TOWARDS AUTOMATED LYMPHOMA PROGNOSIS BASED ON PET IMAGES

Gisele L. Pappa
Computer Science Department, UFMG
Belo Horizonte, Brazil

Hugues Talbot
A2SI, ESIEE, Université Paris-Est
Noisy-le-Grand, France

David Menotti
Computing Department, UFOP
Ouro Preto, Brazil

Michel Meignan
Nuclear Medicine Department, Université Paris 12
Créteil, Paris, France

ABSTRACT

This paper proposes a simple method to identify candidate tumors in a set of Positron Emission Tomography (PET) images obtained from patients suffering from lymphoma, and then extract statistics from the image most active tumor. These statistics are used as input for three machine learning algorithms, which generate models for overall survival and event-free survival. The results obtained by these methods are better than the ones obtained by visual analysis, and competitive or better than the ones obtained by a quantitative measure of prognosis. Besides, the results indicate that there is a lot of redundant information coming from the images, and only 2 out of 10 attributes might be enough to predict prognosis.

1. INTRODUCTION

In the past decade, images obtained from Positron Emission Tomography (PET) became one of the most promising tools for cancer diagnosis and prognosis [1]. PET is a nuclear medicine imagining technique that produces a series of 3D images showing functional processes of the body. During the exam, the patient is injected with a short half-life radioactive tracer, most commonly the fluorodeoxyglucose (FDG) – which is essentially a glucose. As cancer cells keep dividing indefinitely, they present higher glucose intake than surrounding healthy cells. These high concentrations of FDG appear as very bright areas in the image.

Although some quantitative methods have been recently proposed for cancer prognosis based on information extracted from the images [2], most of the time the diagnosis and/or prognosis is made visually, in a process which takes long enough to be worth automating.

There are three difficult problems to be tacked in order to allow an automated analysis of PET images: first, the resolution of the images is relatively low (144 x 144 x *z*) when compared to other types of medical images, e.g. CT or MRI images, due to the physical constraints of the imaging technique. Second, functional images do not bring structural information about the human body, making the image segmentation process more difficult. Third, not all the bright spots in the image necessarily represent tumors. In addition, this paper deals with a type of cancer in which segmentation and tumor identification is even more difficult than in any other type of cancer: lymphoma.

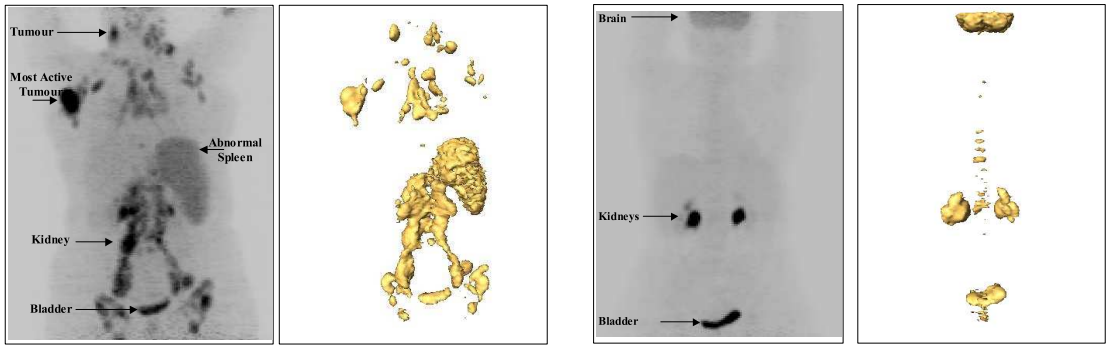
Lymphoma is a general term used to define many types of blood cancer which affect the lymphocytes. Typically, lymphoma cells form tumors in the lymphatic system organs: lymph nodes (the human body has approximately 500-600 lymph nodes, ranging in diameter from 2 to 15 mm), spleen, bone marrow and blood. However, as lymphocytes can move to many parts of the body, the cancer can also affect other organs (extranodal disease), including the liver and the lungs. In other words, there are small tumors that can be located in many areas of the human body.

This papers combines medical knowledge and a classical segmentation technique to extract and identify the most active tumor in a PET image. After the tumor is identified, statistics about the FDG uptake in that tumor are gathered, and three different machine learning techniques are used to create models for cancer prognosis. The proposed method is tested in a controlled data base composed of 92 patients with Diffuse Large B-cell Lymphoma (DLBCL).

