SEGMENTATION OF DYNAMIC TOTAL-BODY [^{18}F]-FDG PET IMAGES USING UNSUPERVISED CLUSTERING

Maria K. Jaakkola*[1], Maria Rantala*[1], Anna Jalo[2,3], Teemu Saari[1,4], Jaakko Hentilä[1], Jatta S. Helin[2,3], Tuuli A. Nissinen[2,3], Olli Eskola[5], Johan Rajander[6], Kirsi A. Virtanen[1,4], Jarna C. Hannukainen[1], Francisco López-Picón[1,2,3], and Riku Klén[1,4]

Abstract. Clustering time activity curves of PET images has been used to separate clinically relevant areas of the brain or tumours. However, such applications on segmenting PET images in organ level are much less studied due to the available total-body data being limited to animal studies. Now the new PET scanners providing the opportunity to acquire total-body PET scans also from humans are becoming more common, which opens plenty of new clinically interesting opportunities. Therefore, organ level segmentation of PET images has important applications, yet it lacks sufficient research. In this proof of concept study, we evaluate if the previously used segmentation approaches are suitable for segmenting dynamic human total-body PET images in organ level. First we excluded methods that were computationally too demanding for the large datasets from human total-body PET scanners. This criteria filtered out most of the commonly used approaches. Then we tested how well the usable methods segment PET images in organ level, highlight the best approaches together with their limitations, and discuss how further research could tackle the observed shortcomings. We used 40 total-body [^{18}F]fluorodeoxyglucose PET images of rats to mimic the coming large human PET images and a few actual human total-body images to ensure that our conclusions from the rat data generalise to the human data. Our results show that independent component analysis has weaker performance than the other two computationally usable approaches and that certain organs are easier to segment than others. While Gaussian mixture model performed sufficiently, it was by far the slowest one among the tested methods, making k-means combined with principal component analysis the most promising candidate for further development.

Key points

- Majority of the considered clustering methods were computationally too intense even for our total-body rat images. The coming total-body human images are 10-fold bigger.
- Heterogeneous VOIs like brain require more sophisticated segmentation method than the basic clustering without further pre-processing tested here.
- PCA combined with k-means had the best balance between performance and running speed among the tested methods, but without further preprocessing, it is not accurate enough for practical applications.

Affiliations. [1] Turku PET Centre, University of Turku, Turku, Finland
[2] MediCity Research Laboratory, University of Turku, Turku, Finland
[3] PET Preclinical Laboratory, Turku PET Centre, University of Turku, Turku Finland

[4] Turku PET Centre, Turku University Hospital, Turku, Finland
[5] Radiopharmaceutical Chemistry Laboratory, Turku PET Centre, University of Turku
[6] Accelerator Laboratory, Turku PET Centre, Åbo Akademi University, Turku, Finland

Funding. Research of both first authors was supported by donation funds of Faculty of Medicine at University of Turku. JCH reports funding from The Academy of Finland (decision 317332), the Finnish Cultural Foundation, the Finnish Cultural Foundation Varsinais-Suomi Regional Fund, the Diabetes Research Foundation of Finland, and State Research Funding/Hospital District of Southwest Finland. KAV report funding from The Academy of Finland (decision 343410), Sigrid Juselius Foundation and State Research Funding/Hospital District of Southwest Finland. JH reports funding from The Finnish Cultural Foundation Varsinais-Suomi Regional Fund.

Data availability. The codes used in this study are available from Github page `https://github.com/rklen/Dynamic_FDG_PET_clustering`. The example data used in this study have not been published at the time of writing. The reference number of the ethical committee decision related to the human data is 14/1801/2022 (Hospital District of South-Western Finland).

1. Introduction

Dynamic positron emission tomography (PET) studies provide a series of three-dimensional (3D) images over time. In PET imaging, tracer with radioactive component is usually injected into the body and the spreading of the tracer is followed by recording the gamma radiation resulting from the radioactive decay. Therefore, the dynamic PET data comprises of tracer radioactivity over time in each 3D pixel called voxel. The vectors containing the tracer activity over time are called time-activity curves (TACs) and each voxel has its own TAC. To study the activity of a specific tissue, the images are segmented and one or more volumes of interest (VOIs) are taken into further analysis. A VOI is a 3D version of the commonly used term 'region of interest' (ROI). Traditionally, the VOIs have been manually segmented based on additional magnetic resonance imaging (MRI) or computed tomography (CT) scan, but those imaging protocols provide structural information rather than regions with distinct distribution of the used PET tracer. Therefore, structurally similar, but functionally different areas will be ignored in MRI or CT based segmentation, which might lose clinically important information. In addition, manual segmentation is user-dependent and time consuming [1]. An automatic segmentation method for PET images could avoid some of these problems and make PET image analysis more efficient and accurate.

