Automatic Detection and Classification of Regions of FDG Uptake in Whole-Body PET-CT Lymphoma Studies

Lei Bi[a], Jinman Kim[a,*], Ashnil Kumar[a], Lingfeng Wen[a,b], Dagan Feng[a,c], and Michael Fulham[a,b,d]

[a] School of Information Technologies, University of Sydney, NSW, Australia

[b] Department of Molecular Imaging, Royal Prince Alfred Hospital, NSW, Australia

[c] Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China

[d] Sydney Medical School, University of Sydney, NSW, Australia

[*] Corresponding author: jinman.kim@sydney.edu.au

***Abstract*— [^{18}F]-Fluorodeoxyglucose (FDG) positron emission tomography – computed tomography (PET-CT) scans of lymphoma patients usually show disease involvement as foci of increased radiotracer uptake. Existing methods for detecting abnormalities, model the characteristics of these foci; this is challenging due to the inconsistent shape and localization information about the lesions. Thresholding the degree of FDG uptake is the standard method to separate different sites of involvement. But may fragment sites into smaller regions, and may also incorrectly identify sites of normal physiological FDG uptake and normal FDG excretion (sFEPU) such as the kidneys, bladder, brain and heart. These sFEPU can obscure sites of abnormal uptake, which can make image interpretation problematic. Identifying sFEPU is therefore important for improving the sensitivity of lesion detection and image interpretation. Existing methods to identify sFEPU are inaccurate because they fail to account for the low inter-class differences between sFEPU fragments and their inconsistent localization information. In this study, we address this issue by using a multi-scale superpixel-based encoding (MSE) to group the individual sFEPU fragments into larger regions, thereby, enabling the extraction of highly discriminative image features via domain transferred convolutional neural networks. We then classify there regions into one of the sFEPU classes using a class-driven feature selection and classification model (CFSC) method that avoids overfitting to the most frequently occurring classes. Our experiments on 40 whole-body lymphoma PET-CT studies show that our method achieved better accuracy (an average F-score of 91.73%) compared to existing methods in the classification of sFEPU.**

***Index Terms*—Classification, Thresholding, PET-CT, CNN**

1. INTRODUCTION

[^{18}F]Fluorodeoxyglucose positron emission tomography – computed tomography (FDG PET-CT) is regarded as the imaging

modality of choice for the evaluation, staging and assessment of response in many malignancies including the lymphomas [1-3]. The combination of PET and CT in one device combines the sensitivity of PET to detect regions of abnormal function and the anatomical localization of CT [3]. Sites of disease usually display greater FDG uptake than normal structures. The standardized uptake value (SUV) is a semi-quantitative measure of FDG uptake or glucose metabolism and is extensively used in clinics to measure the degree of FDG uptake in sites of disease [3]. Different malignancies have varying degrees of FDG uptake and lymphomas are one of the most glucose-avid cancers that are routinely staged and re-staged with PET-CT. SUV thresholding of PET images is the main approach used to detect sites of abnormal FDG uptake before and after treatment [4, 5]. Regions with an SUV value higher than a specified limit (called the 'threshold value') are identified as regions of interest (ROIs) e.g., tumors [1, 3, 5-8]. Common SUV threshold values include an SUV of ≥2.5 [9], 4.4 [10], 5.3 [11], and a value above the average SUV of a background reference region [12, 13], such as the liver [3, 14] and the mediastinal blood pool in the thoracic aorta [14, 15]. However, (global) SUV thresholding does not take local SUV variations into account and so can include normal tissue.

We define sites of FDG excretion and physiologic uptake (sFEPU) as the globally thresholded sites of expected normal FDG uptake that are thresholded alongside tumors and other abnormal regions in whole-body PET studies. These sFEPU predominately belong to the excretion uptake in the kidneys and both ureters, normal physiological uptake in the brain and the heart, and pooling of FDG in the bladder. A single sFEPU is often split into many smaller fragments, which is a byproduct of global thresholding on heterogeneous structures such as the kidneys, which have varying degrees of FDG present in different locations. Global thresholding can therefore make the image-driven assessment of disease problematic as it can obscure disease in adjacent structures, in particular, in the paravertebral regions, in the mid and lower abdomen where involve lymph nodes lie adjacent to the ureters. The automatic identification and labeling of sFEPU, and their separation from sites of disease would thus therefore improve lesion detection and computer aided diagnosis. The automated detection and labelling of sFEPU is challenging because: (1) there are low inter-class differences, as some sFEPU fragments may only partially represent a class/structure. e.g., a kidney fragment only represents a portion of the whole kidney which makes some image features ineffective for differentiation (see Figure 1b); (2) there is inconsistent localization information about sFEPU fragments due to the random localization of abnormal sites with the body; and (3) there is a large variation in the degree of FDG uptake among different patients where a structure (e.g., heart) may not have been thresholded (due to being under the threshold value) and thus appears 'absent'.

There have been many different approaches that attempt to separate and label different structures on PET-CT studies: (a) abnormality classification/detection, which attempts to classify/detect one type of abnormality, e.g., liver tumors; (b) multi-structure classification where the aim is to detect or semantically label multiple anatomical structures that excludes abnormalities;

and (c) abnormalities and multi-structure classification, which attempts to label different types of structures and abnormalities within the same framework. Existing research in abnormality detection is mainly limited to detecting only a single type of abnormality e.g., liver tumors [16], lung nodules [17], lung tumors [18]. The underlying assumption is that there is only single lesion type in the image. These methods typically require prior knowledge to model the abnormality and to constrain the detection e.g., lung segmentation is usually required for lung tumor detection and the classification accuracy will rely on the segmentation performance [18]. In addition, these methods are unproven for the simultaneous detection of abnormalities on whole-body images as they depend on the segmentation of anatomical structures and a priori knowledge about specific abnormalities. The majority of multi-structure classification approaches are optimized to localize normal structures: Zhan et al. [19] used an active scheduling approach to detect multiple organs, Criminisi et al. [20] used relative spatial features with random forest to localize different organs on CT volumes, and Linguraru et al. [21] used template matching to detect abdominal organs on CT. Methods using probabilistic atlases with deformable registration, geometric transformation and probabilistic averaging have also been used to identify multiple organs [22-26]. The focus on normal structures, however, means that these methods struggle in the presence of the deformations introduced by disease, which affect the size and shape of the involved structure in variable and inconsistent ways [27].