The remainder of this paper is organized as follows. Section 2 discusses how lymphoma prognosis is currently made. Section 3 presents the method proposed to identify the most active tumor, extract characteristics from it and create decision trees and rule models for lymphoma prognosis. Section 4 brings some information about the image data base, while Section 5 presents the results of prognosis obtained and compare it with a visual analysis and a quantitative method. Finally, Section 6 draws some conclusions and presents ideas of future work.

2. CURRENT METHODS FOR LYMPHOMA PROGNOSIS

As the aim of this paper is to propose a method that automates the prognosis of lymphoma based on PET images,



(a) Images from the PET1 (first exam) (b) Images from the PET2 (second exam)

Fig. 1. Original and segmented images from two different PET exams

this section roughly describes how the images are visually analyzed by the physicians in order to predict prognosis. It also describes a quantitative method proposed by [2] for the same purpose.

During the diagnosis and treatment of lymphoma, patients go through a set of at least two or three PET exams. The first exam (from now on referred as PET1) is the one that usually identifies the disease, while the second and third (from now on referred as PET2 and PET4) are executed after 2 and 4 cycles of chemotherapy, and are used as follow-ups of the disease.

The prognosis usually is based on two different variables: event-free survival (EFS), which represents event *versus* no event (where an event includes patient progression, relapse, or death from any cause), and overall survival (OS), which predicts whether the patient lives or not.

In order to predict these two variables, the physicians take into account information about the candidate tumors (hot spots) from all available images, which tell them how the disease evolved. During this process, they are assisted by a quantitative measure of FDG uptake named SUV (standardized uptake value) [3]. The SUV is calculated considering the patient's weight and height, and the amount of FDG injected during the exam. Although the use of SUV is still controversial, in average lesions with SUV greater than 2.5 are considered as a tumor, while SUV values smaller than 2.5 are considered normal (although this value may vary according to the type of cancer being studied).

Fig. 1 shows the images for a lymphoma patient obtained during PET1 and PET2, and their respective segmented images. As observed in Fig. 1(a), the bladder and the kidneys appear as hot spots in the exam, as the injected FDG is expelled by the urinary system. Moreover, this patient also has another disease affecting his spleen, which also appears as a candidate tumor in the segmented image. The automatic classification of lesions like the abnormal spleen in Fig. 1(a) as tumors or not is very difficult, as the physicians use their knowledge and experience about the disease to identify them.

Fig. 1(b) shows the PET image of the same patient after two cycles of chemotherapy. Note that the segmented image shows only the kidneys, the bladder and the brain (which also might be active during the exam). In addition, parts of the vertebral column are also active. This behavior is also normal, and caused by the medicines used during treatment. In other words, there is no tumor left in PET2. In terms of the prognosis variables to be predicted, the patient had an event during the treatment, as its condition improved. However, he passed away in the next 42 months after treatment.

Just by the analysis of the images in Fig. 1 we notice that organs like the bladder, kidneys, heart and brain might be active during all the exams, while the vertebral column might appear in the PET2 and PET4 exams. Furthermore, other diseases or inflammations might also mislead an automated system for tumor identification.

Recently, a quantitative approach for lymphoma prediction was proposed in [2], where a ROC analysis was used to choose the SUV optimal cutoff value to separated EFS from non-EFS and OS from non-OS patients. It concluded that both (1) the maximum value of SUV found in the most active tumor of PET2 (from now on referred as $SUVMax_2$) and (2) the maximum SUV percentage of reduction in the most active tumor from PET2 to PET1 (from now on referred as $SUVMax_{Red12}$) were good prognosis predictors. It established a cut-off value for each of them that separates patients with different outcomes for both OS and EFS. Note that, although an analysis using $SUVMax_2$ obtained better results in predicting OS than $SUVMax_{Red12}$, the authors did not recommended it due to the differences in measurements that can occur in different PET machines (and affect the calculated SUV values). For this reason, the researchers concluded that using the $SUVMax_{Red12}$ as a predictor attribute would be more appropriated than using the absolute value of $SUVMax_2$.

The results showed in [2] were compared to the ones obtained from visual analysis, and improve both the accuracy,

PPV (positive predicted value), NPV (negative predicted value) and AUC (area under the ROC curve). However, the results reported were based on the entire population, having no statistical significance. Besides, this work did not explore the interactions between the different attributes which could be extracted from the images.