One of the possible ways to automatically segment PET images is to cluster TACs and this technique with multiple different variations has been utilised in previous studies. For instance, Wong *et al.* [2] proposed a method similar to k-means clustering, which was based on a weighted least-square distance, whereas Vogel *et al.* [3] utilised

BASC-clustering combining bootstrapping and k-means with sophisticated pre- and post-processing. In addition, spectral clustering has been utilised in several studies as Mouysset *et al.* [4] segmented dynamic PET images with kinetic spectral clustering method and Zbib *et al.* [5] suggested using spectral clustering with automatically inferred parameters. Several other classical clustering approaches have been applied as well. For example, Chen *et al.* [6] used an expectation maximisation based clustering algorithm with Markov Random Field models, Guo *et al.* [7] proposed a hierarchical linkage based algorithm for clustering pixels, and Ashburner *et al.* [8] introduced a modified mixture model algorithm, which clusters based on the shapes of the TACs rather than their absolute scaling. In addition, Parker and Feng [9] used a segmentation algorithm based on Mumford-Shah energy minimisation and pre-processed the data with PCA. While clustering is the most common approach for this task on dynamic images, also contouring [10, 11] and level set analysis [12] has been applied.

However, PET image segmentation clustering methods described above have mainly focused on different areas of the brain, likely because those applications can eventually have clinical relevance to the health care. Indeed, among the examples described above, all but the study by Wong and colleagues [2] focus on clustering VOIs or ROIs within brains. As only a small part of human body fits into a traditional PET scanner, total-body PET images of human have not been available for research. However, new PET scanners, which enable total-body pet scans, have become more and more popular and they have a great potential for novel scientific discoveries and medical applications. Thus, motivation to identify specific organ level VOIs based on TACs has suddenly exploded, yet the previous studies have mainly focused on a single organ or on a very limited area of the body.

The aim of this study was to test if previously used classical clustering methods can be used to segment dynamic total-body PET images. A clustering method needs to a) be computationally light and b) provide biologically meaningful segments in order to be useful tool for automatic segmentation of large modern PET images. The first criteria is important as the modern total-body scans generate tens of millions of TACs per image and the number is likely to further increase as the scanners become more and more accurate and comprehensive. According to our tests, only two clustering approaches, k-means and Gaussian mixture models (GMM), were fast and light enough to analyse the modern images within reasonable time. Thus we had to exclude some previously successfully used clustering methods, such as spectral clustering, from further analyses. We tested these two clustering methods' capability to identify VOIs representing heart, brain, and kidneys defined from each analysed PET image. Two different preprocessing approaches, namely principal component analysis (PCA) and independent component analysis (ICA), were tested and we also briefly investigated if the defined organ specific VOIs consist of multiple clusters.

2. Materials and methods

2.1. **Data.** In this study, we used total-body dynamic 3D PET images of 40 male Sprague-Dawley rats (mean weight 487 grams with standard deviation of 90 grams), which were scanned at Turku PET Centre using Inveon MM Platform by Siemens Molecular Imaging. The scans contained 50 time frames ($30 \cdot 10$s, $15 \cdot 60$s, $5 \cdot 300$s). The used tracer was [^{18}F]fluorodeoxyglucose ([^{18}F]-FDG) and the average dose of it was 20.8 MBq with standard deviation of 1.19 MBq. At the time of imaging, the rats were 8-35 weeks old, under different diets and exercise programs (but fasting for the last 4 hours), and they were in insulin clamp. The images contained $128 \cdot 128 \cdot 159 = 2605056$ voxels making them approximately ten-fold smaller than the human total-body scans. Notably, all the voxels were not clustered. Those voxels with mean radioactivity over time below the average mean activity over all voxels were defined as background and were not used for clustering analysis. Thus on average 20.4% of voxels were used in clustering (Supplementary Figure 1). Detailed listing of numbers and proportions of voxels used for clustering is available at `https://github.com/rklen/Dynamic_FDG_PET_clustering`. Because the images were clustered independently from each other, the activity values in TACs were used as such without further scaling.

In addition to the rat images, we clustered three dynamic total-body PET images of humans (1 female and 2 male). The used tracer was the same [^{18}F]-FDG as in the rat data. All human subjects were healthy, as determined by routine laboratory tests, oral glucose tolerance test, and medical examination prior to inclusion. The weights of the three individuals were 79.9 kg, 70.2 kg, and 64.3 kg, and the injected tracer doses were 100.71 MBq, 106.55 MBq and 106.93 MBq, respectively. The scan was started immediately after bolus injection of [^{18}F]-FDG. The scans had 13 time frames ($1 \cdot 60$s, $6 \cdot 30$s, $1 \cdot 60$s, $3 \cdot 300$s, $2 \cdot 600$s). These images contained $440 \cdot 440 \cdot 354 = 68534400$ voxels, and 8.7-10.0 % of them were used for clustering. The selection criteria for the voxels to be clustered was the same mean-based threshold than with the rat images.

2.2. **Clustering methods.** Multiple interesting segmentation approaches were considered for this study, but eventually excluded due to too long running time or memory consumption for practical applications on modern big images. The excluded clustering methods include spectral clustering, agglomerative clustering (i.e. hierarchical clustering), affinity propagation, HDBSCAN, and linkage_vector from Python package fastcluster. Besides different clustering methods, also other types of segmentation approaches were considered, but excluded due to computational reasons. Such excluded methods include slic, watershed, random walker, morphological geodesic active contouring, and morphological active contours without edges. Only two clustering methods, k-means and Gaussian mixture model, were usable with our test images.

