mean of all data points assigned to it.

4.5.2. Hierarchical clustering

Another common clustering approach is hierarchical clustering. In this work we employ hierarchical *agglomerative* clustering, which is the mode where all data points (supervoxels) start out as separate clusters. The algorithm then consists of recursively merging the most similar pair of clusters until we are left with one big cluster, in this way producing a hierarchy of nested clusterings (Theodoridis & Koutroumbas, 2008).

In order to identify the most similar pair of clusters in each iteration, the proximity g between all possible pairs of clusters (C_i, C_j) is computed as a function of the set of affinities between pairs of observations in C_i and C_j (Theodoridis & Koutroumbas, 2008). This requires us to define a measure of proximity between data points (vectors) *and* between clusters (sets of vectors). Thus, depending on the chosen measure of affinity between data points and linkage between clusters, the clustering algorithm may lead to completely different clustering results.

Denoting d_{mn} the *dissimilarity* between observation m in cluster C_i and observation n in cluster C_j, we can define average linkage (Hastie, Tibshirani, & Friedman, 2009):

$$g_{CL}(C_i, C_j) = \frac{1}{N_i N_j} \sum_{m \in C_i} \sum_{n \in C_j} d_{mn}, \tag{7}$$

where N_i and N_j are the number of observations in cluster i and j, respectively.

The average linkage measures the average dissimilarity between the clusters and is a compromise between measuring the dissimilarity between the most similar observations (single link) and the most dissimilar observations (complete link) in different clusters. This middle-ground approach is known to be less sensitive to noise and outliers as the measure is based on all observations in the clusters. As for the dissimilarity between observations d_{mn}, we examine three different measures: the Euclidean norm, the Manhattan norm, and the Cosine distance, leading to three different average linkage clustering algorithms.

4.5.3. Spectral clustering

The third and final clustering approach that we consider in this work is spectral clustering, which has become one of the most used clustering algorithms in recent years (Von Luxburg, 2007). It exploits the spectrum of the affinity matrix to perform clustering and is designed for non-convex problems (Hastie et al., 2009).

In spectral clustering, we represent our data in the form of a similarity graph. Each vertex corresponds to an observation (supervoxel) and the edges connecting pairs of vertices are weighted by their pair-wise similarity. The problem of clustering can then be formulated as a graph-cut problem where we are looking for a graph partitioning such that edges between subsets have low weight and edges within subsets have high weight (Von Luxburg, 2007).

To construct the graph, we first have to decide on a similarity measure, and one of the most common choices is the radial basis function (Theodoridis & Koutroumbas, 2008). This is a Gaussian similarity function which encodes the relation between observations in a local neighborhood. The function is given by

$$a(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}{2\sigma^2} \right), \tag{8}$$

where σ is a scaling parameter controlling the width of the neighborhood (Von Luxburg, 2007). The affinity matrix $\boldsymbol{A}$ containing the pairwise similarities $a_{ij} = a(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between all n observations can then be used to define the graph Laplacian:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}, \tag{9}$$

where $\boldsymbol{D}$ is the degree matrix, defined as a diagonal matrix with $d_{ii} = \sum_j a_{ij}$. This particular matrix is known to have an important property that can be used to change the representation of the data (Von Luxburg, 2007): For every vector $\boldsymbol{f} \in \mathbb{R}^n$ we know that

$$\boldsymbol{f}'\boldsymbol{L}\boldsymbol{f} = \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} (f_i - f_j)^2. \tag{10}$$

Eq. (10) can be seen as the eigenvalue decomposition of $\boldsymbol{L}$,

$$\boldsymbol{f}'\boldsymbol{L}\boldsymbol{f} = \lambda, \tag{11}$$

which means that the eigenvector $\boldsymbol{f}$ can be thought of as a fuzzy indication vector, indicating a partitioning of the graph resulting in a cut cost corresponding to its eigenvalue λ. Spectral clustering exploits this result by containing the m eigenvectors of $\boldsymbol{L}$ corresponding to the m smallest eigenvalues in a matrix $\boldsymbol{F}_m \in \mathbb{R}^{n \times m}$ and performs k-means clustering on its rows.

In this paper, we employ spectral clustering with the normalized graph Laplacian, defined by

$$\boldsymbol{L}_n = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2}. \tag{12}$$

This matrix has properties similar to $\boldsymbol{L}$ and is usually preferred for reasons discussed by Von Luxburg, 2007.

5. Experiments and results

In this section, we evaluate the above-mentioned clustering algorithms on the task of lung tumor segmentation. We seek a clustering consisting of two clusters (tumor and non-tumor supervoxels) and we analyze the performance of the different clustering algorithms, as well as the influence of the proposed pre-processing steps. That is, we apply the clustering algorithms to z-normalized Box-Cox transformed data (referred to as transformed data) and evaluate the results quantitatively and qualitatively. For ease of comparison, all quantitative results are summarized in Table 2.

5.1. Evaluation measure

To quantitatively compare the segmentation performance of the clustering methods with the manual delineations, we use the voxel-wise dice score. The dice score, D, between two segmentations A and B is given by

$$D(A, B) = 2 \frac{|A \cap B|}{|A| + |B|}, \tag{13}$$

which means that a dice score of 1 corresponds to a perfect match between the segmentations.

We compute both *overall* dice score (treating the labels of all patients as one segmentation), and *patient-wise* dice score where we report the mean dice score and the standard deviation over all patients.

Table 2
Quantitative results of 2-class clustering on the Box-Cox transformed data.

Method	Mean	SD	OA	#Miss
K-means	0.011	0.015	0.015	8/19
Hierarchical_E	0.288	0.294	0.361	4/19
Hierarchical_M	0.461	0.321	0.657	5/19
Hierarchical_C	0.013	0.017	0.013	7/19
Spectral	0.470	0.308	0.668	5/19

Mean, standard deviation (SD) and overall (OA) dice score. #Miss is the number of tumors completely missed in the segmentation.