Hence, in this work, we first automatically identify the most active tumor in the image, and then extract a few statistics based on its values of SUV. Next, we create two different data sets: one to predict EFS and the other to predict OS, and feed them to three different machine learning algorithms that can explore these attribute interactions, as explained in the next section.

3. AUTOMATED PROGNOSIS OF LYMPHOMA

Most of the studies involving machine learning and cancer prognosis are concentrated in genomic, proteomic and clinical data, besides the combinations of them [4]. In this paper, we are particularly interested in data extracted from PET images obtained before treatment (PET1) and after two (PET2) and four (PET4) cycles of chemotherapy. In special, we want to investigate if interactions between attributes extracted from these images can lead to better prognosis than the one proposed in [2]. The method proposed in this paper works in three main steps: (1) image segmentation, (2) feature extraction and (3) prognosis models. Each of these steps is described in the next subsections.

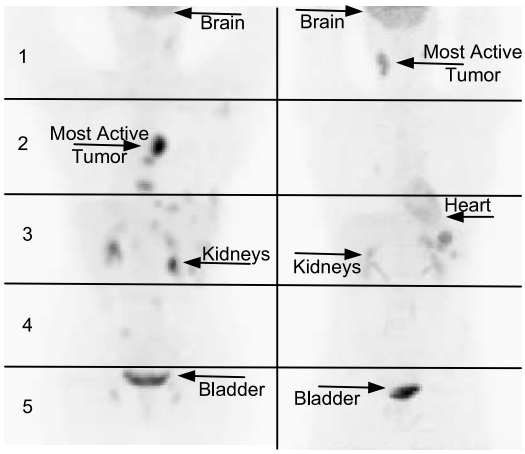


Fig. 2. Images of two patients with the bladder, brain, heart and kidneys identified

3.1. Image Segmentation

In this paper, the goal of the image segmentation process is to extract the most active tumor from the image. Physicians believe that the information contained in this tumor is enough to predict prognosis, and the properties of less aggressive tumors should not influence it. The use of more than one tumor to prognosis should be studied in the near future, but for now we are interested in comparing an automated approached based on the same principles as the visual ones used by the physicians.

The most challenging step during this phase is to extract from the image organs we know will certainly be active during the exam, such as the bladder, the kidneys, the heart and the brain, as we do not want to take them as tumors. PET image segmentation was successfully achieved before using both simple and more sophisticated methods [5, 6]. However, these methods focused on segmentation, without separating candidate tumors in real tumors or not.

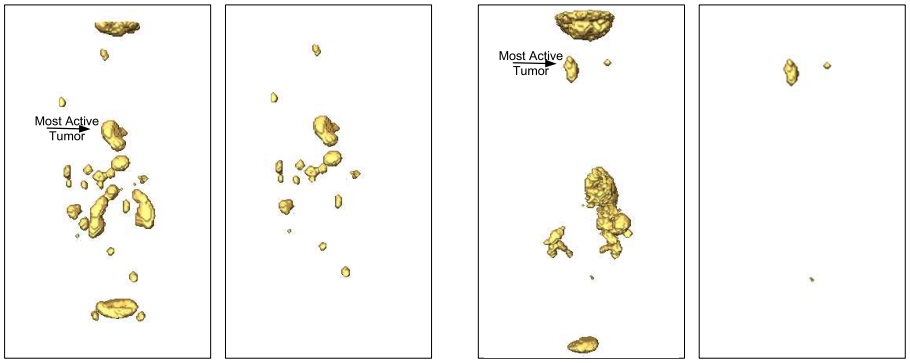
When we first started this work, we intended to apply a conventional gradient watershed segmentation [7], but the time necessary to process each image proved to be prohibitive. Hence, we took the simplest possible approach. We separated tumors from non-tumors the same way the physicians do: thresholding the image with a 2.5 SUV value. These values were obtained by multiplying the raw values in the image by a factor, based on the patient's weight and height, and in the amount of FDG injected on him during the exam.

After thresholding, we identified the connected components of the image and labeled them. An analysis of the resulting images showed that the bladder, kidneys, heart and brain were always between the six largest connected components. This information, together with the positional information of each organ, was used as a heuristic to exclude these organs from the images. Fig. 2 shows two images from two different patients. Although they are not aligned, note that if the image is divided in five equal regions, the brain is always in region 1, the kidneys in region 3 and the bladder in region 5. Regarding the kidneys, we know that their center of mass is separated by a maximum distance of 2 voxels, and can use this information allied to symmetry and positioning to easily exclude them from the image. The heart, when active, is usually in the bottom of region 2 or top of region 3. The vertebral column was not automatically excluded from the image but, in general, it showed not to be the most active spot in the PET2 images.