K-means clustering with principal component analysis (PCA). Principal component analysis (PCA) is a popular unsupervised multivariate analysis technique, which can reduce high dimensional data into a smaller number of components. These components are called Principal Components and they explain most of the variance in the original

multivariate data with the least loss of information. PCA uses a transformation matrix formed from the eigenvectors of the covariance matrix of the vector data to decorrelate the data. In this study, the PCA was performed on the TAC data with the Python function PCA, which implements the probabilistic PCA model of Tipping and Bishop [13]. Because we dealt with a large number of voxels, the singular value decomposition (SVD) solver in function PCA runs a randomised SVD by the method of Halko *et al.* [14] and Martinsson *et al.* [15].

K-means clustering with independent component analysis (ICA). Independent component analysis (ICA) is a method that can be used to find a linear representation of non-Gaussian data so that the components are statistically as independent as possible. This representation can capture the essential structure of the data in many applications, such as feature extraction and signal separation. In this study, ICA was performed on the TAC data with the Python function FastICA, whose implementation bases on Hyvärinen *et al.* [16].

K-means is a method that aims to separate n samples into k clusters based on the distance between the samples and the initialised cluster centroids. The k-means implementation used in this study was the Python function k-means, which minimises within-cluster sum-of-squares called inertia:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2), \tag{1}$$

where x_i is the ith sample and μ_j is the jth cluster centroid from cluster C. K-means assigns each sample to the closest centroid using the inertia criterion (1). After that, k-means calculates new centroids by taking the mean value of all the samples in each cluster. This process is repeated until the difference between the old and the new centroids is less than a threshold. K-means clustering uses random centroid initialisation, thus repeated replications may yield different results. An additional drawback of k-means is that the number of clusters must be defined in advance. PCA and ICA can be combined with k-means (or any other clustering approach) by preprocessing the data with either of them prior to the actual clustering. Such preprocessing reduces the dimension of each data point to be clustered to the chosen number of independent or principal components.

Gaussian mixture model (GMM). The third tested method is a probabilistic model called Gaussian mixture model (GMM). Mixture models generally assume that any distribution can be conveyed as a mixture of distributions of some known parametrisation. In the case of GMM, the data is assumed to have been generated by more than one Gaussian distributions. Clustering happens by assigning a label to the Gaussian that contributes the largest probability. A GMM is solved using Expectation-maximisation (EM) algorithm, which calculates maximum likelihood estimates of the parameters of each component distribution. The EM algorithm is guaranteed to converge, so a locally optimal solution is always obtained. The GMM algorithm implementation used in this

study was a Python function GaussianMixture. All the utilised clustering related functions were from Python library sklearn associated with version 1.1.2 of scikit-learn, and the used Python version was 3.10.

2.3. **Test design.** In order to validate the clustering results, we needed known VOIs for heart, brain, and kidneys from each image. Those were obtained by manually drawing the organs with Carimas software (version 2.10) from the 40 PET images utilised in this study. Importantly, the gold standard to define VOIs is to obtain them directly from the PET images based on tracer activity instead of corresponding CT images visualising static structure. Thus we only used PET images to obtain VOIs. The rounded average numbers of voxels with their standard deviations (in parenthesis) of heart, brain and kidneys were 2834 (757), 4846 (888), and 3054 (1634), respectively. The corresponding mean volumes were 1359, 2325, and 1465 mm^3. The average TACs of these VOIs are visualised in Supplementary Figure 2. Notably, the kidneys did not always fully fit into the scanned area, which explains the high variation in their sizes.

The performance of each method for each organ was evaluated using Jaccard index, which describes how well the cluster and the manual segmentation of the analysed VOI overlap (value 1 indicates perfect overlap and 0 no overlap at all). Jaccard index is formally defined as

$$Jaccard = \frac{|A \cap B|}{|A \cup B|}, \tag{2}$$

where A is the set of voxels belonging to the cluster representing the analysed organ, and B is the set of voxels manually segmented to form the VOI for the organ. Notations $|A \cap B|$ and $|A \cup B|$ indicate the number of voxels in the intersection and union of A and B, respectively. For each organ, the cluster with the highest Jaccard index was selected to represent it. The statistical significance between the Jaccard indices from two methods was calculated using two-tailed Wilcoxon signed-rank test with the conventional p-value cutoff of 0.05 for significance.

As all of the tested methods require a number of clusters to use, we tested cluster numbers 5, 10, 15, ..., 45 for all methods with the 10 images, which were acquired first. Then, for each method separately, we selected the average of the best (defined as the highest Jaccard index) organ specific cluster numbers for the further analyses with the remaining 30 images.

A computer with 16GB of RAM and Intel Pentium Gold processor G6405T (CPU 3.50GHz) was used in the analyses and the reported running times are obtained with it.

3. Results

3.1. **Number of clusters.** First, the number of clusters for each method was defined. The tested cluster numbers were between 5 and 45, and 10 images were analysed. Our results show (Figure 1) that PCA had the best performance with the largest number of clusters and GMM's performance was at its best with the smallest number of clusters.

For further analyses, we used 25 clusters for PCA, 22 for ICA, and 18 for GMM. Notably, GMM produced errors with some analyses, particularly with large cluster numbers, so those results are missing from Figure 1.