5.2. K-means clustering

As is apparent from Table 2, k-means clustering into two clusters leads to poor performance with respect to dice score and we completely miss 8 out of 19 tumors. Taking a closer look at the clustering result by mapping the labels back to the image domain, we see in Fig. 6 that the clusters roughly represent "air/lung" and "tissue/bone/tumor" and not "tumor", "non-tumor". This is not uncommon in the unsupervised setting, where the model is not steered to produce specific classes.

In order to further analyze the performance of k-means clustering, we successively increase the number of desired clusters up to $k = 30$ and determine the best possible performance that can be achieved in each step: If the one "best cluster" out of the produced number of clusters is selected to represent "tumor" and the union of the remaining clusters is treated as "non-tumor", we can compute the maximum dice score for each configuration, shown by the blue curve in Fig. 7. As is apparent from the plot, the dice score increases as the number of clusters increases and we can achieve dice scores higher than 0.7 if we use a high enough k. In practice, the selection of the "best cluster" could be performed by medical experts, but we resort to finding the cluster that gives the best performance using label information (the label information is only used for evaluation). However, the merging of all non-tumor clusters is a non-trivial task and is not feasible in practice.

To examine the effect of the Box-Cox transform, we have also included the results of clustering the non-transformed data (orange curve) in Fig. 7. This curve converges at a lower dice score, which is related to k-means' known problems with elongated clusters and tendency to cluster the data into compact and uniform sized clusters. This experiment further confirms our suspicion that the Box-Cox transformation improves clusterability and we, therefore, consider only the transformed data in the remaining experiments.

5.3. Hierarchical clustering

Table 2 presents the results of clustering transformed data into two clusters using the different hierarchical clustering algorithms. We see that clustering with Manhattan distance measure achieves the highest overall (0.657) and mean (0.461 ± 0.321) dice score. Further, we see that the Euclidean distance measure achieves significantly lower dice scores, but misses only four out of 19 tumors, whereas the Cosine distance measure yields low dice scores and a high number of missed tumors.

Fig. 8 shows the clustering results mapped back to the image domain for five tumor slices in five different patients. It is apparent that both the Euclidean and Manhattan distance measure seem to cluster the data roughly into "tumor" and "non-tumor", whereas the Cosine distance measure leads to a poor segmentation performance, similar to k-means.

5.4. Spectral clustering

In spectral clustering, the affinity matrix is computed using the radial basis function, which is standard practice (Theodoridis & Koutroumbas, 2008). As the scaling parameter σ decides the width of the neighborhoods in which we encode the relations between observations, the parameter choice is critical for the clustering result. Here, we apply the rule of thumb given in Jenssen, 2009:

$$\sigma = 0.15 \cdot \text{median}\{d_{ij}\}_{i,j=1}^{n}, \tag{14}$$

where d_{ij} is the Euclidean distance between feature vector i and j. Nonetheless, in our experiments, we observed that the results are robust to the choice of σ. Since the radial basis function results in a connected graph, the eigenvector corresponding to the smallest eigenvalue (=1) is constant (Von Luxburg, 2007). We, therefore, ignore the smallest one and look at the subsequent eigenvectors. Fig. 9 shows a plot of the 2nd, 3rd, and 4th smallest eigenvectors mapped back to the image domain for one slice in five different patients. From the first row, we see that the cheapest cut (2nd smallest eigenvector) corresponds to (soft) partitioning the graph into two subsets roughly representing "air surrounding the patient" and "patient". Moving on to the third eigenvector (second row in Fig. 9), we see that it appears to detect the tumors. The most common approach in spectral clustering is to use as many eigenvectors in the final k-means step as there are classes in the data. However, as the third eigenvector seems to have the most information about the tumors, we first cluster the data into two clusters based on this eigenvector alone. This yields an overall dice score of 0.668 and a mean dice of 0.470 ± 0.308. No improvements were observed when including the second and fourth eigenvectors.

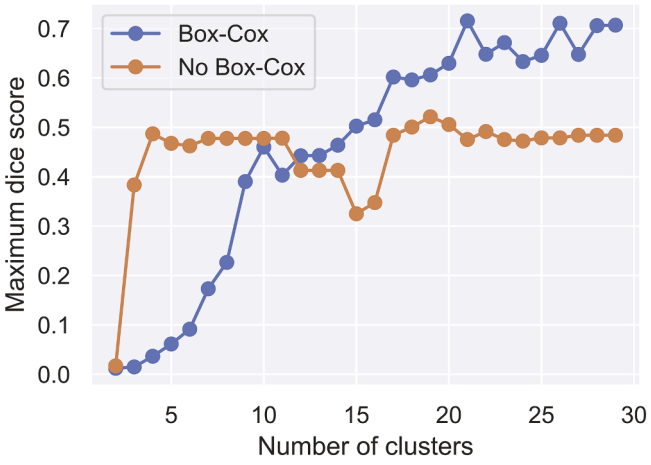

Fig. 7. K-means as function of number of clusters. Performance of k-means clustering with respect to maximum overall (OA) dice score as a function of number of clusters for standard scaled data (orange) and Box-Cox transformed data (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.5. Effect of tumor size

In our data set we have 19 tumors ranging in size from 1944 mm^3 to

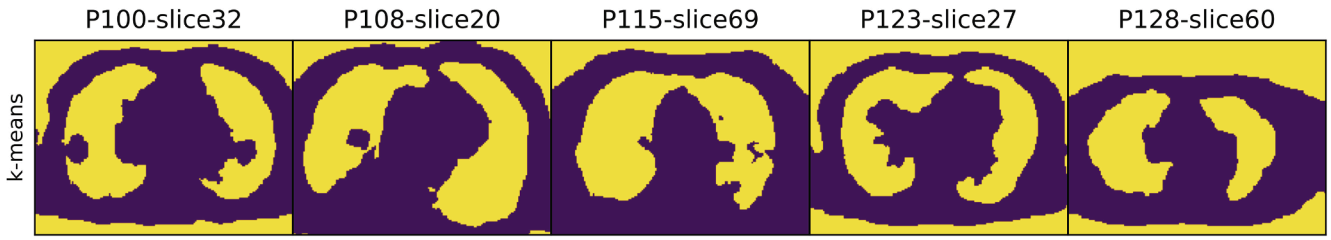

Fig. 6. Visualization of k-means clustering. Result of mapping the k-means $k = 2$ clustering labels back to the image domain and displaying the segmentations for five tumor containing slices in five different patients. The two clusters roughly represent "air/lung" (yellow) and "tissue/bone/tumor" (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