There has been limited work on the simultaneous classification (detection and separation) of abnormalities and multiple normal structures. In general, this work has involved in localizing individual regions, extracting discriminative features, and then using supervised classification algorithms to label each region. Lartizien et al. [28] used a combination of texture features, filter based feature selection, and support vector machines (SVMs) to separate several types of lymphoma and non-lymphoma regions in PET-CT images. However, input ROIs required manual delineation, which is highly operator dependent, time-consuming, and is poorly reproducible across different user groups. Wu et al. [29] used region growing to detect potential abnormalities and then used SVMs to classify these regions into different classes. Similarly, Song et al. [30] used a multi-stage classification framework that combined SVM with conditional random field (CRF) for detecting the lungs, mediastinum, lung tumors and lymph nodes. In a later work, Song et al. [31] used a weighted sparse representation with image patches for classification. However, all these works were designed to work with specific anatomical regions in PET-CT images such as the thorax [30, 31] and head and neck area [29]. Furthermore, these methods relied on contextual features to separate different structures and were dependent on the accurate localization of the normal structures. Such methods are not suitable for whole-body PET-CT lymphoma studies where there can be innumerable sites of disease seen across the region examined.

In previous work, we conducted preliminary studies to address the simultaneous classification of abnormalities and multiple normal structures on whole-body PET-CT studies [32-34]. Our approach was to detect all the potential abnormalities e.g.,

thresholding and then iteratively filtering out normal structures rather than model lesions that can have inconsistent shapes and localization information. Abnormalities can be detected in a reverse manner through the filtering (removal) of normal structures. In the initial work, we used PET-CT features [32] to classify and separate the sFEPU fragments, where we selectively used PET, CT or PET-CT features based on the image characteristics of different structures. We extended this work to cluster the thresholded fragments thereby increasing the discriminative power of the features derived from clustered fragments when compared to using individual fragments [33]. We also investigated the optimal feature representation to individual structures using a structure based feature selection strategy together with a SVM for classification [34]. These previous approaches relied on using individual thresholded fragments which lack discriminative power especially for small fragments (as shown in Figure 1b)[32, 34]. The clustering based method assembled the fragments of the same structure (see Figure 1c) but was not able to describe the structures since the cluster only partially represented the actual structure and left large semantic differences between the clusters and the actual structure (see Figure 1 a, c and e)

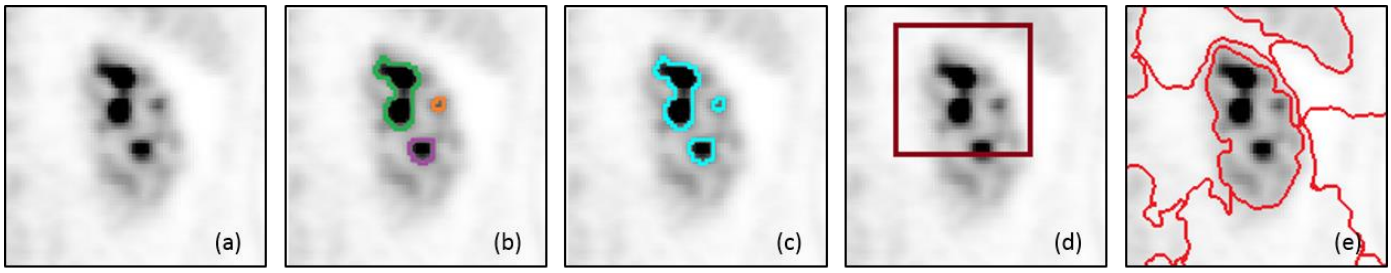

Figure 1: Feature extraction using different methods: (a) the PET image; (b) traditional classification method where different color contours represent different regions (fragments); (c) clustering-based method; (d) sliding window (patch) based classification method; and (e) our proposed superpixel-based classification method.

In this study, we propose a novel algorithm that uses a multi-scale superpixel-based encoding method (MSE) and a class-driven feature selection and classification model (CFSC) for sFEPU classification in whole-body PET-CT lymphoma studies. We derived class-driven features from multi-scale superpixel regions encoded with domain transferred deep convolutional neural networks (CNN) features for classification. Our algorithm differs from other methods as follows:

(1) Our MSE approach enables the grouping of the sFEPU fragments which then permits the extraction of optimal features on multi-scale superpixels, thereby increasing the discriminative power compared to using individual fragments. When compared with traditional methods reliant on sliding windows, our approach minimizes the risk of merging unrelated pixels by aggregating pixels conservatively into superpixels to capture local redundancy in the data. The use of multi-scale superpixels allows us to classify sFEPU of various sizes such as small lesions and large anatomical structures, which is more relevant for sFEPU classification when compared with single-scale superpixels or sliding window.

(2) We leverage a domain transferred deep CNN to encode individual superpixel regions. The CNN feature extractor allows us to obtain a feature representation of the original image that is more descriptive of the spatial characteristics of the superpixels when compared with handcrafted features such as scale-invariant feature transform (SIFT) features [35].

(3) CFSC enables the selection of the optimal features for classifying individual sFEPU. In contrast to the traditional feature selection methods, CFSC allows us to select optimal features locally (to individual sFEPU classes) and globally (among all classes) resulting in better inter-class differentiation and classification performance.

(4) Our algorithm operates on all of the structures in whole-body PET-CT images rather than body regions such as the thorax or the abdomen, which is more clinically relevant.

Our preliminary results were reported in our conference paper [36] and our approach has been updated for this paper. We now use linear spectral clustering (LSC) superpixel generation instead of the simple linear iterative clustering (SLIC) method, which allows the superpixels to account for image-wide properties (e.g., intensity variability) thereby enabling adherence to important image edges while ignoring less important ones (optimized across the whole image): this is not possible with region-wise SLIC [37]. We also replaced the texture features with a new features set derived from a domain transferred deep CNN features. Transfer learned CNN features have consistently shown benefits in the medical image domain [38] and this replacement importantly increases the feature discrimination of the spatial characteristics of individual superpixels that can better identify sFEPU fragments. Furthermore, we replaced the sparse and dense (SD) based classification approach, which was prone to overfit to the dominant classes, with a class-driven feature selection and classification model (CFSC). The new CFSC maximized the difference among individual classes to produce balanced results. We have also carried out a more thorough evaluation and comparison of our approach to related state-of-the-art methods. The rest of the paper is organized as follows: Section 2 gives the detailed description of our method and evaluation materials; Sections 3 and 4 outline the Results and Discussion; and the summary of our contributions are found in Section 5.