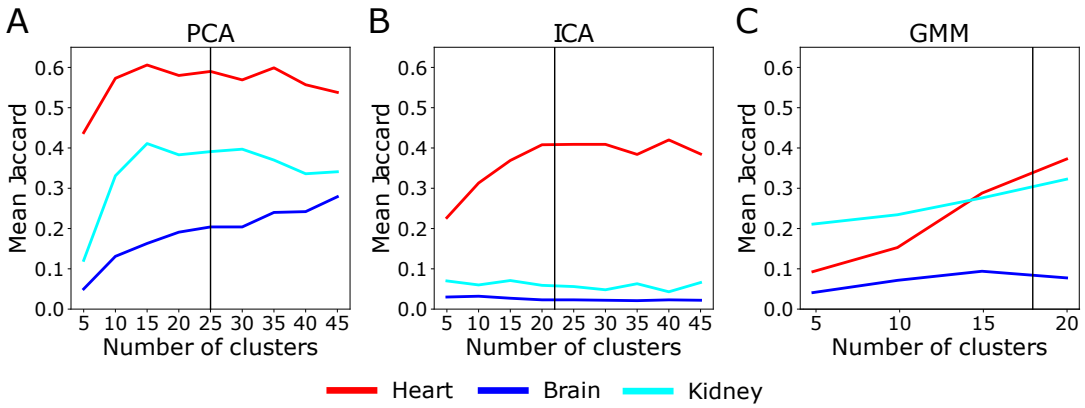

FIGURE 1. Mean Jaccard index over 10 images (y-axis) as a function of the number of clusters (x-axis) for (A) k-means with principal component analysis (PCA), (B) k-means with independent component analysis (ICA), and (C) Gaussian mixture models (GMM). The accuracy represented by mean Jaccard index is calculated separately for heart (red lines), brain (blue lines), and kidney (cyan lines). The selected number of clusters for further analyses is indicated with a black vertical line.

3.2. **Clustering results.** Next, the remaining 30 images were clustered using the method specific cluster numbers defined in the previous step. ICA combined with k-means failed systematically to identify brain and kidney, and also with the VOI representing heart, its performance was poorer than those of PCA and GMM (Figure 2A). All Jaccard indices are available in numeric format at `https://github.com/rklen/Dynamic_FDG_PET_clustering`. The performance of GMM comes closer to PCA than expected based on the 10 images analysed earlier as the mean Jaccard indices differed no more than 0.08 units in any of the organs (Figure 2B). Notably, in the context of the heart, PCA has steady performance and GMM has higher variation over the images, whereas in the context of the kidneys, GMM has steady performance and PCA has considerable variation in its performance (Figure 2A and Figure 2C). The differences between the methods were statistically significant in most organs, except in comparisons PCA vs GMM for kidneys and ICA vs GMM for heart (Table 1).

Figure 3 illustrates the manually segmented VOIs together with clusters representing them from all the tested methods in one example image. The results for brain are poor because it is clustered with wide areas all over the rat with all the tested methods. The same observation can be made from all the images (Supplementary Figure 3). Interestingly, all of the three methods systematically cluster the spinal cord together with the

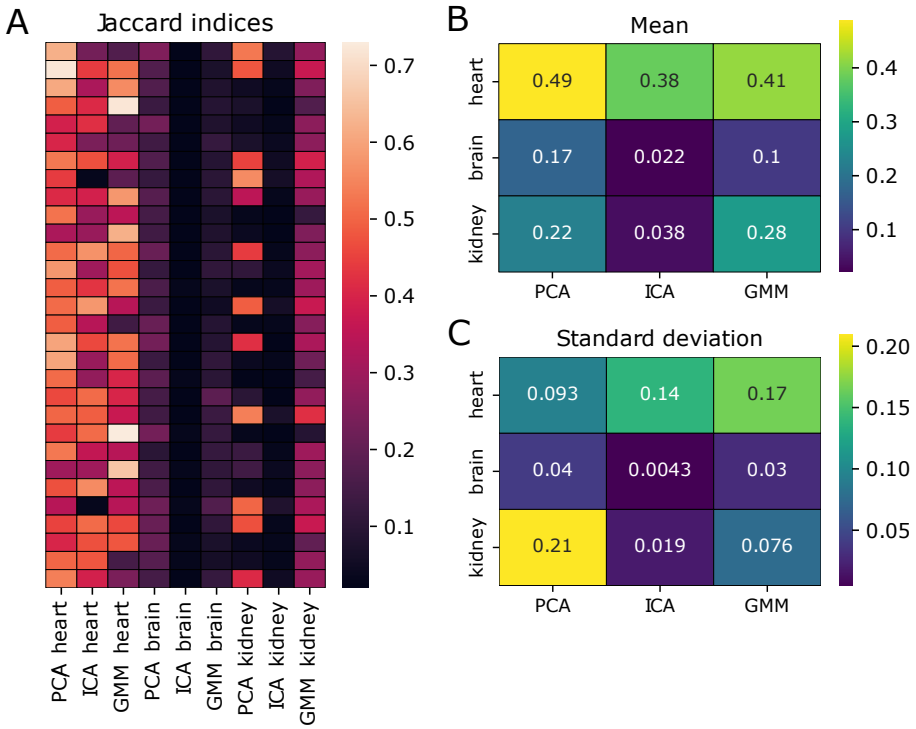

FIGURE 2. (A) Jaccard indices encoded by colour for different organs and methods (cols) in each 30 test image (rows). The number of clusters was 25 for PCA, 22 for ICA, and 18 for GMM. Subfigures (B) and (C) indicate the mean and standard deviation of the Jaccard indices over the images for all organs (rows) and methods (columns).

