Automated Thresholded Region Classification Using a Robust Feature Selection Method for PET-CT

Lei Bi[1], *Jinman Kim*[1], *Lingfeng Wen*[2], *Dagan Feng*[1,4] and *Michael Fulham*[1,2,3]

[1]School of Information Technologies, University of Sydney, Australia
[2]Department of PET and Nuclear Medicine, Royal Prince Alfred Hospital, Australia
[3]Sydney Medical School, University of Sydney, Australia
[4]Med-X Research Institute, Shanghai Jiao Tong University, China

ABSTRACT

Fluorodeoxyglucose Positron Emission Tomography – Computed Tomography (FDG PET-CT) is the preferred imaging modality for staging the lymphomas. Sites of disease usually appear as foci of increased FDG uptake. Thresholding is the most common method used to identify these regions. The thresholding method, however, is not able to separate sites of FDG excretion and physiological FDG uptake (sFEPU) from sites of disease. sFEPU can make image interpretation problematic and so the ability to identify / label sFEPU will improve image interpretation and the assessment of the total disease burden and will be beneficial for any computer aided diagnosis software. Existing classification methods, however, are sub-optimal as there is a tendency for over-fitting and increased computational burden because they are unable to identify optimal features that can be used for classification. In this study, we propose a new method to delineate sFEPU from thresholded PET images. We propose a feature selection method, which differs from existing approaches, in that it focuses on selecting optimal features from individual structures, rather than from the entire image. Our classification results on 9222 coronal slices derived from 40 clinical lymphoma patient studies produced higher classification accuracy when compared to existing feature selection based methods.

Index Terms— Feature Selection, Classification, PET-CT.

1. INTRODUCTION

Positron emission tomography-computed tomography, using ^{18}F-Fluorodeoxyglucose (FDG PET-CT), is arguably the best and most accurate imaging modality to stage, assess treatment response and also to detect disease relapse in the lymphomas. Sites of disease are indicated by varying levels of FDG uptake in PET images in lymph nodes, which may or may not be enlarged on the anatomical data provided by the co-registered CT images, the spleen, soft tissues and the bone marrow [1].

The standardized uptake value (SUV), which is a semi-quantitative measure of FDG uptake, is widely used in the evaluation of malignant conditions [2]. Thresholding is the most common method that is used to identify sites of abnormal FDG uptake in lymphoma patients [2,3] and a $SUV \geq 2.5$ and a value that is of 50% of the SUV_{max} are commonly applied. FDG uptake that is not due to disease and is related to normal FDG excretion from the kidneys and pooling in the bladder and physiological uptake in the brown fat, the brain, the heart and active muscles is routinely seen in clinical FDG PET-CT scans. We refer to such regions as sites of FDG excretion and physiologic uptake (sFEPU). sFEPU can make image interpretation problematic (see Fig. 1), obscure disease in adjacent structures and limit the ability to measure the entire disease burden. Further, it is important to accurately identify sFEPU for computer aided diagnosis.

The correct identification and labeling of sFEPU will improve the evaluation of malignant disease generally, image interpretation and visualization. In our previous work, sFEPU were separated and labelled by using PET or CT features [4] or contextual features via region grouping [5]. Texture features, as well as descriptive features or spatially-focused features have been widely used for their ability to characterize different tumors and structures in PET data [4-6]. However, many of these features may be irrelevant or redundant which may degrade classification performance with over-fitting and increased computational burden [6,7].

In recent work, feature selection has been reported as a means to address this issue of selecting optimal features. By introducing manual annotations, Lartizien et al [6] proposed a computer aided staging framework to separate malignant regions in lymphoma with PET-CT images. The optimal textual based image features were extracted by a filter based feature selection method, and followed by support vector machine (SVM) in evaluation. However, the required manual delineation of regions as input, which is operator-dependent and time-consuming, may limit its application in

the clinical environment and is not suitable for automated computer aided diagnosis.