2. Methods and Materials

2.1 Materials and Ground Truth Construction

Our dataset consisted of 40 whole-body PET-CT studies from 11 lymphoma patients provided by the Department of Molecular Imaging, Royal Prince Alfred (RPA) Hospital, Sydney, NSW, Australia. There were 6 females and 5 males (age: 17 – 82 years old, body weight: 44 – 90 kg). The 40 scans were divided across the 11 patients as follows: 1 patient with 6 scans, 6 patients with 4 scans, 2 patients with 3 scans, 2 patients with 2 scans. All studies were acquired on a Biograph TruePoint PET-CT scanner

(Siemens Medical Solutions, Hoffman Estates, IL, USA). Approximately 400 MBq of [^{18}F]-FDG was injected intravenously; the uptake period was 60 min. The acquisition time of PET was 1.5 - 4 min per bed position, depending on the patient's weight. PET images were reconstructed using the 3-D ordered-subset expectation maximization (3-D-OSEM) method with 21 subsets and 3 iterations and point spread function (PSF) based resolution recovery. CT-derived attenuation correction, random counts correction, [^{18}F] decay correction, and Siemens proprietary scatter correction were incorporated in the reconstruction. The reconstructed PET has a matrix size of 168×168 pixels with a pixel size of 4.07mm^2 and the reconstructed CT has a matrix size of 512×512 pixels with a pixel size of 0.98mm^2. Both PET and CT had a slice thickness of 3mm. During the preprocessing, the PET images were linearly interpolated to the same voxel size as the CT images. Upsampling of PET images, was chosen over downsampling of CT images, to avoid losing pixel information. The bed and linen were automatically removed from the co-registered CT images using an adaptive thresholding and image subtraction method with a given bed template [39]. All data were de-identified.

The ground truth were regions identified using PERCIST thresholding (see Section 2.2 and Section 2.3) together with the diagnostic report of the PET-CT scan. PERCIST thresholding was applied to the PET data to generate a binary mask. Using the diagnostic reports, experienced operators manually labelled 655 objects in the binary mask as different sFEPU fragments, which included: 49 brain (BR), 39 bladder (BL), 38 heart (HE), 107 left kidney (LK), 116 right kidney (RK), and 306 other hypermetabolic (HY) fragments (Figure 2f). The hypermetabolic fragments were sites of active lymphoma in lymph nodes, spleen, bone marrow etc. or other sites where there can be markedly increased FDG uptake in brown fat, sites of inflammation, and physiologic uptake in bowel. There are more fragments than the number of studies due to the normal expected fragmentation of large anatomical structures into smaller fragments after thresholding.

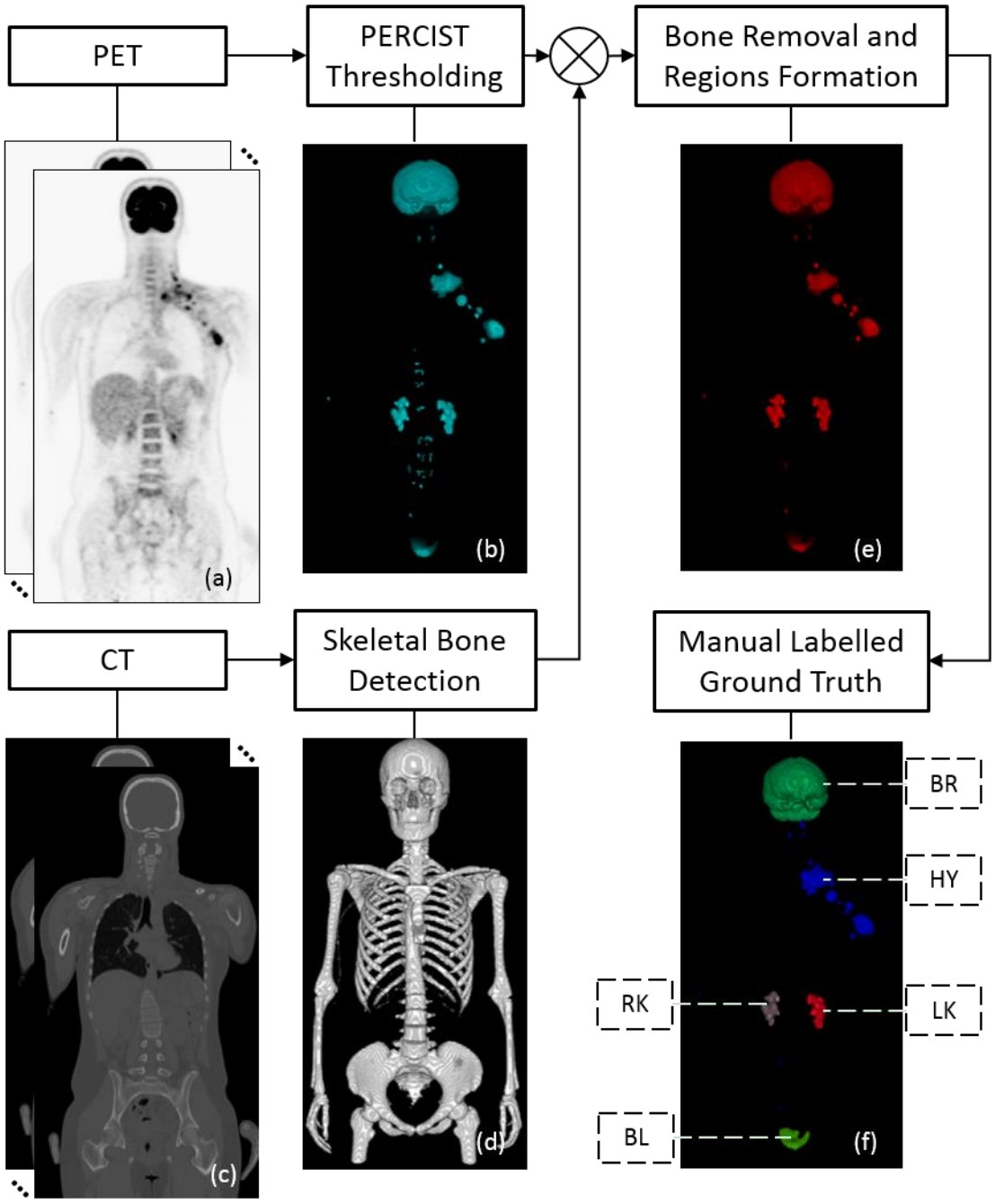

Figure 2: Generation of ground truth: (a) Coronal PET image – there are multiple regions of increased FDG uptake in sites of lymphoma in the left axilla and left supraclavicular fossa; the excretion uptake in the kidneys; normal physiological uptake in the brain; pooling of FDG in the bladder (showing in different slices) and uptake in the bones; (b) results of PERCIST thresholding from coronal PET; (c) coronal CT from corresponding PET image; (d) bony skeleton calculated from CT; (e) thresholding result after removal of bony skeleton; (f) manually labeled ground truth. Note: (b, d, e, f) are represented in 3D volume for visual clarity.