	heart	brain	kidneys
PCA vs ICA	0.002	0.000	0.000
PCA vs GMM	0.038	0.000	0.064
ICA vs GMM	0.339	0.000	0.000

TABLE 1. P-values from Wilcoxon's test for all organs (columns) and method comparisons (rows) rounded to three decimals.

brain (Supplementary Figure 3), which is reasonable considering their functional connection with nervous system. Other intriguing observations common for all the tested methods are that 1) the heart tends to cluster together with harderian glands (heart-shaped areas behind eyes, Figure 3, Supplementary Figure 4), and 2) brain cluster often includes also brown fat in neck area (Figure 3, Supplementary Figure 3). High radioactivity of harderian glands and brown fat with [^{18}F]-FDG has been reported in the literature [17, 18]. GMM is less prone to cluster heart with the harderian glands than the two

k-means approaches, but this kind of clustering occurred also with GMM in about half of the test images (Supplementary Figure 4). In some cases, kidneys or heart consist of multiple clusters not spanning over other areas (Supplementary Figure 5). This phenomenon lowers the Jaccard indices artificially in those cases, yet the identified sub VOIs with distinct TAC profiles can be clinically interesting findings, though hard to validate. See section Discussion for further debate about the limitations of the validation.

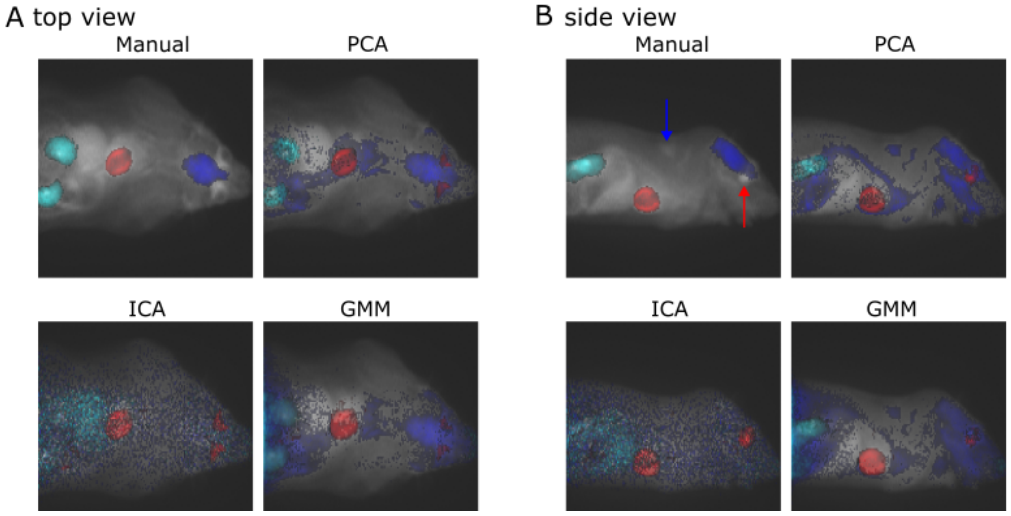

FIGURE 3. An example image with VOIs identified either manually (top-left panel), or using PCA (top-right panel), ICA (bottom-left panel) or GMM (bottom-right panel). The image is flattened into 2D from (A) top or (B) side view by taking the sum of voxel values over the dimension not visible in the 2D representation. The underlying gray scale image is the PET image and the VOIs are plotted on the top with VOIs corresponding to heart, brain and kidneys encoded with red, blue, and cyan, respectively. The harderian glands are highlighted with a red arrow and brown fat with a blue arrow from the side view image (B) with the manual segmentation.

PCA failed to segment kidneys in 18 test images (Jaccard < 0.15). In some instances, kidneys or parts of them were clustered with heart, as illustrated in Figure 4 together with a successful example. Another common issue was kidneys clustered together with larger abdomen area including intestines (Supplementary Figure 6). Supplementary Figure 7 shows that PCA detected kidneys more successfully from the images, where the kidneys fitted better into the scanned part of a body.

3.3. **Human images.** We also tested the three clustering approaches on human images. While the sample size of three dynamic 3D images prevents any statistical analysis, these examples are in-line with our observations from the rat data. PCA appears as the most