Thus, in this study, we propose a new automatic approach to label sFEPU based on selecting an optimal subset of image features from individual structures. Initially, PET images were used to extract 51 descriptive and texture features for their proven ability to discriminate tumors [8]. Then the structure-based feature selection method embedded with a maximum relevance – minimum redundancy (MRMR) algorithm was used to select an optimal subset of features. These features were then used in classification and labelling. When compared to previous work, our approach differs as follows: (1) our method is fully automated and it does not require any manual delineation; and (2) we introduce the novel concept of selecting optimal features based on structures that enable greater discrimination features and better classification performance.

2. METHODS

2.1. Materials and Ground Truth Construction

Our dataset consisted of 47 whole-body PET-CT studies from 11 lymphoma patients acquired from the Department of PET and Nuclear Medicine at Royal Prince Alfred Hospital (Camperdown, NSW, Australia). Seven scans with extensive lymphoma were excluded since all the thresholding regions from different sFEPU groups were connected to each other. Hence, a total of 40 PET-CT studies were used in this work (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 64-slice PET-CT scanner (Siemens Medical Solutions, Hoffman Estates, IL, USA).

The bed and linen were automatically removed from the co-registered CT images by an adaptive thresholding and image subtraction method with a given bed template [9]. Training and ground truth data were constructed according to the PET Response Criteria in Solid Tumors (PERCIST) threholding method [3]. PERCIST is a robust method for deriving SUV threshold [2,10,11] and we have adopted the automated PERCIST calculation in our previous work [12].

To exclude normal uptake in the bone, we segmented the bony skeleton from CT (threshold of >150 Hounsfield Units [5]) and then removed these structures from the PERCIST thresholded PET images via image subtraction. A morphological filter was then applied on the resulting binary mask to remove noise (exemplified in Fig. 1b). The remaining sites were considered to be sFEPU and we manually labelled them as belonging to the brain (BR), bladder (BL), heart (HE), left kidney (LK), right kidney (RK) or other (OT) structures (please see Fig. 1c). The "other" class contained potential abnormalities (based on patients' clinical report), regions of physiological uptake, brown fat and lymph node inflammation. A total of 467 thresholded regions were manually labeled as the ground truth. To provide more data for evaluation we separated each region into slices (in the coronal plane).

Regions smaller than 4×4 pixels were not considered in this work since features measurement might not be meaningful in small regions [8]. A total of 9222 slices were used in this study, including 3191 BR, 1205 BL, 1224 HE, 1012 LK, 1285 RK and 1305 OT region slices.

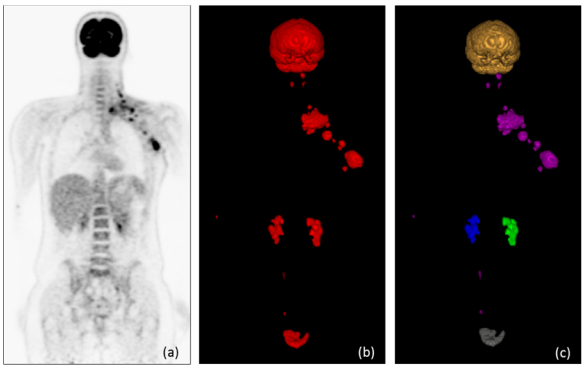

Figure 1. An example of sFEPU. (a) coronal PET image; (b) thresholding result after removing bony skeleton uptake seen in brain, kidneys, bladder and also in the disease in lymph nodes in the left axilla and left supra-clavicular fossa; and (c) manually labeled ground truth, where brown, grey, green, blue and purple represent BR, BL, LK, RK and OT sFEPU.

2.2. PET Features Extraction

All the PET images were processed with SUV according to the injected dose of FDG and patient body weight. The descriptive and texture features were then extracted from PET SUV images according to the work reported by Orlhac et al [8]. In this work, we extracted these features from PET slices mentioned in section 2.1. Fifty one features were calculated including spatial and statistical descriptive (SSD) features, gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), neighborhood gray-level different matrix (NGLDM) and gray-level zone length matrix (GLZLM) features. There were 6 GLCM, 11 GLRLM, 5 NGLDM and 11 GLZLM features. The SSD features include average locations in transverse, coronal, saggital planes, SUV_{max}, SUV_{mean}, SUV standard deviation (SUV_{std}), $SUV_{skewness}$, $SUV_{kurtosis}$, SUV_{median}, SUV_{min}, SUV_{sum}, maximum lengths in two directions, perimeter, circularity, area, area ratio to bounding box and roundness.