Fig. 3 shows the segmentation process for two patients (whose original images are shown in Fig. 2), contrasting the threshold segmented image and the image after removing the known active organs. Observe that the most active tumor is easily identified after the components are removed.

3.2. Feature Extraction

After the most active tumor in the image was identified, we extracted the following statistical attributes from the images: the maximum values of SUV uptake in PET1, PET2 and PET4; the average values of SUV uptake for PET1,



(a) Patient 1 (b) Patient 2

Fig. 3. PET1 Images from the same patients presented in Fig. 2 after segmentation and automatic exclusion of known active organs

PET2 and PET4; and the percentage of reduction in the maximum and averaged SUV values occurred from PET2 and PET4 to PET1.

In total, 10 attributes were used to create data sets to predict both EFS and OS. Note that the location of the most active tumor might change from the first to the second and third exams, i.e. the activity in the current most active tumor might be reduced (making another tumor the most active) or a new and more active tumor might appear. In this case, the statistics were extracted considering the difference between the old and new most active tumors.

3.3. Prognosis Models

The prognosis models that the physicians expect us to create from data are interpretable ones, which could be validated and understood by their medical knowledge. Taking that into account, we applied the C4.5 [8] and the Ripper [9] algorithms to the data, in order to generate a decision tree and a rule list. After analyzing the first results, we also applied the AdaBoost-M1 algorithm [10], to check if it would be able to improve the accuracy rates and positive predicted values obtained. The three algorithms were applied with their default parameters, using the WEKA data mining tool [11].

4. THE LYMPHOMA IMAGE DATA BASE

The data base of images considered in this study is composed of 92 patients suffering from Diffuse Large B-cell Lymphoma (DLBCL) lymphoma. The data was collected between January 2000 and December 2005, by a multicenter study involving four Departments of Hematology of the *Assistance Publique - Hôpitaux de Paris*.

During the 42 months follow up, 60 patients had no events, and 32 progressed or died (i.e. event-free survival (EFS) class distribution 60/32 examples). At the same time, considering overall survival, 71 patients survived and 21 died (death of any cause is counted here, with overall survival (OS) class distribution 71/21 examples).

5. COMPUTATIONAL RESULTS

In this section we report computational results obtained by running the C4.5, Ripper and AdaBoost-M1 algorithms with a 10-fold cross-validation procedure to predict both EFS and OS. While in [2] they looked for the best SUV-based measure to predict EFS and OS simultaneously using information from PET1 and PET2 images, here we consider EFS and OS prediction separately, and take into account information from PET1, PET2 and PET4 exams. Furthermore, the machine learning algorithms also give us a good insight about the interactions between different features extracted from the images, which was not done in [2]. The results of PPV, NPV, sensitivity, specificity, predictive accuracy and AUC for EFS and OS are reported in Tables 1 and 2, respectively.

The results found were compared with those obtained through a visual analysis (made by two physicians) and those reported in [2]. This latter study used a cutoff value of 65.74% for the SUV uptake reduction from PET1 to PET2 to predict both OS and EFS. In other words, the patients with reductions inferior to 65.74% would suffer an event and not survive treatment.

As observed in Tables 1 and 2, a visual analysis of the images predicted EFS with an accuracy of 65.2%, and overall survival with an accuracy of 68.5%. The study made in [2], in turn, obtained an EFS prediction accuracy of 76.1%, and overall survival accuracy of 84.8%.

It is important to point out that the majority of studies published about cancer diagnosis and prognosis are in medical journals and, as investigated by [4], most of them lack appropriated levels of validation and test. The results presented in [2] were not validated. Hence, in order to

Table 1. Results obtained for predicting EFS survival

Method	PPV	NPV	TPR	TNR	Acc	AUC
Visual Analysis	0.50	0.741	?	?	65.2	NA
$SUVRed12$-65.74%	0.813	0.75	0.95	0.406	76.09	0.689
$SUVRed12$	0.73 ± 0.02	0.642±0.03	0.916±0.05	0.367±0.07	72.67±3.09	0.642±0.03
C4.5	0.69 ± 0.02	0.523±0.14	0.90±0.05	0.242±0.05	67.55±2.93	0.605±0.06
Ripper	0.735 ± 0.03	0.567±0.14	0.90±0.04	0.367±0.09	71.78±4.00	0.633±0.05
AdaBoost-M1	0.864 ± 0.03	0.75 ±0.13	0.957±0.03	0.483±0.11	84.77±4.10	0.7±0.06