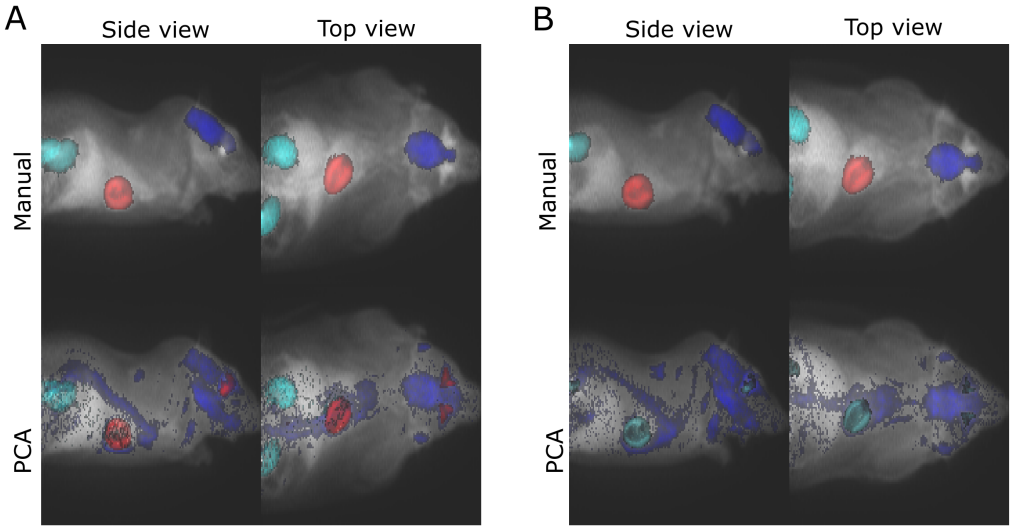

FIGURE 4. In image (A) PCA succeeded to identify kidneys (Jaccard 0.54), whereas in image (B) it failed in it with Jaccard index 0.05. In both subfigures, the two upper panels include manual segmentations of the VOIs (red for heart, blue for brain, and cyan for kidneys), and the bottom panels visualise the PCA clusters representing corresponding VOIs (again, red for heart, blue for brain, and cyan for kidneys). The underlying gray scale image is the PET image. For visualisation purposes, the images were flattened into 2D format by taking the sum over the dimension missing from the 2D representation.

accurate approach, while ICA provides the least intuitive clustering (Figure 5, Supplementary Figure 8).

3.4. **Processing time.** Among the tested methods, GMM was 23 times slower than the fastest approach ICA with a mean processing time of approximately 22 minutes per image. Unsurprisingly, the k-mean based approaches were more similar to each other with mean clustering times of 79 seconds (PCA) and 57 seconds (ICA). The standard deviations of the running times for PCA, ICA, and GMM were 14 seconds, 7 seconds, and 5 minutes, respectively. The full list of processing times is available at `https://github.com/rklen/Dynamic_FDG_PET_clustering`. Notably, with the three human images, PCA was faster than ICA and GMM was only about 10-fold slower than the k-means approaches. The mean running times (rounded to full minutes) with human images for PCA, ICA, and GMM where 7 minutes, 10 minutes, and 73 minutes, respectively.

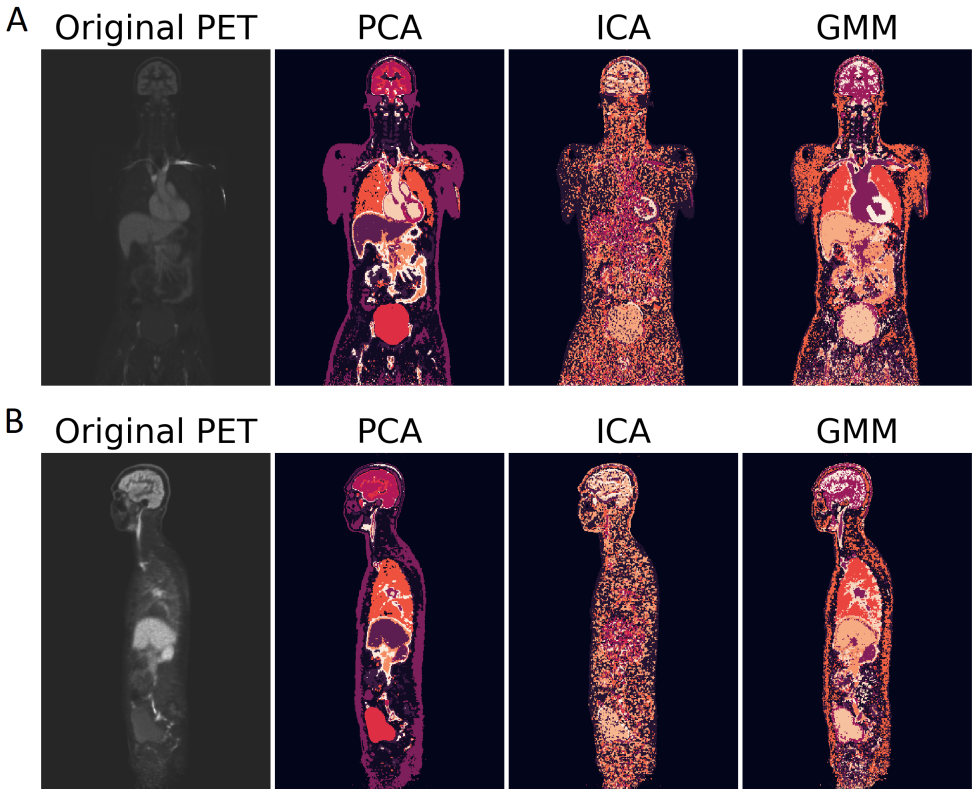

FIGURE 5. (A) Front and (B) side view of the clusters from the middle of a dynamic 3D human total-body PET image. The visualised slice in both subfigures is the slice number 200 out of 400. Notably, in the original PET image (first panel), the voxel activities are sums over time and brighter colour indicate higher activity, whereas in the cluster images (the remaining three right-most panels) different colours indicate different clusters in no particular order.

4. Discussion and conclusions

Here we have tested three clustering approaches to identify predefined VOIs from dynamic total-body PET images based on voxel specific TACs. Data from rats and humans were used, but the specie did not have high impact on our conclusions. Our results show that none of the tested methods is accurate enough for practical applications without further method development, but at least the tested methods were still computationally usable also with the large human images.

