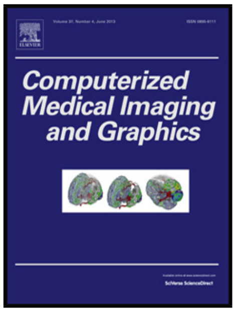
Accepted Manuscript



Title: Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier

Author: Desbordes Paul Ruan Su Modzelewski Romain Vauclin Sébastien Vera Pierre Gardin Isabelle

PII: S0895-6111(16)30132-X
DOI: http://dx.doi.org/doi:10.1016/j.compmedimag.2016.12.002
Reference: CMIG 1486

To appear in: *Computerized Medical Imaging and Graphics*

Received date: 29-4-2016
Revised date: 10-10-2016
Accepted date: 16-12-2016

Highlights :

- We propose a new feature selection strategy in two steps called GARF (Genetic Algorithm based on Random Forest) to select the most relevant subset of features from a large amount of characteristics extracted from positron emission tomography images and clinical data.
- A genetic algorithm is used to perform this selection according to a new multiparametric fitness function depending on a random forest misclassification rate, areas under receiver operating characteristic curves and a sparsity constraint.
- Experimental results show that excellent performances are obtained by our feature selection strategy compared to those obtained by 3 other methods.

Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier

Desbordes Paul[a,b], Ruan Su[a], Modzelewski Romain[a,c], Vauclin Sébastien[b], Vera Pierre[a,c], Gardin Isabelle[a,c]

[a]Litis - QuantIF, *University of Rouen, 22, boulevard Gambetta, 76000 Rouen, France*
[b]Dosisoft, *45/47, avenue Carnot, 94230 Cachan, France*
[c]*Henri Becquerel Centre, 1, rue d'Amiens, 76038 Rouen Cedex France*

Abstract

The outcome prediction of patients can greatly help to personalize cancer treatment. A large amount of quantitative features (clinical exams, imaging . . .) are potentially useful to assess the patient outcome. The challenge is to choose the most predictive subset of features. In this paper, we propose a new feature selection strategy called GARF (Genetic Algorithm based on Random Forest) extracted from Positron Emission Tomography (PET) images and clinical data. The most relevant features, predictive of the therapeutic response or which are prognoses of the patient survival 3 years after the end of treatment, were selected using GARF on a cohort of 65 patients with a local advanced oesophageal cancer eligible for chemo-radiation therapy. The most relevant predictive results were obtained with a subset of 9 features leading to a random forest misclassification rate of 18 ± 4% and an Areas Under the of Receiver Operating Characteristic (ROC) Curves (AUC) of 0.823 ± 0.032. The most relevant prognostic results were obtained with 8 features leading to an error rate of 20 ± 7% and an AUC of 0.750 ± 0.108. Both predictive and prognostic results show better performances using GARF than using 4 other studied methods.

Keywords: Feature Selection, Oesophageal cancer, Random Forest, Genetic Algorithm, Radiomics

1. Introduction

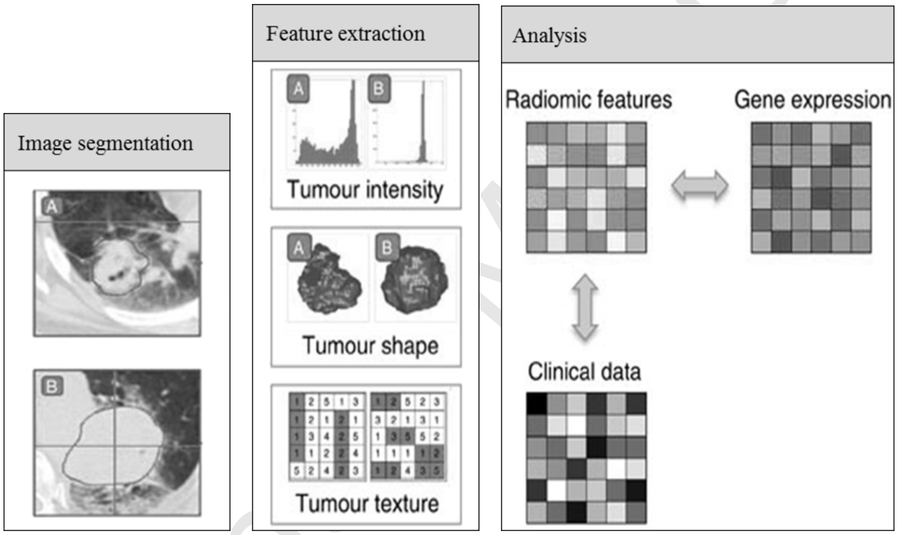
Outcome prediction is the foundation for tailoring and adapting a treatment planning in cancer therapy. Medical imaging plays a fundamental role in assessing the response to a treatment. In oncology, the Standard Uptake Value (SUV) of 18-FluoroDeoxyGlucose (FDG) measured by PET is widely used for diagnosing, staging and monitoring response to therapy [1]. Predictive (prediction of treatment response) and prognostic studies (prediction of the survival) using image features derived from 1st order statistics, such as Metabolic Tumour Volume (MTV) or Total Lesion Glycolysis (TLG: $\text{SUV}_{\text{mean}} \times \text{MTV}$), have been carried out. In solid tumours, predictive and prognostic values were found for these features [1].