2.2 Automatic PERCIST-based Thresholding

PERCIST is a robust method for calculating SUV thresholds. It is based on SUV (normalized by the lean body mass, denoted as SUV_{LBM}) together with a reference volume of interest (VOI) [3, 6, 40, 41] – a 3cm diameter sphere placed on the right lobe of the liver to measure the average FDG. We considered the fragments that were above this threshold value to be the sFEPU. We automatically calculated the PERCIST threshold value by applying our prior work on automatic PERCIST thresholding [42] on the PET image to generate a binary mask $T_{PERCIST}$ (Figure 2b).

2.3 Bony skeleton detection

We removed bony structures from the thresholded results using the anatomical bone information from the corresponding CT slice. A binary skeletal mask $T_{Skeletal}$ was generated using a threshold of >150 Hounsfield Units (HU) on the CT [43] (Figure 2d). $T_{Skeletal}$ was then subtracted from $T_{PERCIST}$. A morphological filter was applied to the resulting binary mask to remove artifacts (Figure 2e).

2.4 Overview of the Classification Framework

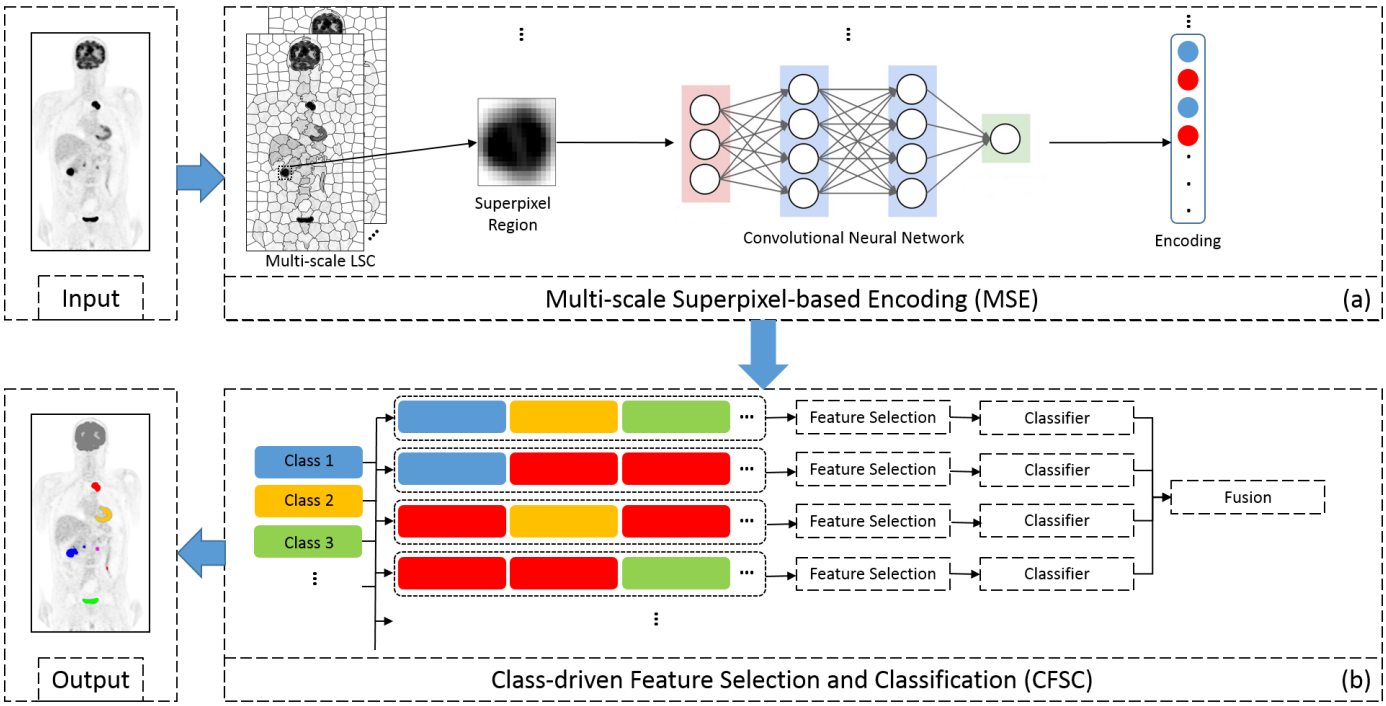

Figure 3: Flow diagram of the framework.

The outline of our proposed classification framework is shown in Figure 3. Multi-scale superpixel-based encoding was applied on PET image with domain transferred deep convolutional neural networks to encode individual superpixel regions (Section 2.5). Our class-driven feature selection and classification model was then applied on individual superpixel regions to produce the final labelling (Section 2.6).

2.5 Multi-scale Superpixel-based Encoding (MSE)

An input PET image slice was segmented into N small superpixels by the linear spectral clustering (LSC) [37] algorithm (Figure 3a). LSC was adopted due to its low computational costs and its ability to preserve image-wide properties to produce superpixels. We extracted superpixels at different scales to manage different sized sFEPU by changing the grid interval size, which allowed us to detect structures (such as tumors) on small scales and large structures (such as the brain) on a larger scale. For each scale, we encoded the individual superpixel via transfer learning. We used the deep learned convolutional neural networks (CNN) model [44] with a VGGNet architecture (developed by Visual Geometry Group, VGG-F) [45] trained on natural images (ImageNet [46]) as a feature extractor, to encode the PET superpixels as a 4096-dimension feature vector. The VGGNet architecture was used for its better performance on image classification problems [47]. We resized the superpixels to be the same size as the VGGNet input (224×224) and we padded the non-superpixel area with a fixed background intensity value of 0. In this way, we ensured our method preserved the original semantic and shape information of the superpixels. We also included spatial information by calculating the centroid of the superpixel in the transverse, coronal, and sagittal planes. We avoided duplication of highly-correlated features and increase variance by reducing the feature dimension to 200 (covering approximate 90% of variance) using principal component analysis [48]. Our superpixels were generated on 2D PET image slices to fit the VGGNet input and we adopted the VGG-F pre-trained model from the MatConvNet library. Our feature extraction used 5 scales of superpixels ranging from 50 to 250 with an increment of 50; these values were selected to balance the computation efficiency with the classification accuracy.