2.3. Maximum Relevance – Minimum Redundancy (MRMR) Features Ranking

The filter based feature selection method is robust on PET images [6] and the maximum relevance – minimum redundancy (MRMR) algorithm is an established filter based feature selection method developed by Peng et al [13]. The aim of the MRMR algorithm is to select the optimal subset that has high relevance to the labels and low redundancy to the selected features based on the use of mutual information. In this work, we used this concept to measure the importance of the features.

Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x,y)$, which was defined as follows:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (1)$$

Suppose we have set **S** containing all the selected features and set **R** containing remaining features. **F** is the whole feature sets and $\{\mathbf{S} \cup \mathbf{R}\} = \mathbf{F}$. Then the relevance function D can be defined as:

$$D = I(f,l) \quad (2)$$

where f is a feature we are testing ($f \in \mathbf{R}$) and l is the class label. Further, the redundancy function C can be calculated between the testing feature f and all the selected features:

$$C = \frac{1}{|\mathbf{S}|}\sum_{f_i \in \mathbf{S}} I(f,f_i) \quad (3)$$

where $|\mathbf{S}|$ is the number of features in a selected features set. After combining equation (2) and (3), feature f_k ($f_k \in \mathbf{R}$) that maximizes the relevancy while minimizing redundancy was calculated as:

$$\max_{f_k \in \mathbf{R}}\left\{I(f_k,l) - \frac{1}{|\mathbf{S}|}\sum_{f_i \in \mathbf{S}} I(f_k,f_i)\right\}. \quad (4)$$

MRMR starts from an empty set and iteratively ranks all *N* features.

2.4. Structure Based Feature Selection and Classification

Fig. 2 presents our structure-based feature selection workflow to classify sFEPU. Initially, the training dataset consisting of all sFEPU labels were copied and changed into multiple binary training datasets which only keeping one sFEPU label and the rest of labels. These training datasets were then used to extract optimal features based on the MRMR algorithm. Before these features were placed into the binary SVM, we trained these features with a radial basis function (RBF) kernel to non-linearly map the data into a higher dimension space [14]. This helps to make the training data more identifiable in a computationally efficient way, where a linear kernel usually has poor performance in a non-linear classification task while a polynomial kernel is computationally expensive [4]. The RBF kernel parameters were optimized with a default grid search analysis method, which is available in the LIBSVM [14]. The testing datasets use the same extracted feature indices and parameters as the training datasets for classification. However, to label the new region only based on the maximum probability from each of binary SVM output, this may lose prediction accuracy, since the confident scale and classification difficulty could be different for each binary SVM group. To maximize the difference between each binary SVM group, we summed each binary SVM outputs with a multi-class SVM output, described at the bottom of Fig. 2. The final labelling of region r is based on the maximum probability score calculated as:

$$arg\max_{l \in \mathbf{L}}\{\mathbf{P}_l(r) + \boldsymbol{\rho}_l(r)\} \quad (5)$$

where **P** and **ρ** are the probability matrixes, calculated from multi-class SVM and binary SVM respectively ($\mathbf{L} = \{BR, BL, HE, LK, RK, OT\}$).