More recently, other features have been proposed for describing ^{18}FDG uptake heterogeneity within the lesion. This heterogeneity would be representative the aggressiveness of the tumour. *El Naqa et al.* [2] have proposed to extract features from the SUV-Volume Histogram (SVH), such as SUV_x (minimum SUV of the $x\%$ highest SUV) and V_x (percentage volume having at least $x\%$ of SUV). These features were found to be relevant in studies of cervix and head and neck cancers. Furthermore, in this paper, *El Naqa et al.* have found that texture indices extracted from the Gray-Level Cooccurrence Matrix [3] (GLC matrix), characterising the intensity relationships between pairs of neighbouring pixels, are some of the most important predictive characteristics in cervix cancer. Other texture matrices have also been proposed in the literature, such as the Gray Level Difference Matrix [4] (GLD matrix) characterising the intensity differences between neighbours, the Gray Level Run Length Matrix [5] (GLRL matrix) and the Gray Level Size Zone Matrix [6] (GLSZ matrix) characterising the size ranges of intensities in one direction or in all the directions respectively. At the end, it is possible to extract several texture indices per matrix leading to a large number of characteristics.

Tixier et al. [7] studied the predictive value of 38 features, including texture indices, extracted from ^{18}FDG PET images on a cohort of patients with oesophageal cancer. Results were based on ROC and measurement of the associated AUC. Authors have found that GLC matrix features (second angular moment, local contrast, entropy, correlation, homogeneity and dissimilarity) and GLSZ matrix features (zone length non uniformity, gray level non uniformity) are relevant to predict patients' response to treatment. In Hatt et al. [8], 6 1^{st} order features were extracted for each tumour from a cohort of 45 patients with an oesophageal cancer treated by chemo-radiation therapy. After statistical analysis based on ROC curves, Kaplan Meier univariate analysis and Cox multivariate analysis, MTV is presented as a significant prognostic feature to predict overall survival at 1 year. In Tan et al., [9] 192 1^{st} order and texture features were extracted for each tumour from a cohort of 22 patients with an oesophageal cancer treated by chemo-radiation therapy. AUC from ROC curves were used to evaluate each feature. Three features resulting from the GLC matrix are predictive of the treatment response: inertia, correlation and cluster tendency (AUC ≥ 0.76).

The increasing number of features have led to the development of a new theory called Radiomics which supposed that quantitative analysis of medical images through automatic or semi-automatic software can provide more and better information than a practionner [10] (see Figure 1). However, an increasing number of features is not necessarily synonymous with performance improvement. *Orlhac et al.* [11] have shown that some texture indices are highly correlated with MTV on 3 types of tumours. In the same way, *Tixier et al.* [7] have shown that GLRL matrix features are highly correlated with GLSZ matrix ones. This redundancy of information is likely to reduce the prediction performances. Therefore, the correlations between features have to be considered during the selection process. Because of the high number of studied features and their nonlinear relationship with patient outcome (i.e. responder or non-responder), machine learning methods could be of great interest.

Globally, 3 types of feature selection methods can be distinguished [12]: filter, wrapper and embedded methods. The filter method is based on general properties such as the correlation between a feature and the prediction. RELIEF



(a) (b) (c)

Figure 1: Strategy for extracting radiomics data from images from [10]. (a) Experienced physicians delineate the tumours on images. (b) Features are extracted from the tumour volume, quantifying tumour intensity, shape and texture. (c) For analysis, the radiomics features are compared with other data.