2.6 Class-driven Feature Selection and Classification (CFSC)

Our CFSC model was based on the popular filter based feature selection methods maximum relevance – minimum redundancy (MRMR) algorithm [49], which has been shown to be robust when applied to PET images [28]. In particular, the MRMR algorithm selects the optimal subset of features that are highly relevant to the labels and has low redundancy with other selected features (as determined by mutual information). The process was as follows:

1. We initially gathered the training samples (labelled superpixels) into a single set $\boldsymbol{D} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_l, \dots \boldsymbol{X}_{ln}\}$ where $\boldsymbol{X}_l$ is the set of the training samples extracted from class $l \in \mathbf{L}$ and ln is the number of classes.
2. We divided $\boldsymbol{D}$ into two sets: $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ such that $\boldsymbol{D_1} = \boldsymbol{X}_l$, i.e. $\boldsymbol{D_1} \subset \boldsymbol{D}$ with only one class label and $\boldsymbol{D_2} = \boldsymbol{D}/\boldsymbol{D_1}$. All the samples in $\boldsymbol{D_1}$ are labeled as +1 while the samples in $\boldsymbol{D_2}$ are labeled as -1.
3. From $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, we extracted the optimal feature set $\boldsymbol{\psi}_l$ with binary labels $\mathbf{L_b} = \{+1, -1\}$ according to:

$$\max_{f_k \in \mathbf{R}} \{M(f_k, \mathbf{L_b}) - \frac{1}{|\mathbf{S}|} \sum_{f_i \in \mathbf{S}} M(f_k, f_i)\} \quad (1)$$

where f_k is a feature we are testing, **S** is the set of all currently selected features, **R** is the set of all other features, and M is a probabilistic based mutual information measurement [49]. The optimal features set were selected via cross-validation via support vector machine (SVM) on the training data.

4. We used the same feature set $\boldsymbol{\psi_l}$ for the input sample (unlabeled superpixels) sp. We then trained a binary classifier C_l to separate $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$. The trained classifier was then tested on sp.
5. We then calculated a probability score $\boldsymbol{\rho}_l(sp)$ based on the output of the classifier C_l. The probability of sp being classified as positive (+1) can be considered as the probability of sp to be classified as label l and we denoted this by $\boldsymbol{\rho}_l^+(sp)$. Similarly, the probability of sp being classified as negative (-1) is the probability that it is not classified as label l and we denoted this as, $\boldsymbol{\rho}_l^-(sp)$.
6. We repeated steps 2-5 for all l in **L**.
7. We combined the final probability score obtained at each iteration with a multi-class probability score. We then extracted optimal features $\boldsymbol{\psi}_m$ according to:

$$\max_{f_k \in \mathbf{R}} \{M(f_k, \mathbf{L}) - \frac{1}{|\mathbf{S}|} \sum_{f_i \in \mathbf{S}} M(f_k, f_i)\} \quad (2)$$

and trained a one-versus-one multi-class classifier C_m by using all the training samples $\boldsymbol{D}$ together with their corresponding labels. The input sp was using the same optimal features as $\boldsymbol{\psi}_m$ and tested with classifier C_m.

8. We obtained the multi-class probability score $\mathbf{P}_m(sp, l)$ representing the probabilities of sp to be classified as label l. Then the final labeling of sp was calculated as:

$$\mathrm{L}(sp) = \underset{l \in \mathbf{L}}{arg\max} (\mathbf{P}_m(sp, l) + \boldsymbol{\rho}_l^+(sp) - \boldsymbol{\rho}_l^-(sp)) \quad (3)$$

We used a support vector machine (SVM) with a radial basis function (RBF) kernel for the classification (both C_l and C_m). The RBF kernel maps the data, non-linearly, into a higher dimension space [50, 51] where the data are more easily separable. In contrast, linear kernels usually have poor performance in non-linear classification tasks while polynomial kernels are usually

computationally expensive [52]. The RBF kernel parameters were optimized with a default grid search analysis method, which is available in LIBSVM [51].

2.7 Multi-scale Superpixel Integration

In order to manage the labelling of each region in 3D volume, we used a majority voting scheme to derive the labels for 3D regions. 2D multi-scale superpixel probabilities were first integrated into pixel-level probabilities, by averaging them across different scales:

$$\emptyset(\varphi, l) = \frac{\sum_{\sigma \in \mathbf{G}} \partial(\varphi, \sigma, l)}{|\mathbf{G}|} \qquad (4)$$

where $\mathbf{G}$ represents different scales and $\partial(\varphi, \sigma, l)$ represents the probabilities (derived from CFSC) for a pixel φ at scale σ to be label l. After that, the region in 3D volume was labelled based on the majority vote of all the pixels within the region.

3. Experiments and Results

3.1 Experiment Setup

We performed the following experiments: (a) an evaluation of the performance of using superpixels for sFEPU classification; (b) an analysis of the performance of individual components; and (c) a comparison with existing methods on sFEPU classification. The first and second experiments were performed on individual 2D image slices. For the third experiment, we made the different methods comparable by using the integration method in Section 2.7 to produce classification results on 3D volumes. We compared the labels from our classification with those in the ground truth (Section 2.1). All experiments followed a leave-one-patient-out cross-validation approach, where we ensured that no patient PET-CT scans were in both the training and test set. We used the F-score, which is the balanced value of precision and recall, for measuring the performance on each class. A receiver operating characteristic (ROC) curve was also used to visualize the performance of using different scales of superpixels. For the first experiment, we compared using superpixels for sFEPU classification on individual classes. Five different scales of superpixels were used in the experiments, ranging from 50-250 per slice (increments of 50). For the second experiment, we compared our method with: (i) SP-Texture-SVM – superpixel encoded with texture features and classified with SVM; (ii) SP-SVM – superpixel with SVM; (iii) SP-MRMR – superpixel with MRMR for feature selection. The experiment was conducted on a superpixel scale of 50 (i.e., 50 superpixels per slice) for all methods, because provided the overall best performance across all methods and avoided the positive influence from using a multi-scale superpixel integration approach. Our texture features included gray-level co-occurrence matrix (GLCM) features and gray-level run length matrix (GLRLM) features, which have proven performance on PET images [28, 53]. For the third experiment, we compared our method with: (i) SP-SD [36] – sFEPU classification via multi-scale superpixels with sparse and dense representations; (ii) Grouping [33] – a clustering based classification method; and (iii) Patch-

SVM – multi-scale sliding window with SVM. The sliding window based method used 5 windows of a similar size to that of the superpixels. We also included the best performing scale from our method for additional comparison.