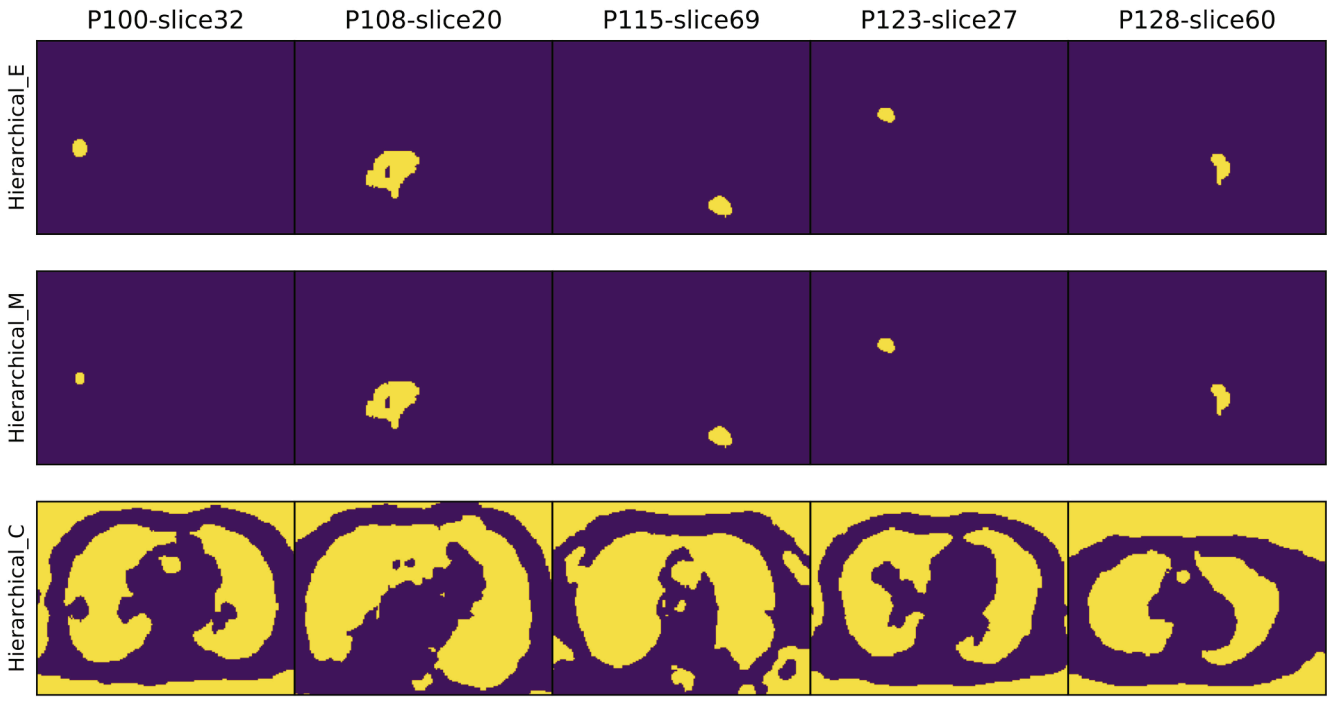

Fig. 8. Visualization of hierarchical clustering. Result of mapping the Euclidean (top), Manhattan (middle) and Cosine (bottom) hierarchical clustering labels back to the image domain and displaying the segmentations for five tumor-containing slices in five different patients. For Euclidean and Manhattan hierarchical clustering, the two clusters roughly represent tumor (yellow) and non-tumor (purple). For Cosine hierarchical clustering, the clusters roughly represent "air/lung" (yellow) and "tissue/bone/tumor" (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