(RELevance In Estimating Features) [13] [14] is considered as one of the most successful filter algorithms, where a margin-based criterion is used to rank the features. Authors in [15] propose the Feature Assessment by Sliding Thresholds (FAST) method, based on the AUC generated by sliding threshold values in one dimensional feature space.

The wrapper method evaluates the subsets of characteristics and detects the interaction between features. SFS (Sequential Forward Selection) [16] [17] and SFFS (Sequential Forward Floating Selection methods) [18] [17] are two representative wrapper algorithms. To solve the nesting problem of SFS, SFFS performs an exclusion step after each inclusion step. More recently, the Hierarchical Forward Selection (HFS) [19] was proposed. This approach selects subset of features based on intrinsic properties of the Support Vector Machine (SVM) [20] classifier.

Finally, the embedded method combines more closely the feature selection strategy, the model creation and evaluation. For example, CART (Classification And Regression Tree) method has a built-in mechanism to perform feature selection [21]. The expense is the loss of simple interpretability of the interactions of features. Guyon et al. [22] propose a selection method using SVM based on Recursive Feature Elimination (RFE). Another example of this approach is the Least Absolute Shrinkage and Selection Operator (LASSO) method [23] constructing a linear model, which penalizes the regression coefficient using a sparse constraint, shrinking many of them to zero. Any features which have non-zero regression coefficients are then selected.

Yet, these methods have not been widely used in the context of PET imaging and, generally, it is univariate and multivariate analyses that are used to study PET features [1]. From Computed Tomography (CT) images of 464 patients with a lung cancer, *Parmar et al.* have extracted 440 features to evaluate and compare the accuracy of 14 feature selection methods combined with 12 classifiers [24]. The best performances were found with the Random Forest algorithm (RF) [25]. The machine learning methods have also been used in single photon emission computed tomography imaging, Huertas-Fernández et al. [26] used a SVM classifier to develop a predictive model for Parkinsons disease.

In this paper, we propose a feature selection strategy in order to define the most relevant subsets of features allowing to predict the treatment response and the patient overall survival. The difficulty comes from the fact that the selection is performed on a large amount of different types of features (PET and clinical) which are heterogeneous. We don't have any *a priori* knowledge about the efficiency of the studied features. Hence, we first performed a Spearman's correlation analysis [27] to group correlated features and to choose the representative of each group. At the end of this first step, there is still an important number of features. Thus, we developed a new feature selection strategy based on a Genetic Algorithm (GA) [28] associated to a new multi-parametric fitness function taking into account a RF misclassification rate, AUC measurement and a sparsity constraint. This new strategy is called GARF. For this study, a database of 65 patients with an oesophageal cancer is studied. This method is compared to other feature selection methods (SFS, RFE, HFS and LASSO). This paper is organised as follows. Section 2 introduces the method developed to extract characteristics and to perform the feature selection strategy. Section 3 presents the experimental results on the patients' database followed by a discussion in Section 4.

2. Materials and Method

Our GARF method consists in the selection of the most relevant predictive and prognostic subsets of features among those previously extracted. Firstly, this section presents the database used. Secondly, our feature selection strategy is presented.

2.1. Image Data

In this retrospective study, data from 65 patients (N) with a locally advanced oesophageal cancer eligible for chemo radiation therapy are used to evaluate

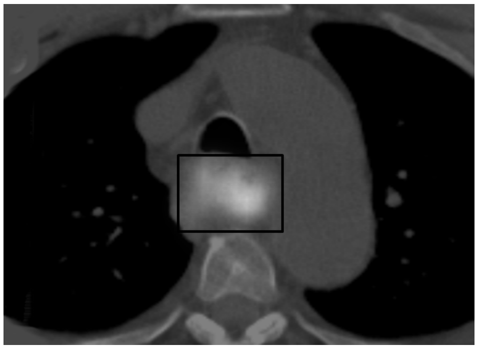


Figure 2: PET/CT slice of the chest with a framed esophageal tumour.

the feature selection strategy. Thirteen features are extracted from the patients' medical records (Table 1), such as patient's usual weight, disease stage, malnutrition evaluation, etc. All patients underwent a pre-treatment ^{18}FDG PET. All the images come from the same nuclear medicine department on a PET/CT Biograph Sensation 16 (Siemens, Erlangen, Germany). The voxel size is $4 \times 4 \times 2\text{mm}^3$. A PET/CT chest slice is given in Figure 2 as example.