3.2 Superpixels Performance

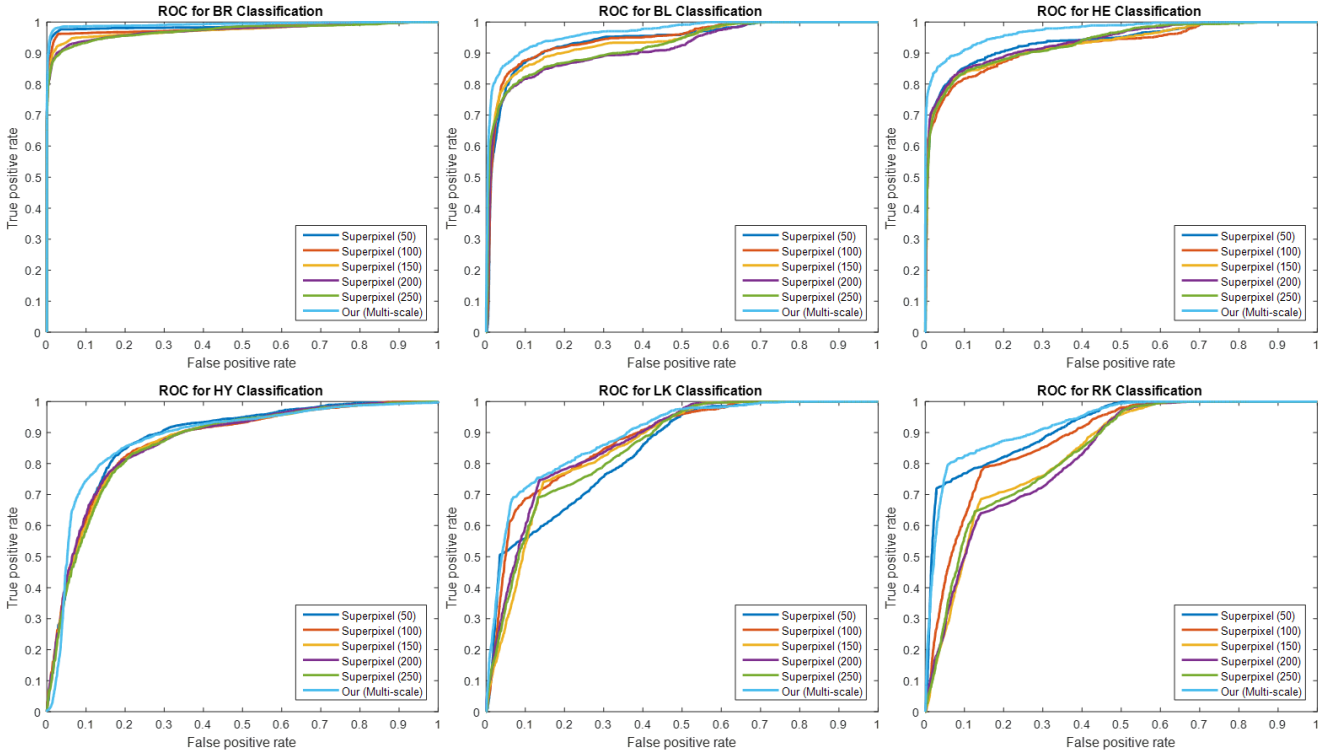

Figure 4: Receiver operating characteristic (ROC) curves of sFEPU classification with different scales of superpixels

In Figure 4 the sFEPU classification performance across different scales of superpixels measured via ROC curve is shown. Overall, the proposed multi-scale method had a higher classification accuracy compared to using any single-scale superpixels. Larger superpixel regions (fewer superpixels per slice e.g., 50 superpixels per slice) performed better than smaller superpixel regions (e.g., 250 superpixels per slice).

3.3 Component Analysis

Table 1: The classification performance of different methods measured via F-score

Method (%)	BR	BL	HE	HY	LK	RK	Average	Std
SP-Handcraft-SVM	94.81	74.43	74.24	58.06	49.58	71.91	70.51	15.59
SP-SVM	96.18	79.33	80.91	61.20	51.91	74.16	73.95	15.62
SP-MRMR	97.44	80.26	86.09	70.21	68.29	78.36	80.10	10.74
Our	**98.22**	**85.38**	**89.74**	**73.49**	**77.53**	**84.61**	**84.83**	**8.78**

Table 1 shows the classification results of the various methods. Our method performed best on classification across the different sFEPU classes and it had the highest average F-score of 84.83%, which is 4.73% higher than the 2nd best method.

3.4 Overall Performance

Table 2: The classification performance of different methods measured via F-score.

Methods (%)	BR	BL	HE	HY	LK	RK	Average	Std
Patch-SVM	79.34	76.29	83.72	84.00	**93.40**	95.32	85.34	7.57
SP-SD [36]	**93.33**	80.46	87.06	89.38	93.39	**97.00**	90.10	5.86
Grouping [33]	62.16	80.56	82.35	86.34	93.14	89.62	82.36	10.92
Our – Best Scale	81.36	85.71	77.55	85.66	93.21	91.77	85.87	5.97
Our – Multi-scale	85.96	**93.98**	**92.68**	**93.22**	89.34	95.20	**91.73**	**3.44**

In Table 2, we list the classification performance of all methods and our method had the better performance on most of the sFEPU classes including BL, HE, HY and the highest average and lowest standard deviation of 91.73% and 3.44%.

4. DISCUSSION

We show that a larger superpixel region (fewer superpixels in a coronal slice e.g., 50 superpixels per slice) performs better than a smaller (e.g., 250) superpixel region for most of the sFEPU classifications, which can be explained by the larger superpixel region contributing more discriminative features that are crucial for classification (see Figure 4). In addition, smaller superpixel regions are useful for classifying smaller sFEPU fragments such as for HE and LK, which can be attributed to the smaller superpixels regions being more adaptive to the smaller fragments. Our multi-scale approach performed best in classification and this suggest that multi-scale integration provides complementary information to target sFEPU fragments of various sizes.