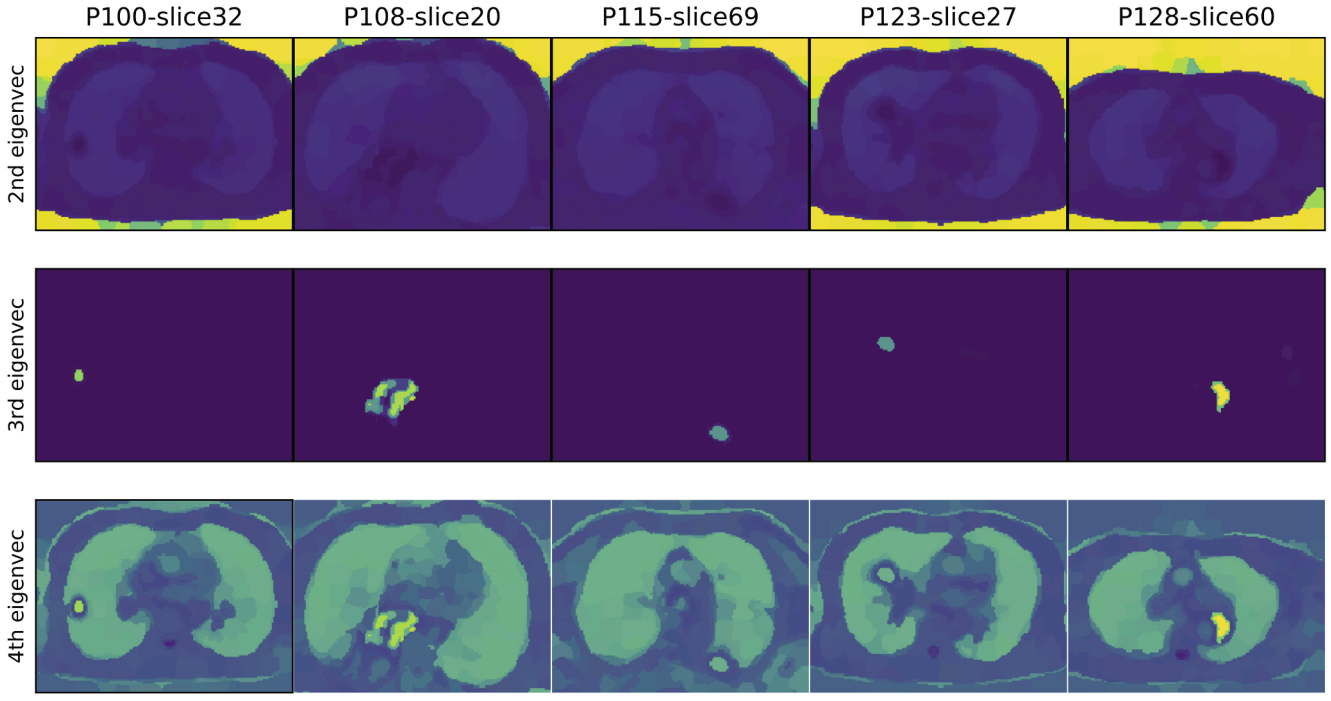

Fig. 9. Visualization of eigenvectors. In spectral clustering, the eigenvectors of the Laplacian can be thought of as fuzzy indication vectors, indicating a partitioning of the graph resulting in a cut cost corresponding to their eigenvalues. This figure shows the eigenvectors corresponding to the 2nd, 3rd and 4th smallest eigenvalue mapped back to pixel space in five tumor slices for five different patients. The 2nd smallest eigenvector (first row) seems to partition the graph into two subsets roughly representing "air surrounding the patient" and "patient". The third eigenvector (second row) appears to pick up on the tumors whereas the fourth eigenvector does not contain additional information about the tumors. Note that we exclude the first eigenvector as it is constant for a connected graph.

$195,744$ mm^3. In order to analyze the effect of tumor size on the clustering dice, we define two thresholds that divide the tumors into three groups; eight small-sized tumor ($< 8000\text{mm}^3$), six medium-sized tumors ($\in [8000, 80,000]\text{mm}^3$) and five large-sized tumors ($> 80,000\text{mm}^3$). The box-plot in Fig. 10 presents the segmentation performance with respect to dice score for these three groups. Note that we omitted k-means

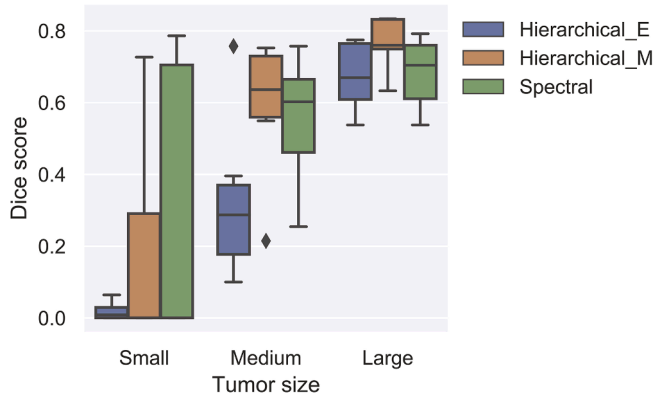

Fig. 10. Effect of tumor size. Grouped box-plot showing the effect of tumor size on the segmentation dice score for Euclidean hierarchical (blue), Manhattan hierarchical (orange) and spectral (green) clustering. The tumors are divided into three groups; small (left), medium (middle) and large (right) size tumors. The mean dice score generally increases while the variance decreases for larger tumors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

clustering and hierarchical clustering with Cosine distance in this comparison as they did not manage to pick up on the tumors. As can be seen in the box-plot, there is a trend towards better dice score with larger tumor size for all methods. The mean dice score generally increases while the variance decreases, yielding more robust predictions for larger tumors. Specifically, we see that none of the algorithms provide reliably high scores for small tumors, but that spectral clustering is able to capture *some* of the small tumors with good dice scores. Further, the difference in segmentation performance (with respect to dice score) among the algorithms decreases with increasing tumor size. Euclidean hierarchical clustering, for instance, gets a dice score close to zero for small tumors but seems to perform equally good as spectral clustering for the large tumors.

5.6. Analysis of segmentation errors

As the dice score treats false negatives and false positives equally, it is also important to evaluate the types of mistakes that each clustering algorithm makes. The bar plot in Fig. 11 presents the number of true positive (TP), false negative (FN) and false positive (FP) voxels in each of the segmentations obtained from the different clustering algorithms. The most interesting result from this analysis is that spectral clustering, which overall achieves the highest dice score, turns out to have the highest number of FNs *and* the lowest number of TPs, providing an overly optimistic segmentation (under-estimation of the tumor volume). Manhattan hierarchical clustering, on the other hand, which, according

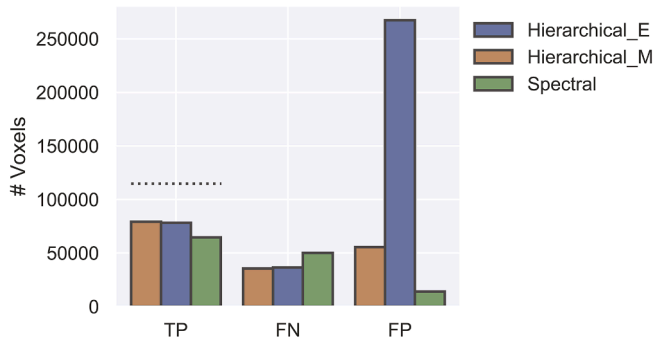

Fig. 11. Segmentation errors. Bar plot presenting the number of TP: true positive (tumor) voxels; FN: false negative (non-tumor) voxels; FP: false positive voxels. The dotted line indicates the total number of TP tumor voxels.

to the dice score has comparable performance, shows to actually have the highest number of TPs and lowest number of FNs, in this way being the method detecting most tumor voxels. Euclidean hierarchical clustering, which has the lowest dice score of the three methods to be compared, has a comparable (to Manhattan hierarchical clustering) number of TPs and FNs, but has too many FPs, resulting in an overly pessimistic segmentation. This means that even though the Euclidean hierarchical clustering fails completely for small tumors according to the dice score, it does *not* necessarily miss the tumors in the segmentation. Note that the sum of the segmentation mask and the ground truth mask in the denominator in Eq. 13 makes the dice score more sensitive to over-segmentation of small tumors compared to larger tumors.