Therapeutic response was evaluated one month after the end of treatment by clinical examination, endoscopy with biopsies and PET/CT imaging. Patients are separated into 2 categories: those with a Complete metabolic Response to treatment (CR) and non-responders or with residual disease (NCR). In our cohort, 41 patients (62%) are considered as CR, while 24 (38%) are considered as NCR. These data are used for the predictive study. For the prognostic study, the Overall Survival (OS) is estimated after a follow-up of 3 years after the end of treatment. At the end of the follow-up, 16 patients were alive (24%) and 50 had deceased (76%).

2.2. PET Feature Extraction

Taking into account the specificity of PET images, we proprose to use 45 features (see Table 1) extracted according to the following workflow. Firstly, the MTV is defined using a contrast-based adaptive threshold algorithm [29]. The mean MTV is 19.6 ± 20.5 cm^3 ($2.5 - 141$ cm^3). From this volume, 19 1$^{\text{st}}$ order and shape features are extracted such as SVH features [2], COV (Coefficient Of Variation [30]) or sphericity [31].

Secondly, 26 texture indices are extracted from 3 texture matrices: 10 from the GLC matrix (averaged over the 13 3D directions [32]), 5 from the GLD matrix and 11 from the GLSZ matrix. To compute these matrices a linear gray-level resampling is applied on the MTV according to [33] (see Equation 1):

$$I(i) = \text{round}\left(D \times \text{SUV}(i)\right) \tag{1}$$

where $I(i)$ is the new gray level value of voxel i with an initial intensity SUV(i) and D is the intensity step, set to 0.5.

Finally, these PET image features added to the 13 features extracted from the medical record lead to a number $F_i = 58$ multimodal initial features.

2.3. Feature Selection Strategy

Our feature selection strategy GARF is composed of 2 steps (see Figure 3) and can be defined as a wrapper method. Firstly, correlated features are eliminated after a Spearman's rank analysis performed on the 58 features. Secondly, the most relevant subsets of features are defined throught a genetic algorithm [28] with a multi-parametric fitness function based on a random forest classification [25], AUC measurement and a sparsity constraint.

2.3.1. Elimination of Correlated Features

As 58 features are extracted, the detection of the most relevant subsets by testing each combination is difficult. In order to keep uncorrelated features and eliminate redundant ones, we propose as a 1$^{\text{st}}$ step, a Spearman's rank correlation analysis [27] calculating the correlation coefficient ρ such as:

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} \tag{2}$$

Where N is the number of patients and d the difference between ranks of two features. Each patient's feature (such as tumour volume, patient's age or weight) are sort in an ascending order. So, the rank corresponds to their

Table 1: List of the initial tumour features.

Kind of features	Characteristics
Clinical	Patient age, Patient gender, Albumin level (g/l), NRI (Nutritional Risk Index), Malnutrition[*], Patient initial weight (kg), Usual weight (kg), Weight loss (%), Tumour location (up, mid, low), Histology (ADC or SCC), TNM stage, World Health Organisation (WHO) stage, Endoscopic tumour length (cm)
First order statistics	SUV_{max}, SUV_{mean}, SUV_{peak}, MTV, Sum of SUV, TLG, Standard Deviation (SD), COV, Sphericity, Skewness, Kurtosis, Energy, Entropy, SUV_{10}, SUV_{90}, $\text{SUV}_{10\text{-}90}$, V_{10}, V_{90}, $\text{V}_{10\text{-}90}$
Texture [**]	*GLCM* [3]: Variance, Energy, Entropy, Correlation, Dissimilarity, Contrast, Homogeneity, Inverse Differential Moment (IDM), Cluster Shade (CS), Cluster Tendency (CT) *GLSZM* [6]: Short Zone Emphasis (SZE), Long Zone Emphasis (LZE), Low Gray level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Short Zone Low Gray-level Emphasis (SZLGE), Long Zone Low Gray-level Emphasis (LZLGE), Short Zone High Gray-level Emphasis (SZHGE), Long Zone HighGray-level Emphasis (LZHGE), Zone Percentage (ZP), Gray Level Non Uniformity (GLNUz), Zone Length Non Uniformity (ZLNU) *GDLM* [5]: Coarseness, Contrast, Busyness, Complexity, Strength

[*] absence if NRI > 97.5, average if $83.5 \leq$ NRI ≤ 97.5 and severe if NRI < 83.5

[**] Mathematical expression of features in Table 1 of supplemental data from [11]