The difference between the SP-Texture-SVM and SP-SVM approaches (see Table 1) illustrates the benefits of using domain transferred deep CNN features for classification. When compared to the SP-SVM, the SP-MRMR approach greatly improved classification accuracy with an average increase of 6.15% F-score and this underlines the importance of feature selection processes in classification. When compared to the SP-MRMR, our approach further improved the classification accuracy, in particular for the BL, HE, LK and RK classes. We explain this as follows - although the SP-MRMR approach was able to identify the most relevant features among different sFEPU classes, the selected features were sub-optimal for individual structures, especially the less dominant classes. Our CFSC approach, in comparison, selected optimal features more robustly on the local and global levels (among different classes) that resulted in a more stable performance across different classes. The large variation between Table 1 and Table 2, is to be expected, because image slices (Table 1) usually carry less semantic information than 3D volumes (Table 2). Table 2 also shows that the inclusion of the multi-scale superpixel levels enables superior classification accuracy when compared to the best single-scale superpixel level (an average of 5.87% increase) and is consistent with the results from the first experiment.

The difference between SP-SD and Patch-SVM shows the advantages of using superpixels over sliding windows and emphasizes the importance of not merging unrelated pixels for features extraction. The SP-SD and our multi-scale approach consistently achieved better performance than the grouping method across different classes. The grouping method required a fixed parameter to define the range of the grouping, which led to over- or under-grouped results with poorer classification performance. In contrast, the SP-SD and our multi-scale approach can label individual regions on different scales based on the superpixel region probability, which minimizes the risk of one-off classification. We suggest that the improvement of our method over the SP-SD relates to the use of class-driven feature selection and classification (CFSC) model. The combination of a sparse and dense based classification model was feasible for sFEPU classification only on a global level with less accurate results for less dominant classes such as the BL class. In comparison, our CFSC model optimized the features and the classification performance for individual sFEPU classes and thus had a more consistent performance, with the highest average accuracy (91.73%) and the lowest standard deviation (3.44%).

5. Conclusion and Future work

We proposed a new classification method that automatically classifies and labels sites of FDG excretion and physiologic uptake in whole-body PET-CT images. Our experiments with 40 clinical lymphoma PET-CT cases show that our method classifies sFEPU classes more accurately than conventional methods. Our improved accuracy relates to using MSE and CFSC to derive class-driven features from multi-scale superpixel regions encoded with domain transferred deep convolutional neural networks. We will now expand our current work to the classification of more clinical datasets in patients with lymphoma before and after therapy and work on embedding this approach into a clinical workflow.

References

[1] R. Hong, J. Halama, D. Bova, A. Sethi, and B. Emami, "Correlation of PET standard uptake value and CT window-level thresholds for target delineation in CT-based radiation treatment planning," *International Journal of Radiation Oncology* Biology* Physics*, vol. 67, pp. 720-726, 2007.

[2] L. Freudenberg, G. Antoch, P. Schütt, T. Beyer, W. Jentzen, S. Müller, *et al.*, "FDG-PET/CT in re-staging of patients with lymphoma," *European journal of nuclear medicine and molecular imaging*, vol. 31, pp. 325-329, 2004.

[3] R.L. Wahl., "From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors," *Journal of Nuclear Medicine*, vol. 50, pp. 122S-150S, 2009.

[4] H. Yu, C. Caldwell, K. Mah, and D. Mozeg, "Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning," *Medical Imaging, IEEE Transactions on*, vol. 28, pp. 374-383, 2009.

[5] S. Vauclin, K. Doyeux, S. Hapdey, A. Edet-Sanson, P. Vera, and I. Gardin, "Development of a generic thresholding algorithm for the delineation of 18FDG-PET-positive tissue: application to the comparison of three thresholding models," *Physics in medicine and biology*, vol. 54, p. 6901, 2009.

[6] K. Hirata, K. Kobayashi, K.-P. Wong, O. Manabe, A. Surmak, N. Tamaki, *et al.*, "A semi-automated technique determining the liver standardized uptake value reference for tumor delineation in FDG PET-CT," 2014.

[7] U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hellwig, C. Rübe, *et al.*, "Comparison of different methods for delineation of 18F-FDG PET–positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer," *Journal of Nuclear Medicine*, vol. 46, pp. 1342-1348, 2005.

[8] A. C. Paulino and P. A. Johnstone, "FDG-PET in radiotherapy treatment planning: Pandora's box?," *International Journal of Radiation Oncology* Biology* Physics*, vol. 59, pp. 4-5, 2004.

[9] D. Hellwig, T. P. Graeter, D. Ukena, A. Groeschel, G. W. Sybrecht, H.-J. Schaefers, *et al.*, "18F-FDG PET for mediastinal staging of lung cancer: which SUV threshold makes sense?," *Journal of Nuclear Medicine*, vol. 48, pp. 1761-1766, 2007.

[10] J. F. Vansteenkiste, S. G. Stroobants, P. De Leyn, P. J. Dupont, J. Bogaert, A. Maes, *et al.*, "Lymph node staging in non-small-cell lung cancer with FDG-PET scan: a prospective study on 690 lymph node stations from 68 patients," *Journal of Clinical Oncology*, vol. 16, pp. 2142-2149, 1998.

[11] A. S. Bryant, R. J. Cerfolio, K. M. Klemm, and B. Ojha, "Maximum standard uptake value of mediastinal lymph nodes on integrated FDG-PET-CT predicts pathology in patients with non-small cell lung cancer," *The Annals of thoracic surgery*, vol. 82, pp. 417-423, 2006.

[12] R. J. Francis, M. J. Byrne, A. A. van der Schaaf, J. A. Boucek, A. K. Nowak, M. Phillips, *et al.*, "Early prediction of response to chemotherapy and survival in malignant pleural mesothelioma using a novel semiautomated 3-dimensional volume-based analysis of serial 18F-FDG PET scans," *Journal of Nuclear Medicine*, vol. 48, pp. 1449-1458, 2007.

[13] K. R. Zasadny, P. V. Kison, I. R. Francis, and R. L. Wahl, "FDG-PET determination of metabolically active tumor volume and comparison with CT," *Clinical Positron Imaging*, vol. 1, pp. 123-129, 1998.

[14] N. Paquet, A. Albert, J. Foidart, and R. Hustinx, "Within-patient variability of 18F-FDG: standardized uptake values in normal tissues," *Journal of Nuclear Medicine*, vol. 45, pp. 784-788, 2004.

[15] P. Ghosh and M. Kelly, "Expanding the power of PET with PERCIST," *White Paper, Siemens Healthcare*, 2010.