Table 2. Results obtained for predicting OS survival

Method	PPV	NPV	TPR	TNR	Acc	AUC
Visual Analysis	0.382	0.862	?	?	68.5	NA
$SUVMax_{Red12}$-65.74%	0.733	0.87	0.944	0.524	84.78	0.689
$SUVMax_{Red12}$	0.86±0.03	0.617±0.145	0.943±0.02	0.467±0.12	83.67±3.40	0.70±0.06
C4.5	0.85±0.02	0.7±0.13	0.957±0.02	0.433±0.09	83.66±2.97	0.707±0.06
Ripper	0.82±0.02	0.55±0.16	0.957±0.02	0.283±0.08	80.44±2.75	0.62±0.04
AdaBoost -M1	0.864±0.03	0.75±0.13	0.946±0.03	0.5±0.12	84.72±4.57	0.741±0.06

make them comparable with the ones obtained by the machine learning algorithms, we selected the $SUVMax_{Red12}$ attribute, and gave it as an input to the Ripper algorithm, which found the best cut-off for each data partition.

It is also important to emphasize that, contrary to most machine learning studies, where we are interested in the sensitivity and specificity of the methods, for physicians the most important measure lies on the PPV. The PPV is calculated as the number of true positives divided by all the examples classified as being positive. In the case of cancer prognosis, it will indicate, from the patients classified as survivors, the percentage of those who actually survived.

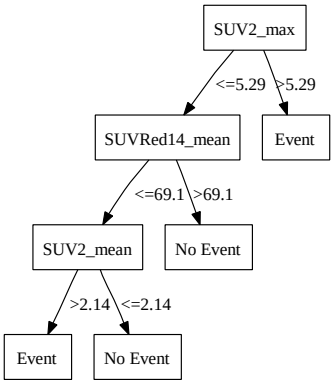


Fig. 4. Decision tree generated by C4.5 to predict EFS

```
if(SUVMean_Red12 <= 68.37%)then Event
                           else No Event
```

Fig. 5. Rule set generated by Ripper to predict EFS

Let us first analyze the results reported in Table 1. We observe that, according to a t-test with confidence level 0.05, when using the C4.5 and Ripper algorithms, the results obtained are considered statistically the same as the results obtained by the analysis using only the $SUVMax_{Red12}$. On the other hand, when using Adaboost-M1, it obtained better accuracy and PPV than the ones obtained by $SUVMax_{Red12}$ (cells indicated in the table by a dark gray background). However, the model generated by Adaboost-M1 is a combination of many other interpretable models, and cannot be easily interpreted. Hence, it does not appeal as much to the physicians as the other ones. Considering the results reported in Table 2, they are also all statistically the same as the ones obtained by the analysis using only the $SUVMax_{Red12}$ attribute.

Yet, the most interesting analysis we can perform from the previous experiments regards the models produced by both Ripper and C4.5, to check if they use only the $SUVMax_{Red12}$ as a predictor or if different attributes can lead to similar prognosis values. Figs. 4 and 5 show the models produced by both C4.5 and Ripper to predict EFS. As observed, the C4.5 does not use the $SUVMax_{Red12}$ in its decision tree, but instead the $SUVMax_2$, $SUVMax_{Red14}$ and $SUVAvg_2$.

In contrast, the model produced by Ripper contains only one rule, which uses the $SUVMean_{Red12}$. Although most studies agree that the use of the maximum value of the SUV may not be the ideal, due to the partial volume effect problem (PVE) [12], they argue that it is a better predictor than the average value of SUV in a tumor. Here, Ripper prefers the averaged value of SUV to the maximum one. However, recall that the quality measures obtained by the Ripper model are equivalent to those obtained using the maximum SUV value as a predictor.

Figs. 6 and 7 show the models produced by both C4.5 and Ripper to predict OS. Again, $SUVMax_2$ appears in the root of the decision tree, followed by $SUVMax_4$. Here, the only rule produced by Ripper also uses the value of $SUVMax_2$ with a threshold of 5.65 to distinguish survival from death.