Fig. 12 shows the effects of the different types of mistakes in the image domain. In general, we see that Euclidean hierarchical clustering does not miss the tumors, but tends to over-segment the tumor class by including other organs, resulting in a large number of FPs. Spectral clustering, on the other hand, tends to under-segment the tumor volume, whereas Manhattan hierarchical clustering captures the most tumor voxels without having an alarmingly high number of FP. The two rightmost columns in Fig. 12 show two slices from the same patient at different positions and further illustrate the over-segmentation issue of the hierarchical algorithms (note that slice 5 does not contain tumor voxels).

An interesting observation in P118-slice50 in Fig. 12 is that there is a "hole" in the tumor, which is a common phenomenon for large tumors in PET imaging. The apparent "hole" is most likely caused by necrosis, occurring due to shortage of oxygen supply to the tumor. We see that the Manhattan hierarchical clustering succeeds in exploiting the combined information from both modalities and provides a closed segmentation, as desired.

Regarding the complete misses, we find that Euclidean hierarchical clustering completely misses four out of the nineteen tumors, where two of the misses come from the same patient. Spectral clustering and Manhattan hierarchical clustering miss the same four tumors, in addition to one more (the same one for both). By inspecting the number of overlapping tumor voxels between the two modalities (using the ground truth segmentation masks), we find that the five tumors that are completely missed by the algorithms are among the six tumors with the least number of overlapping tumor voxels across modalities. Moreover, we find that for two of the missed tumors, the maximum SUV within the ground truth segmentation is lower than 1.3, which is a particular low uptake value in PET.

In a clinical setting, the detection of tumors is arguably of utmost importance and over-segmentation is preferred. Only focusing on the number of tumors that are missed completely and therefore choosing the Euclidean hierarchical clustering, would, however, result in a large number of false-positive voxels. This means that the clinicians would be presented with many potential tumors that they have to evaluate, which in turn could lead to real tumors getting missed. Spectral clustering and Manhattan hierarchical clustering, on the other hand, achieve a low number of misses, while at the same time producing a low number of false-positive voxels.

5.7. Benefit of clustering across patients

An important contribution of this paper is the across-patients clustering to improve segmentation performance. In order to quantify the benefit of clustering across patients, we conduct an experiment where we Box-Cox-transform and cluster the supervoxel features of each patient separately. Fig. 13, illustrates the results and visualizes the difference in performance when considering patient-by-patient clustering versus across-patients clustering for the two best performing models. Across-patients clustering achieves considerable improvements for most tumors. For spectral clustering, for instance, the mean dice score increases from 0.169 ± 0.295 patient-by-patient to 0.470 ± 0.308 across-patients. The results for the individual patients can also be found in

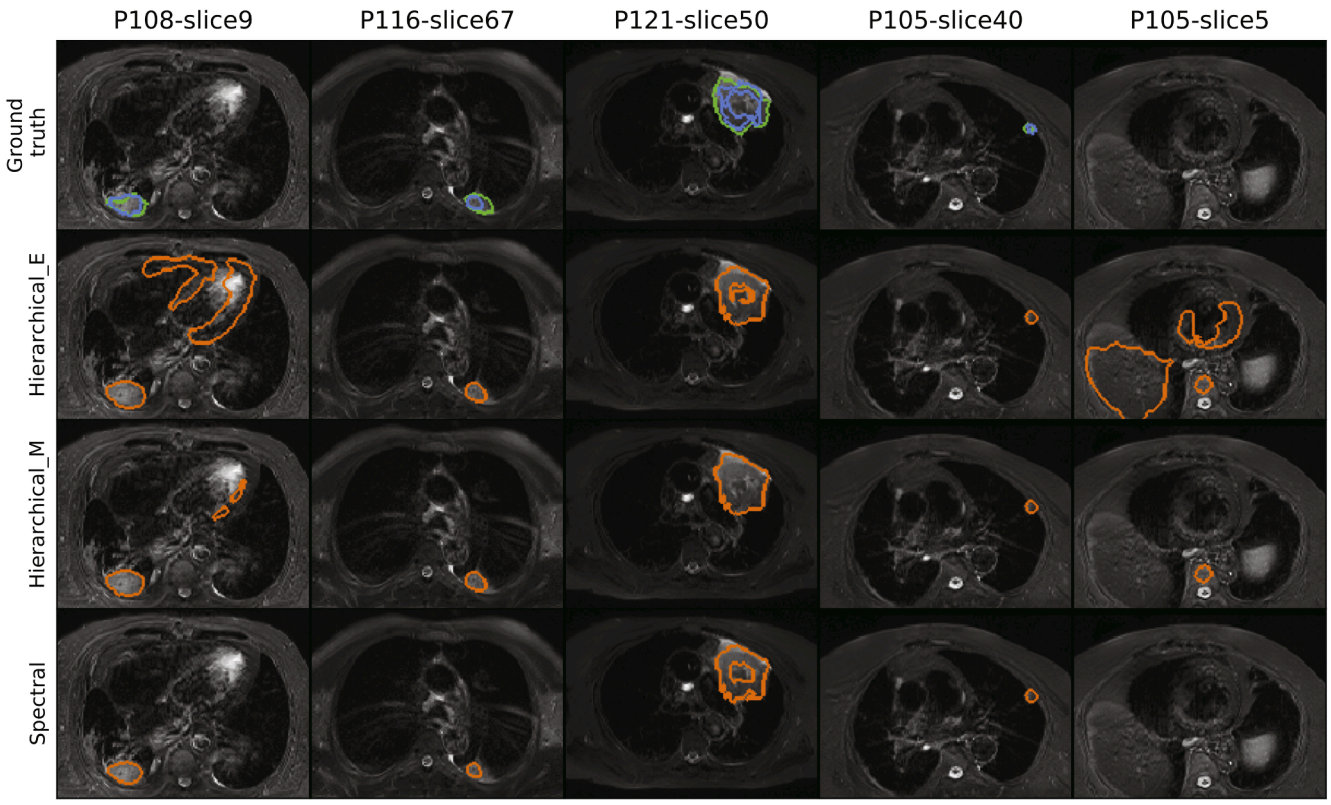

Fig. 12. Segmentation errors in image domain. Overlaying segmentation results (orange) on the MRI for a few selected slices for the different clustering algorithms. The first row shows the ground truth (PET delineation in blue and MRI delineation in green). Note that columns 4–5 show two slices from the same patient at different positions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