[16] D. Pescia, N. Paragios, and S. Chemouny, "Automatic detection of liver tumors," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, 2008, pp. 672-675.

[17] F. Zhang, Y. Song, W. Cai, M.-Z. Lee, Y. Zhou, H. Huang, *et al.*, "Lung nodule classification with multilevel patch-based context analysis," *Biomedical Engineering, IEEE Transactions on*, vol. 61, pp. 1155-1166, 2014.

[18] C. Ballangan, X. Wang, S. Eberl, M. Fulham, and D. Feng, "Automated lung tumor segmentation for whole body PET volume based on novel downhill region growing," in *SPIE Medical Imaging*, 2010, pp. 76233O-76233O-8.

[19] Y. Zhan, X. S. Zhou, Z. Peng, and A. Krishnan, "Active scheduling of organ detection and segmentation in whole-body medical images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, ed: Springer, 2008, pp. 313-321.

[20] A. Criminisi, J. Shotton, and S. Bucciarelli, "Decision forests with long-range spatial context for organ localization in CT volumes," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2009, pp. 69-80.

[21] M. G. Linguraru and R. M. Summers, "Multi-organ automatic segmentation in 4D contrast-enhanced abdominal CT," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, 2008, pp. 45-48.

[22] X. Han, M. S. Hoogeman, P. C. Levendag, L. S. Hibbard, D. N. Teguh, P. Voet, *et al.*, "Atlas-based auto-segmentation of head and neck CT images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, ed: Springer, 2008, pp. 434-441.

[23] X. Zhuang, K. Rhode, S. Arridge, R. Razavi, D. Hill, D. Hawkes, *et al.*, "An atlas-based segmentation propagation framework using locally affine registration–application to automatic whole heart segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, ed: Springer, 2008, pp. 425-433.

[24] M. Fenchel, S. Thesen, and A. Schilling, "Automatic labeling of anatomical structures in MR FastView images using a statistical atlas," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, ed: Springer, 2008, pp. 576-584.

[25] A. Shimizu, "Multi-organ segmentation in three dimensional abdominal CT images," *Proc of Computer Assisted Radiology and Surgery, 2006*, 2006.

[26] C. Yao, T. Wada, A. Shimizu, H. Kobatake, and S. Nawano, "Simultaneous location detection of multi-organ by atlas-guided eigen-organ method in volumetric medical images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 1, pp. 42-45, 2006.

[27] C. Li, X. Wang, Y. Xia, S. Eberl, Y. Yin, and D. D. Feng, "Automated PET-guided liver segmentation from low-contrast CT volumes using probabilistic atlas," *Computer methods and programs in biomedicine*, vol. 107, pp. 164-174, 2012.

[28] C. Lartizien, M. Rogez, E. Niaf, and F. Ricard, "Computer-Aided Staging of Lymphoma Patients With FDG PET/CT Imaging Based on Textural Information," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, pp. 946-955, 2014.

[29] B. Wu, P.-L. Khong, and T. Chan, "Automatic detection and classification of nasopharyngeal carcinoma on PET/CT with support vector machine," *International journal of computer assisted radiology and surgery*, vol. 7, pp. 635-646, 2012.

[30] Y. Song, W. Cai, J. Kim, and D. D. Feng, "A Multistage Discriminative Model for Tumor and Lymph Node Detection in Thoracic Images," *Medical Imaging, IEEE Transactions on*, vol. 31, pp. 1061-1075, 2012.

[31] Y. Song, W. Cai, H. Huang, X. Wang, S. Eberl, M. Fulham, *et al.*, "Similarity guided feature labeling for lesion detection," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, ed: Springer, 2013, pp. 284-291.

[32] L. Bi, J. Kim, D. D. Feng, and M. Fulham, "Classification of thresholded regions based on selective use of PET, CT and PET-CT image features," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 2014, pp. 1913-1916.

[33] L. Bi, J. Kim, D. Feng, and M. Fulham, "Multi-stage Thresholded Region Classification for Whole-Body PET-CT Lymphoma Studies," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, ed: Springer International Publishing, 2014, pp. 569-576.

[34] L. Bi, J. Kim, L. Wen, D. Feng, and M. Mulham, "Automated Thresholded Region Classification Using A Robust Feature Selection Method For PET-CT," in *International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, 2015.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91-110, 2004.

[36] L. Bi, J. Kim, A. Kumar, D. Feng, and M. Mulham, "Adaptive Supervoxel Patch-based Region Classification in Whole-Body PET-CT," presented at the Medical Image Computing and Computer Assisted Intervention (MICCAI) - Computational Methods for Molecular Imaging (CMMI), 2015 (In Press).

[37] Z. Li and J. Chen, "Superpixel Segmentation using Linear Spectral Clustering," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 1356-1363.

[38] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," 2016.

[39] J. Kim, Y. Hu, S. Eberl, D. Feng, and M. Fulham, "A fully automatic bed/linen segmentation for fused PET/CT MIP rendering," in *Society of Nuclear Medicine Annual Meeting Abstracts*, 2008, p. 387P.

[40] M. Niyazi, S. Landrock, A. Elsner, F. Manapov, M. Hacker, C. Belka, *et al.*, "Automated biological target volume delineation for radiotherapy treatment planning using FDG-PET/CT," *Radiat Oncol*, vol. 8, p. 180, 2013.

[41] L. Bi, J. Kim, L. Wen, A. Kumar, M. Fulham, and D. D. Feng, "Cellular automata and anisotropic diffusion filter based interactive tumor segmentation for positron emission tomography," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 5453-5456.

[42] L. Bi, J. Kim, L. Wen, and D. D. Feng, "Automated and Robust PERCIST-based Thresholding framework for whole body PET-CT studies," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 5335-5338.

[43] S. Hu, E. A. Hoffman, and J. M. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *Medical Imaging, IEEE Transactions on*, vol. 20, pp. 490-498, 2001.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[45] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[48] I. Jolliffe, *Principal component analysis*: Wiley Online Library, 2002.
[49] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1226-1238, 2005.
[50] B. Schölkopf, K.-K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, *et al.*, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2758-2765, 1997.
[51] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, p. 27, 2011.
[52] M. Kakar and D. R. Olsen, "Automatic segmentation and recognition of lungs and lesion from CT scans of thorax," *Computerized Medical Imaging and Graphics*, vol. 33, pp. 72-82, 2009.
[53] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat, "Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis," *Journal of Nuclear Medicine*, vol. 55, pp. 414-422, 2014.