The results obtained in Tables 1 and 2 showed that, apart from the results obtained by AdaBoost-M1 for EFS, all the

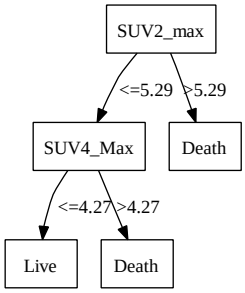


Fig. 6. Decision tree generated by C4.5 to predict OS

```
if (SUVMax2 >= 5.65)then Death
                    else Survival
```

Fig. 7. Rule set generated by Ripper to predict OS

others are statistically the same. Looking at the models generated, we would chose the simplest one as the best prognosis predictor, as all other information available is redundant, as it would lead to similar results. Following this idea, both models produced by Ripper would be used for prognosis. These models validate the results found in [2], indicating $SUVMax_2$ and $SUVMean_{Red12}$ as sufficient information for prognosis. However, note that in [2] they indicate $SUVMax_{Red12}$ and not $SUVMean_{Red12}$ as a good predictor. Further investigation on the use of the mean SUV value of a tumor or the maximum SUV value of a tumor, taking into account the PVE, is needed.

6. CONCLUSIONS

This study proposed a simple method for most active tumor automatic identification, feature extraction and machine learning models for lymphoma prognosis. The results were compared with both visual analysis and the quantitative approach proposed by [2], and showed that there is a lot of redundant information in the images. Besides, it validated the study of [2], confirming the use of $SUVMax_2$ as a good predictor, and suggesting the use of $SUVMean_{Red12}$ instead of $SUVMax_{Red12}$ for EFS prediction.

As a next step, we will investigate a more sophisticated and fast method for segmentation that identifies not only the most active tumor, but all the tumors in the image. This is not a simple task because, as explained earlier, there might be hot spots in the images caused by inflammations and other diseases, which are difficult to be distinguished from a real tumor. We then intend to study prognosis models based on data from all tumors, and not only the most active one. We also need to investigate how to remove the vertebral column from the PET2 images.

Acknowledgments

Many thanks to Emmanuel Itti and Gaetano Paone for the information regarding PET visual analysis. Also thanks to Sophie Lin for providing the data on PET quantitative prognosis analysis. This work was partially supported by ANR grant SURF-NT05-2_45825.

7. REFERENCES

[1] Y.S. Jhanwar and D.J. Straus, "The Role of PET in Lymphoma," *J Nuclear Medicine*, vol. 47, no. 8, pp. 1326–1334, 2006.

[2] C. Lin et al., "Early 18F-FDG PET for Prediction of Prognosis in Patients with Diffuse Large B-Cell Lymphoma: SUV-Based Assessment Versus Visual Analysis," *J Nuclear Medicine*, pp. 000–012, 2007.

[3] J.A. Thie, "Understanding the Standardized Uptake Value, Its Methods, and Implications for Usage," *J Nuclear Medicine*, vol. 45, no. 9, pp. 1431–1434, 2004.

[4] J.A. Cruz and D.S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, vol. 2, pp. 59–78, 2006.

[5] J. Liao and J. Qi, "Dynamic pet image segmentation using multiphase level set method," in *IEEE Nuclear Science Symposium Conference Record*. 2006, vol. 4, pp. 2047 – 2052, IEEE.

[6] H. Guan, T. Kubota, X. Huang, et al., "Automatic hot spot detection and segmentation in whole body fdg-pet images.," in *ICIP*. 2006, pp. 85–88, IEEE.

[7] F. Meyer and S. Beucher, "Morphological segmentation," *Journal of Visual Communication and Image Representation*, vol. 1, no. 1, pp. 21–46, Sept. 1990.

[8] J. R. Quinlan, *C4.5: programs for machine learning*, Morgan Kaufmann, 1993.

[9] W.W. Cohen, "Fast effective rule induction," in *Proc. of the* 12^{th} *Int. Conf. on Machine Learning (ICML-95)*, Tahoe City, CA, jul 1995, pp. 115–123.

[10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *ECCLT*, 1995, pp. 23–37.

[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java implementation*, Morgan Kaufmann, 2 edition, 2005.

[12] Marine Soret, Stephen L. Bacharach, and Irene Buvat, "Partial-Volume Effect in PET Tumor Imaging," *J Nuclear Medicine*, vol. 48, no. 6, pp. 932–945, 2007.