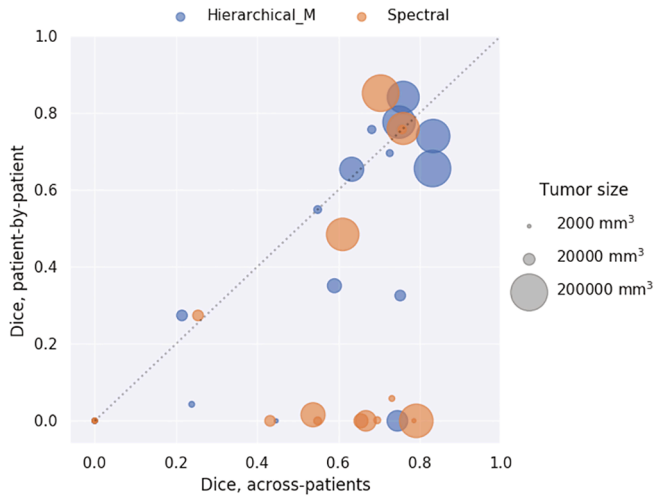

Fig. 13. Patient-by-patient clustering versus across-patients clustering. Scatter plot showing dice score for patient-by-patient clustering against dice score for across-patients clustering. For tumors under the diagonal, the across-patients clustering achieved better dice scores than the patient-by-patient clustering did.

Table 3.

5.8. Noise robustness

Differences in imaging protocols and acquisition conditions can result in variations in the signal-to-nose-ratio. To evaluate the noise robustness of the two best performing methods, we therefore simulate a reduced signal-to-noise-ratio by adding noise to the images before performing the segmentation. Following Jayender, Chikarmane, Jolesz, and Gombos, 2014, we add white Gaussian noise with standard deviation equal to 5%, 15% and 25% of the base intensity of each voxel in the PET and MRI images. Fig. 14 shows an example slice in PET (top) and

Table 3
Patient-by-patient versus across-patients clustering. Dice scores for patient-by-patient (P-by-p) and across-patients (Across-p) clustering for the two best performing models.

	Dice score			
	Spectral		Hierarchical_M	
Patient	P-by-p	Across-p	P-by-p	Across-p
100	0.0	0.549	0.550	0.549
102	0.0	0.0	0.0	0.0
103a	0.0	0.0	0.0	0.0
103b	0.0	0.0	0.0	0.0
104	0.0	0.786	0.0	0.447
105	0.0	0.656	0.351	0.590
108	0.485	0.612	0.777	0.750
109	0.0	0.0	0.0	0.0
112	0.016	0.538	0.654	0.633
114	0.274	0.254	0.274	0.215
115	0.0	0.432	0.326	0.752
116	0.760	0.760	0.841	0.760
118	0.852	0.704	0.656	0.832
121	0.0	0.0	0.0	0.0
123	0.058	0.732	0.043	0.239
125	0.0	0.668	0.0	0.745
128	0.757	0.757	0.757	0.682
129	0.0	0.792	0.740	0.834
131	0.002	0.696	0.696	0.727
Mean	0.169	0.470	0.351	0.461
Std	0.295	0.308	0.328	0.321

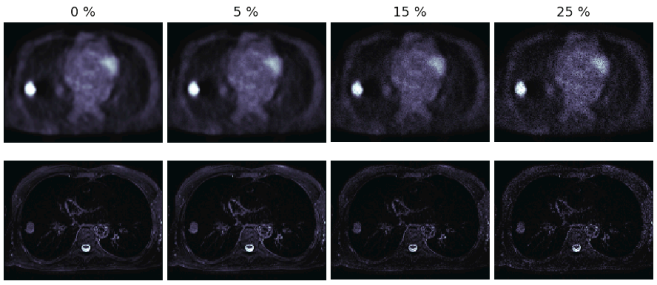

Fig. 14. Noise analysis. Example slices in PET (top) and MRI (bottom) corrupted with white Gaussian noise with standard deviation equal to 5%, 15% and 25% of the voxel intensities.

MRI (bottom) with increasing noise level towards right.

The segmentation maps were computed in the same way as for the original PET/MRI images and compared to the ground truth. The experiments were repeated ten times (with different random seeds in the noise generation) and the results are reported in Fig. 15. Overall, we see that the spectral clustering seems to be robust to 5% and 15% noise, whereas it becomes unstable for 25% noise, with a drop in mean dice and a considerable increase in standard deviation. For Manhattan hierarchical clustering on the other hand, we see that the mean dice score drops significantly and that the standard deviation is high for all noise levels.

6. Outlook and limitations

From our results, we can see that the algorithms detect most of the tumors, but that there still is a relatively high number of tumor voxels that are wrongly segmented. From Fig. 5, it is evident that it is impossible to perfectly cluster tumor voxels and non-tumor voxels into two separate clusters, and the reason for this is twofold. Firstly, some of the supervoxels contain both tumor and non-tumor voxels, and secondly, the chosen features are not able to completely separate tumor-containing and non-tumor-containing supervoxels.