The ultimate purpose of studying the clustering of PET images is to aid the process of quantification of PET data. In the end, the clustering could be used to automatically

segment PET images into VOIs that could be used in kinetic modelling. This is an early investigation on clustering total-body images and many new applications are still possible. This kind of total-body clustering could possibly find connections between TACs of different tissue types, which could give insight on different problems in medicine. For example, total-body clustering could be experimented in finding metastatic tumors and otherwise just different tissues that have the same uptake of the tracer.

We suspect that among the tested VOIs, heart is the easiest one to segment with clustering due to the high and relatively homogeneous maximum activity values of its TACs (Supplementary Figure 2). Similarly, among the tested VOIs, the brain was the most difficult to detect with clustering and it has functionally heterogeneous areas and lower radioactivities than the other two organs. This suggests that the overall activities have an important role in the clustering process. Notably, PCA had unstable performance with kidneys depending on how well the kidneys fitted into the scanned area, but it is unclear why only PCA was sensitive to this. However, it is also probable that whole organs may not fully fit into human total-body scans. In this study, we used PET images containing [^{18}F]-FDG tracer, which is not commonly combined with dynamic imaging in clinical usage. However, for our technical evaluation this is not a problem, as clustering should not be sensitive to the selected tracer as long as the manual segmentation used as gold standard is drawn from the same images used for clustering.

One possible approach to prevent entirely unrelated parts from clustering together is to split the images into smaller subsections or otherwise limit the clustering to nearby voxels. Alternatively, splitting the unconnected clusters into separate ones could be automatically done with connected component analysis as a post-processing step. Besides preventing scattered clusters, this would also allow using slightly more computationally demanding methods and, in the case of dividing the image into subimages, different number of clusters could be used in different image regions. The last benefit seems important, as in this study, the brain benefited from larger number of clusters than heart or kidney when PCA approach was used. Particularly complex and heterogeneous VOIs like brain could benefit from this approach. The cons of looking only for connected clusters include, for example, difficulties with associating metastases with the primary tumour, and pair organs like kidneys not clustering together. A post-processing step where tiny (particularly single voxel) unconnected parts of a bigger cluster would be removed from their original cluster and merged into the attached cluster that most resembles them could be a reasonable compromise with little to loose between the pros and cons of looking for connected clusters.

Other interesting topics for further research include the automatic definition of number of clusters, the possibility to combine nearly located clusters if they resemble each other, pre-processing of the imaging data, and the use of kinetic models. The optimal number of clusters is likely heavily dependent on the size of the image and the utilised tracer, so we do not expect that the cluster number defined in this study generalises well to other studies. Different pre-processing approaches have the advantage of typically being computationally lighter than sophisticated clustering methods. For example, noise reduction by using mean activities over the voxel neighbourhood instead of the measured activities of a voxel as

such could improve the clustering accuracy particularly for ICA, which performed poorly with brain and kidneys, possibly due to being sensitive to noise. Also scaling the TACs so that the clustering is done based on their shapes rather than absolute activity levels is an interesting option already applied in the literature [19, 8]. It could affect which VOIs are easy and which ones difficult to detect with clustering. Also considering issues common to PET data, such as spill-over effect from closely located organs and partial volume effect, could improve the results.

Notably, the original motivation to use GMM arises from the voxels of a VOI in a static 3D PET image typically having a distribution easily modelled with a mixture of Gaussian distribution. For dynamic images where the distributions are fitted into TACs this is not necessarily the case. Thus, GMM might be better suited for segmenting static 3D PET images than dynamic 4D PET images.

Due to the differences in the scanned area and evaluation metrics, it is difficult to compare our results to the results obtained in the previous studies. However, higher accuracies have been achieved in the brain region level with sophisticated yet computationally demanding methods [4, 20]. This highlights the need for a better clustering approach designed for very large images.

Different machine learning approaches are fast to use after the training has been done, which make them an appealing approach for segmenting very large total-body images, though the available training data is still very limited. However, the scope of their segmentation is different from that of in-situ approaches like clustering. Learning-based machine learning methods can be used to segment VOIs that are expected to be present in all images (such as full organs), whereas in-situ approaches are suited to segment VOIs that differ from the rest of the image in the particular image under study (e.g. part of liver behaving unexpectedly). The research question defines which type of segmentation is desirable.

This is an early stage study with several limitations in the test data and evaluated methods. Besides the clustering approaches being used as such without any further processing discussed here, the only tracer utilised was [^{18}F]-FDG. While it is a widely used one, other tracers, particularly those not based on fluorine, could provide drastically different data as they accumulate differently. We do not expect different clustering methods to be accurate with different tracers, but another tracer could introduce different challenges not present in this study. In addition, our data was limited to 40 images from rats and 3 images from humans, while ideally a larger number of human total-body images would have been used. For the sake of straightforward validation, only three VOIs were considered here. In real application, the number of functionally distinct areas of total-body PET images is much greater, though most studies likely focus on only some of them. Other issue with the validation is that the manual segmentation is done on sum 3D images neglecting the time aspect. Thus, VOIs with different time-patterns, but similar total radioactivities won't be distinguished and segmentation methods for static 3D PET images would likely appear equally strong to the methods for dynamic images if compared here. However,

doing the manual segmentation for validation purposes from a video of a 3D object is not feasible at the time of writing.