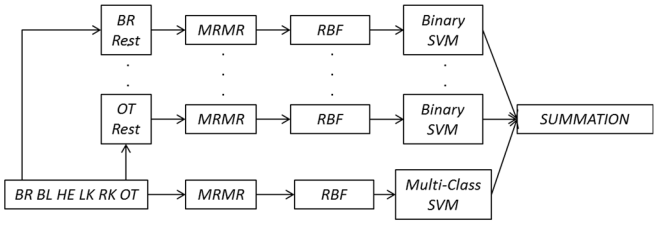

Figure 2: The structure-based feature selection and classification process: initially, multi-class training dataset was transformed into multiple binary datasets to only keep one class versus the rest of classes. Then, features were extracted from each binary datasets and trained from a RBF kernel. Finally, the output of the binary SVM was summed with the multi-class SVM output to maximize the probability score and to label each of regions.

3. RESULTS AND DISCUSSION

3.1 Experiments Setup

We considered a region to be classified correctly if it was labelled as the same as the ground truth. We randomly partitioned the data into two sets and we ensured that there were no patient studies that appeared in both datasets. The first set was used to construct the training set and the second set was used for evaluation. Then we reversed the roles of the two sets. Here, we split the data into two sets to speed up the optimization process. This also ensured that the selected features were not over-fitted compared to splitting the data into many datasets. We compared our proposed Structure-based feature selection (StrFS) with the classification results from using multi-class SVM without feature selection (denoted as NoFS), with conventional feature selection method MRMR and using multiple binary SVMs with MRMR (BSVMs+MRMR). Further, we compared our method with common feature selection methods including the conditional informative feature extraction (CIFE) [15] and F-score based feature selection methods (F-select) [16]. MRMR and CIFE are commonly used as filter based feature selection methods while F-select is a wrapper-based feature selection approach. The same SVM classifier were used for all methods and the optimized RBF parameters were retrieved for each method as mentioned in Section 2.4. Top N features were defined beforehand for MRMR, CIFE and StrFS; we ran N ranging from 1 to full features with the best performing results selected for each of the methods. For BSVMs+MRMR and StrFS, we used the same N as MRMR.

3.2 Results and Discussion

The summary of the classification results for all the feature selection methods are shown in Table 1. Due to the space limitations, the three representative methods - NoFS, best conventional feature selection method (MRMR) and the proposed method - are detailed in the confusion matrix in Table 2. When compared to the other methods, the StrFS produced the best results for all measurements. We attributed this improvement to the use of the MRMR algorithm and the structured based feature selection approach, where the MRMR algorithm effectively identified

the most relevant features and the structured based feature selection approach was able to emphasize the most relevant features with a higher probability score. Further, both StrFS and MRMR were able to produce their respective best performance when *N*=10, whereas CIFE needed *N*=25, resulting in higher efficiency in computation.

Table 1: Overall accuracy (Accu), average precision (Prec), sensitivity (Sens) and specificity (Spec) of proposed method and other methods

Method (%)	Accu	Prec	Sens	Spec
NoFS	86.99	86.83	84.00	97.33
BSVMs+MRMR	85.72	84.50	82.17	97.17
MRMR	93.60	92.50	92.50	98.50
CIFE	89.76	88.50	87.50	97.83
F-select	81.34	83.17	81.33	96.33
StrFS	**95.21**	**94.33**	**94.17**	**98.83**

Table 2: Classification results with our StrFS method compared to NoFS and MRMR. (Note Ground Truth is denoted as G.T.)

Method (Overall)	G.T.	Prediction (%)					
		OT	***BR***	***BL***	***HE***	***LK***	***RK***
NoFS (86.99%)	OT	**64.90**	13.64	9.04	4.90	1.84	5.67
	BR	0.38	**99.62**	-	-	-	-
	BL	7.47	-	**92.53**	-	-	-
	HE	27.94	-	-	**71.57**	0.49	-
	LK	14.13	-	0.89	1.98	**83.00**	-
	RK	9.03	-	0.31	-	-	**90.66**
MRMR (93.60%)	OT	**82.61**	3.07	4.52	3.98	1.00	4.83
	BR	1.32	**98.68**	-	-	-	-
	BL	7.14	-	**92.86**	-	-	-
	HE	5.72	-	-	**91.50**	2.78	-
	LK	4.25	-	-	1.38	**94.37**	-
	RK	5.76	-	-	-	-	**94.24**
StrFS (95.21%)	OT	**82.15**	3.37	4.75	4.06	1.07	4.60
	BR	0.16	**99.84**	-	-	-	-
	BL	5.56	-	**94.44**	-	-	-
	HE	1.55	-	-	**95.67**	2.78	-
	LK	3.16	-	-	1.09	**95.75**	-
	RK	3.19	-	-	-	-	**96.81**