The purity of the supervoxels could in theory be enhanced by increasing the number of supervoxels. However, this comes at the cost of increased computational complexity. In our experiments, a number of 1500 supervoxels per patient was chosen as a middle-ground between supervoxel purity and computational cost. Nevertheless, we can not guarantee that this is the best setting and improved supervoxel generation is left for future work.

The median intensities within the supervoxels are in some cases insufficient to detect a supervoxel as "tumor supervoxel". Other features, such as shape, texture, and histogram features may be able to help the discrimination. Radiomics is a process that extracts large amounts of these types of quantitative image features from medical images and has shown potential to improve tumor classification (Wu et al., 2016). However, how to exploit these large amounts of features in an unsupervised manner is challenging because the variance in the features does not necessarily reflect the classes of interest (tumor and non-tumor). Future efforts should focus on searching for alternative features to improve the discrimination between tumor and non-tumor supervoxels.

Further, there are potential limitations connected to the nature of the data acquisition. Firstly, because of respiratory motion, the PET and MRI can not be assumed perfectly co-registered. An unsupervised co-registration was performed to improve the tumor overlap, but missmatches are still present in the data set. Another limitation is the relatively small sample size.

7. Conclusion

In this paper, we proposed a framework for across-patients supervoxel-based unsupervised lung tumor segmentation in PET/MRI. We

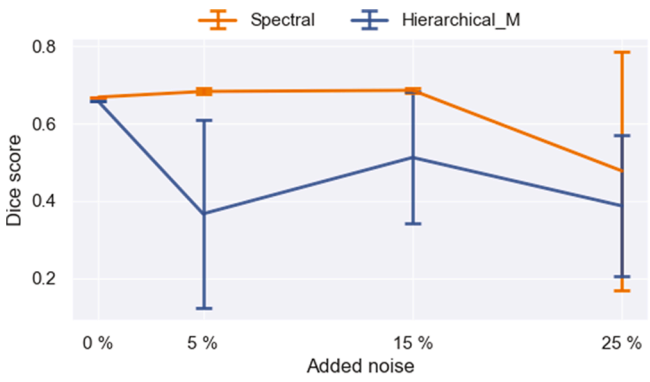

Fig. 15. Noise robustness. Line plot with error-bars showing the effect on the dice score with increasing noise levels for Manhattan hierarchical (blue) and spectral (orange) clustering. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analyzed the segmentation results for several commonly used clustering algorithms within the framework, investigating their advantages and shortcomings. Results demonstrate that spectral clustering and Manhattan hierarchical clustering have the potential to segment tumors in PET/MRI by producing a low number of missed tumors while maintaining a low number of false-positives. In the presence of low to moderate noise levels, spectral clustering provides stable results whereas Manhattan hierarchical clustering seems to be more sensitive to perturbations in the voxel intensities. The results further highlight the importance of performing clustering across patients, and an analysis of the clustering errors illustrates that it is a particular challenge to segment small-size tumors in the presence of imperfect co-registration. Moreover, the framework represents a step towards generic unsupervised tumor segmentation also beyond the lung tumor segmentation task.

CRediT authorship contribution statement

Stine Hansen: Conceptualization, Software, Methodology, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Samuel Kuttner:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Michael Kampffmeyer:** Conceptualization, Methodology, Writing - review & editing, Writing - original draft. **Tom-Vegard Markussen:** Data curation, Writing - review & editing. **Rune Sundset:** Validation, Writing - review & editing. **Silje Kjærnes Øen:** Data curation, Resources, Writing - review & editing. **Live Eikenes:** Data curation, Resources, Writing - review & editing. **Robert Jenssen:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Karl Øyvind Mikalsen and Stian Normann Anfinsen for insightful discussions. This work is supported by the Northern Norway Regional Health Authority (Grant No. HNF1349-17), the Central Norway Regional Health Authority (Grant No. 46056912), and the Norwegian Research Council (Grant No. 303514)

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2010). Slic superpixels. Technical Report.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 2274–2282.

Bagci, U., Udupa, J. K., Mendhiratta, N., Foster, B., Xu, Z., Yao, J., Chen, X., & Mollura, D. J. (2013). Joint segmentation of anatomical and functional images: Applications in quantification of lesions from pet, pet-ct, mri-pet, and mri-pet-ct images. *Medical Image Analysis, 17*, 929–945.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boroojeni, K. G., Amini, M. H., Bahrami, S., Iyengar, S., Sarwat, A. I., & Karabasoglu, O. (2017). A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. *Electric Power Systems Research, 142*, 58–73.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*, 211–243.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*, 394–424.

Caldwell, C. B., Mah, K., Ung, Y. C., Danjoux, C. E., Balogh, J. M., Ganguli, S. N., & Ehrlich, L. E. (2001). Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on ct: the impact of 18fdg-hybrid pet fusion. *International Journal of Radiation Oncology* Biology* Physics, 51*, 923–931.

De Bruijne, M. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis.

Ehman, E. C., Johnson, G. B., Villanueva-Meyer, J. E., Cha, S., Leynes, A. P., Larson, P. E. Z., & Hope, T. A. (2017). Pet/mri: where might it replace pet/ct? *Journal of Magnetic Resonance Imaging, 46*, 1247–1262.

Even, A. J., Reymen, B., La Fontaine, M. D., Das, M., Mottaghy, F. M., Belderbos, J. S., De Ruysscher, D., Lambin, P., & van Elmpt, W. (2017). Clustering of multi-parametric functional imaging to identify high-risk subvolumes in non-small cell lung cancer. *Radiotherapy and Oncology, 125*, 379–384.

Flechsig, P., Mehndiratta, A., Haberkorn, U., Kratochwil, C., & Giesel, F. L. (2015). Pet/mri and pet/ct in lung lesions and thoracic malignancies. In *Seminars in nuclear medicine* (pp. 268–281). Elsevier. Vol. 45.

Foster, B., Bagci, U., Mansoor, A., Xu, Z., & Mollura, D. J. (2014). A review on segmentation of positron emission tomography images. *Computers in Biology and Medicine, 50*, 76–96.