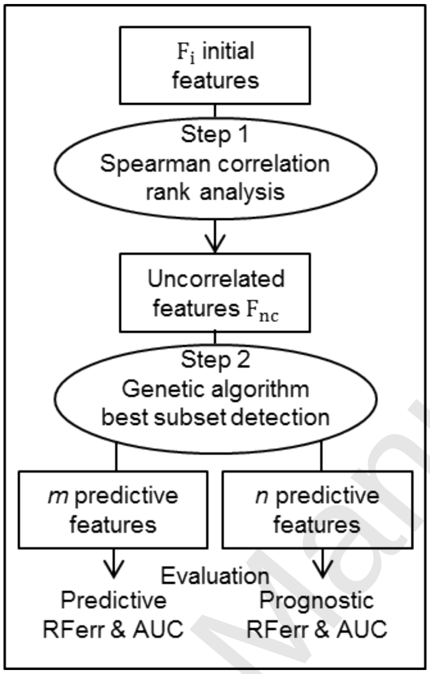


Figure 3: Feature selection strategy.

position in the sorted list. Features are compared one by one and considered as significantly correlated if the absolute value of the Spearman's correlation coefficient ($|\rho|$) is higher or equal to 0.8 with a p-value (p) smaller than 5% 170 [11]. Correlated features are placed in a group verifying these conditions. The mean $|\rho|$ value and the associated standard deviation are calculated for each group. Each correlation group is represented by the feature having the highest ρ-value compared to other members of its group. The total number of the selected features by this step, F_{nc}, is equal to the number of correlation groups. 175 Uncorrelated features represent groups of single features.

Among the remaining features in the second step, our feature selection strategy selects the most relevant prognostic and predictive subsets using a random forest method and a genetic algorithm.

2.3.2. Random Forest Classification

The RF method [25] is a machine learning technique which classifies data using decision trees. This method is widely used in many fields with interesting performances particularly in medical imaging [24]. The principle is to build a multitude (k) of independent trees built from an initial sample corresponding to N patients with F studied features. The initial training sample can be represented by a matrix of size (N, F). After construction, the forest is used for classification creating estimated labels (i.e. label 0 as responder, label 1 for non-responder). Then, the final labelling of patients is done using a majority vote from the forest.

Two random processes are used for the forest construction. Firstly, each tree of the forest is built from a bootstrap sample of k patients randomly picked with replacement. It means that a bootstrap can include several times the same patient's data. Secondly, from this sample, a decision tree is constructed as a binary tree. For each tree node, a subset of f features picked randomly among the F features is defined. f is equal to rounded $\sqrt{F}$ [34]. The most discriminating feature is picked in this subset according to the Gini criterion measuring the statistical dispersion [35]. This step is repeated for all nodes until all the observations are well separated with respect to their belonging (ground truth). A cross-validation method is used for the evaluation of the classifier based on a sampling technique. It means that the dataset is divided into K subsamples. $K-1$ are used to build the training samples, while the last one is considered as the test sample. This operation is realised K times with rotation in order to use each subsample as validation set. By comparing truth and estimated labels, it is possible to calculate accuracy for each test sample. The accuracy of the model corresponds to the mean and the standard deviation (SD) of the K computed accuracies.

2.3.3. Selection of the most relevant subsets of features

The aim of the GA is to converge to a solution minimising the score obtained by the fitness function, leading to the selection of the most relevant predictive subset of size m and the most relevant prognostic subset of size n. Thus, a multi-parametric fitness function is used, based on 3 criteria:

- Minimisation of the misclassification rate ($\mathrm{RF}_{\mathrm{err}}$) obtained by a random forest classification evaluated by a K-fold cross validation method with $K = 5$,
- Maximisation of the AUC measurement performed after a combination of features by logistic regression,
- Minimisation of the number of features using a sparsity constraint [36] (nbFeat).

According to these criteria the proposed expression of the fitness function, f_i, is:

$$f_i = \frac{\mathrm{nbFeat} + \alpha(1 - \mathrm{AUC}) + \beta \mathrm{RF}_{\mathrm{err}}}{\alpha + \beta + 1} \tag{3}$$

The sparsity constraint nbFeat is equal to $\dfrac{F_s \times \log(F_{nc}) - \log(F_{nc})}{F_{nc} \times \log(F_{nc}) - \log(F_{nc})}$, where F_s is the number of active chromosomes in the studied element. This constraint is normalised to have the same order of magnitude as the two other criteria. α and β are weights ($\in [0, 10]$) used to regulate the AUC measurement and the $\mathrm{RF}_{\mathrm{err}}$, respectively.