In Table 2, the high classification accuracy for all methods when classifying the brain was expected as the brain has unique feature characteristics compared with other structures. The relatively low classification accuracy for all methods when classifying the "other" class was likely due to the unstable features, the inclusion of many different structures and potential abnormalities in various locations e.g., ureters and brown fat. When compared to the NoFS, the MRMR approach greatly improved the classification accuracy, which underlines the importance of feature selection processes in classification. When compared to the MRMR approach, the StrFS further improved the classification performance, especially in classifying the bladder, heart, left and right kidney. This is likely to be although the MRMR approach could identify most relevant features, these features were selected on a global level which could be sub-optimal for individual structures. The structured based feature selection approach, meanwhile, could select optimal features more robustly, on local and global levels thus demonstrating the discriminative ability to partition the structures into individual sFEPU.

4. CONCLUSION AND FUTURE WORK

In this work, we propose a new classification method to classify and label the sFEPU automatically, based on deriving optimal structure-based features. Our experiments with 9222 coronal slices derived from 40 clinical PET-CT lymphoma studies resulted in higher classification accuracy when compared to other conventional feature selection approaches. For future studies, we will evaluate our method on greater number of clinical studies and disease types.

REFERENCES

[1] Freudenberg, L., et al. "FDG-PET/CT in restaging of patients with lymphoma. *Eur. J. Nucl. Med. Mol. I.,* 31(3), 2004.

[2] Hirata, K., et al., "A Semi-Automated Technique Determining the Liver Standardized Uptake Value Reference for Tumor Delineation in FDG PET-CT. " *PloS one*, 2014.

[3] Wahl, RL., et al. "From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors." *J. Nucl. Med.*, 122S-50S, 2009.

[4] Bi, L., et al. "Multi-stage Thresholded Region Classification for Whole-Body PET-CT Lymphoma Studies," *MICCAI*, pp569-76, 2014.

[5] Bi, L., et al. "Classification of Thresholded Regions based on Selective Use of PET, CT and PET-CT Image Features," *EMBC*, pp1913-16, 2014.

[6] Lartizien, C., et al. "Computer aided staging of lymphoma patients with FDG PET/CT imaging based on textural information," *ISBI*, pp.118,121, 2012.

[7] Brown, G., et al. "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection." *J. Mach. Learn. Res*., 13(1): 27-66, 2012.

[8] Orlhac, F., et al. "Tumor Texture Analysis in ^{18}F-FDG PET: Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis." *J. Nucl. Med*., 2014.

[9] Kim, J., et al., "A fully automatic bed/linen segmentation for fused PET/CT MIP rendering," *J. Nucl. Med.*, 387P, 2008.

[10] Bi, L., et al. "Cellular Automata and Anisotropic Diffusion Filter based Interactive Tumor Segmentation for Positron Emission Tomography," *EMBC,* 2013.

[11] Niyazi, M., et al., "Automated biological target volume delineation for radiotherapy treatment planning using FDG-PET/CT." *Radiat. Oncol*., 180, 2013.

[12] Bi, L., et al. "Automated and robust PERCIST-based thresholding framework for whole body PET-CT studies," *EMBC*, 2012.

[13] Peng, H., et al. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE. T. Pattern. Anal*., 2005.

[14] Chang, C C., et al. "LIBSVM : a library for support vector machines," *ACM TIST*, 2(3), 27, 2011.

[15] Lin, D., et al. "Conditional infomax learning: an integrated framework for feature extraction and fusion." *ECCV*, 2006.

[16] Chen, Y W., et al. "Combining SVMs with various feature selection strategies." *Feature extraction*, pp315-24, 2006.