Gordillo, N., Montseny, E., & Sobrevilla, P. (2013). State of the art survey on mri brain tumor segmentation. *Magnetic Resonance Imaging, 31*, 1426–1438.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hossain, M. Z. (2011). The use of box-cox transformation technique in economic and statistical analyses. *Journal of Emerging Trends in Economics and Management Sciences, 2*, 32–39.

Hurkmans, C. W., Borger, J. H., Pieters, B. R., Russell, N. S., Jansen, E. P., & Mijnheer, B. J. (2001). Variability in target volume delineation on ct scans of the breast. *International Journal of Radiation Oncology* Biology* Physics, 50*, 1366–1372.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters, 31*, 651–666.

Jayender, J., Chikarmane, S., Jolesz, F. A., & Gombos, E. (2014). Automatic segmentation of invasive breast carcinomas from dynamic contrast-enhanced mri using time series analysis. *Journal of Magnetic Resonance Imaging, 40*, 467–475.

Jenssen, R. (2009). Kernel entropy component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 847–860.

Ju, W., Xiang, D., Zhang, B., Wang, L., Kopriva, I., & Chen, X. (2015). Random walk and graph cut for co-segmentation of lung tumor on pet-ct images. *IEEE Transactions on Image Processing, 24*, 5854–5867.

Klein, S., Staring, M., Murphy, K., Viergever, M. A., & Pluim, J. P. (2009). Elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging, 29*, 196–205.

Kuttner, S., Lassen, M. L., Øen, S. K., Sundset, R., Beyer, T., & Eikenes, L. (2020). Quantitative pet/mr imaging of lung cancer in the presence of artifacts in the mr-based attenuation correction maps. *Acta Radiologica, 61*, 11–20.

Leibfarth, S., Eckert, F., Welz, S., Siegel, C., Schmidt, H., Schwenzer, N., Zips, D., & Thorwarth, D. (2015). Automatic delineation of tumor volumes by co-segmentation of combined pet/mr data. *Physics in Medicine & Biology, 60*, 5399.

Liu, C.-L., Yin, F., Wang, D.-H., & Wang, Q.-F. (2013). Online and offline handwritten chinese character recognition: Benchmarking on new databases. *Pattern Recognition, 46*, 155–162.

Lucchi, A., Smith, K., Achanta, R., Knott, G., & Fua, P. (2011). Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Transactions on Medical Imaging, 31*, 474–486.

Marstal, K., Berendsen, F., Staring, M., & Klein, S. (2016). Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 134–142).

Moghbel, M., Mashohor, S., Mahmud, R., & Saripan, M. I. B. (2018). Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. *Artificial Intelligence Review, 50*, 497–537.

Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M., MacDonald, J., Thomas, D., Moskaluk, C., Wang, Y., et al. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research, 66*, 7466–7472.

Rayens, W. S., & Srinivasan, C. (1991). Box–cox transformations in the analysis of compositional data. *Journal of Chemometrics, 5*, 227–239.

Roth, H. R., Farag, A., Lu, L., Turkbey, E. B., & Summers, R. M. (2015). Deep convolutional networks for pancreas segmentation in ct imaging. In Medical imaging 2015: Image processing. Vol. 9413. International Society for Optics and Photonics. p. 94131G.

Sauwen, N., Acou, M., Van Cauter, S., Sima, D., Veraart, J., Maes, F., Himmelreich, U., Achten, E., & Van Huffel, S. (2016). Comparison of unsupervised classification methods for brain tumor segmentation using multi-parametric mri. *NeuroImage: Clinical, 12*, 753–764.

Sbei, A., ElBedoui, K., Barhoumi, W., Maksud, P., & Maktouf, C. (2017). Hybrid pet/mri co-segmentation based on joint fuzzy connectedness and graph cut. *Computer Methods and Programs in Biomedicine, 149*, 29–41.

Sbei, A., ElBedoui, K., Barhoumi, W., & Maktouf, C. (2020). Gradient-based generation of intermediate images for heterogeneous tumor segmentation within hybrid pet/mri scans. *Computers in Biology and Medicine, 119*, Article 103669.

Shah, S. K., McNitt-Gray, M. F., Rogers, S. R., Goldin, J. G., Suh, R. D., Sayre, J. W., Petkovska, I., Kim, H. J., & Aberle, D. R. (2005). Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features1. *Academic Radiology, 12*, 1310–1319.

Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T. L., Barrick, T. R., Howe, F. A., & Ye, X. (2017). Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in flair mri. *International Journal of Computer Assisted Radiology and Surgery, 12*, 183–203.

Stoto, M. A., & Emerson, J. D. (1983). Power transformations for data analysis. *Sociological Methodology, 14*, 126–168.

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern recognition* (4th ed.). USA: Academic Press Inc.

Viergever, M. A., Maintz, J. A., Klein, S., Murphy, K., Staring, M., & Pluim, J. P. (2016). A survey of medical image registration – Under review.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*, 395–416.

Vos, P., Barentsz, J., Karssemeijer, N., & Huisman, H. (2012). Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Physics in Medicine & Biology, 57*, 1527.

Wadhwa, A., Bhardwaj, A., & Verma, V. S. (2019). A review on brain tumor segmentation of mri images. *Magnetic Resonance Imaging, 61*, 247–259.

Wu, J., Gensheimer, M. F., Dong, X., Rubin, D. L., Napel, S., Diehn, M., Loo, B. W., Jr, & Li, R. (2016). Robust intratumor partitioning to identify high-risk subregions in lung cancer: a pilot study. *International Journal of Radiation Oncology* Biology* Physics, 95*, 1504–1512.

Wu, W., Parmar, C., Grossmann, P., Quackenbush, J., Lambin, P., Bussink, J., Mak, R., & Aerts, H. J. (2016). Exploratory study to identify radiomics classifiers for lung cancer histology. *Frontiers in Oncology, 6*, 71.

Xu, Z., Bagci, U., Udupa, J. K., & Mollura, D. J. (2015). Fuzzy connectedness image co-segmentation for hybridpet/mri and pet/ct scans. In *Computational Methods for Molecular Imaging* (pp. 15–24). Springer.