Notably, bladder filling during the imaging process has been stated to be a problem for PCA in the literature [10], but our data does not support that claim. While the bladder is not within the scanned area in our rat data, the human images include it, yet PCA provided meaningful clustering on them. However, our human data is limited to three images so further investigation of the topic is needed.

In this study, we have shown PCA combined with k-means to be the best performing method among the tested alternatives. In addition, its running time is in the same scale as the fastest method tested here. However, considering its incapability to capture brain from the images and unstable performance with kidney, there is room and need for a more sophisticated total-body TAC clustering method. While the requirement for scalability to very large datasets presents some challenges, we have discussed multiple computationally light approaches to further improve the results.

References

[1] Brent Foster, Ulas Bagci, Awais Mansoor, Ziyue Xu, and Daniel J Mollura. A review on segmentation of positron emission tomography images. *Computers in biology and medicine*, 50:76–96, 2014.

[2] Koon-Pong Wong, Dagan Feng, Steven R Meikle, and Michael J Fulham. Segmentation of dynamic pet images using cluster analysis. *IEEE Transactions on nuclear science*, 49(1):200–207, 2002.

[3] Jacob W Vogel, Niklas Mattsson, Yasser Iturria-Medina, Olof T Strandberg, Michael Schöll, Christian Dansereau, Sylvia Villeneuve, Wiesje M van der Flier, Philip Scheltens, Pierre Bellec, et al. Data-driven approaches for tau-pet imaging biomarkers in alzheimer's disease. *Human brain mapping*, 40(2):638–651, 2019.

[4] Sandrine Mouysset, Hiba Zbib, Simon Stute, Jean-Marc Girault, Jamal Charara, Joseph Noailles, Sylvie Chalon, Irène Buvat, and Clovis Tauber. Segmentation of dynamic pet images with kinetic spectral clustering. *Physics in Medicine & Biology*, 58(19):6931, 2013.

[5] Hiba Zbib, Sandrine Mouysset, Simon Stute, Jean-Marc Girault, Jamal Charara, Sylvie Chalon, Laurent Galineau, Irene Buvat, and Clovis Tauber. Unsupervised spectral clustering for segmentation of dynamic pet images. *IEEE Transactions on Nuclear Science*, 62(3):840–850, 2015.

[6] Jun L Chen, Steve R Gunn, Mark S Nixon, and Roger N Gunn. Markov random field models for segmentation of pet images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 468–474. Springer, 2001.

[7] Hongbin Guo, Rosemary Renaut, Kewei Chen, and Good Samaritan PET Center. Clustering for three dimensional kinetic pet data. In *Proceedings of IEEE International Conference on Data Mining, Clustering Large Data Sets, Workshop notes*, pages 43–48, 2003.

[8] John Ashburner, Jane Haslam, Christopher Taylor, Vincent Cunningham, and Terry Jones. A cluster analysis approach for the characterization of dynamic pet data. In *Quantification of brain function using PET*, pages 301–306. Academic Press, Ltd, 1996.

[9] Brian J Parker and Dagan Feng. Graph-based mumford-shah segmentation of dynamic pet with application to input function estimation. *IEEE Transactions on Nuclear Science*, 52(1):79–89, 2005.

[10] Renaud Maroy, Raphaël Boisgard, Claude Comtat, Vincent Frouin, Pascal Cathier, Edouard Duchesnay, Frédéric Dollé, Peter E Nielsen, Régine Trébossen, and Bertrand Tavitian. Segmentation of rodent whole-body dynamic pet images: an unsupervised method based on voxel dynamics. *IEEE transactions on medical imaging*, 27(3):342–354, 2008.

[11] Tony Shepherd and Rikard Owenius. Gaussian process models of dynamic pet for functional volume definition in radiation oncology. *IEEE transactions on medical imaging*, 31(8):1542–1556, 2012.

[12] Jinxiu Cheng-Liao and Jinyi Qi. Segmentation of mouse dynamic pet images using a multiphase level set method. *Physics in Medicine & Biology*, 55(21):6549, 2010.

[13] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[14] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[15] Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011.

[16] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[17] Yuji Kuge, Kazuo Minematsu, Yasuhiro Hasegawa, Takenori Yamaguchi, Hideaki Mori, Hajime Matsuura, Naoto Hashimoto, and Yoshihiro Miyake. Positron emission tomography for quantitative determination of glucose metabolism in normal and ischemic brains in rats: an insoluble problem by the harderian glands. *Journal of Cerebral Blood Flow & Metabolism*, 17(1):116–120, 1997.

[18] Anniina Snellman, Jatta S Takkinen, Francisco R López-Picón, Olli Eskola, Olof Solin, Juha O Rinne, and Merja Haaparanta-Solin. Effect of genotype and age on cerebral [18f] fdg uptake varies between transgenic appswe-ps1de9 and tg2576 mouse models of alzheimer's disease. *Scientific reports*, 9(1):1–12, 2019.

[19] Yuichi Kimura, Michio Senda, and Nathaniel M Alpert. Fast formation of statistically reliable fdg parametric images based on clustering and principal components. *Physics in Medicine & Biology*, 47(3):455, 2002.

[20] Mustafa E Kamasak and Bulent Bayraktar. Unsupervised clustering of dynamic pet images on the projection domain. In *Medical Imaging 2006: Image Processing*, volume 6144, pages 1539–1548. SPIE, 2006.