2.4. Classifier Evaluation Criteria

The final subsets of features are evaluated by RF classifications associated with K-fold cross validations ($K = 5$) and ROC curves. Se and Sp are used to compare the estimated and known labels according to the following expressions:

$$\mathrm{Se} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{4}$$

$$\mathrm{Sp} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \tag{5}$$

where True positive corresponds to the number of correct estimations of positive labels (label 1), True negative to correct estimations of negative label

(label 0), False negative to wrong estimations of positive labels and False positive to wrong estimations of negative labels.

3. Experimental Results

3.1. Spearman's Rank Correlation Analysis

Results of the Spearman's rank analysis of the 58 initial features, performed on the training sample, are given in Table 2. Concerning clinical data, patients' usual and current weights are correlated ($|\rho| > 0.96$). Likewise, the albumin level, the NRI and the malnutrition are highly correlated ($|\rho| > 0.84$). Only 9 PET images features are uncorrelated: COV, Skewness, Kurtosis, SUV_{90}, V_{10}, Contrast (GLD matrix), SZE, SZLGE and GLNUz. At the end of this correlation analysis, 9 groups of significant correlated features ($|\rho| \geq 0.8$, $p < 0.05$) are identified leading to a 1st selection of F_{nc} equal to 29/58 features (13 clinical and 16 from images).

3.2. Influence of the Coefficients

The expression of the GA fitness function contains 2 weight parameters α and β varying from 0 to 10. To find the optimal α and β-values minimizing the fitness function, three thousands experiments were done. For the predictive study, optimal α and β-values are 8 and 5, respectively, while for the prognostic study, α and β-values are both equal to 5. In Table 3 are given some examples of the values of the GA fitness function according to several α and β-values, as well as the corresponding AUC, RF misclassification rate and sparsity constraint.

3.3. Feature Selection Results

In the 2nd step, the GA parameters are set to 30 chromosomes and 30 generations. $nTrees = 500$ decision trees are used to build the RF algorithm in the GA fitness function. Higher values of $nTrees$ have been tested without any significant difference being observed. Figure 4 shows predictive results of the fitness function according to the generation with the optimal α and β-values.

Table 2: Groups of correlated features with the mean absolute value of the Spearman's correlation coefficient per group and the associated Standard Deviation (SD). The feature selected to represent each group for the next step is in bold.

Group	Correlated features	$\lvert\rho\rvert\pm$ SD
1	**Usual weight** - Current weight	0.96
2	**NRI** - Albumin level - Malnutrition	0.88 ± 0.04
3	$\mathbf{V_{10\text{-}90}}$ - V_{90}	0.96
4	**Energy** - Entropy	0.99
5	**MTV** - $_{\text{sum}}$SUV - TLG - Correlation (GLCM)	0.89 ± 0.03
6	**HGZE** - SUV_{10} - Variance - CT - SZHGE - SUV_{max} SUV_{peak} - SUV_{mean} - Complexity - SD - SUV_{10-90} Contrast (GLCM) - LGZE	0.91 ± 0.03
7	**IDM** - Homogeneity - LZE - ZP - Dissimilarity Energy (GLCM) - LZLGE - LZHGE - Strength Entropy (GLCM)	0.91 ± 0.03
8	**ZLNU** - Cluster Shade (GLCM)	0.81
9	**Busyness** - Coarseness - Sphericity	0.94 ± 0.02

Table 3: Examples of values of the fitness function (f_i) and the corresponding AUC measurement, RF misclassification rate and sparsity constraint according to several α and β-values. The best predictive and prognostic results are in bold.

	α	β	f_i	AUC	RF_{err} (%)	Sparsity constraint
predictive	0.8	0.8	0.198	0.844	21	0.21
	0.1	0.9	0.201	0.793	20	0.21
	0.7	0.1	0.214	0.821	19	0.25
	0.4	0.6	0.221	0.804	22	0.25
	8	**5**	**0.145**	**0.898**	**15**	**0.29**
prognostic	1.8	0.4	0.198	0.831	21	0.25
	0.4	0.5	0.201	0.821	22	0.21
	5	8	0.214	0.821	23	0.25
	1	0.3	0.220	0.827	23	0.25
	5	**5**	**0.169**	**0.833**	**15**	**0.25**

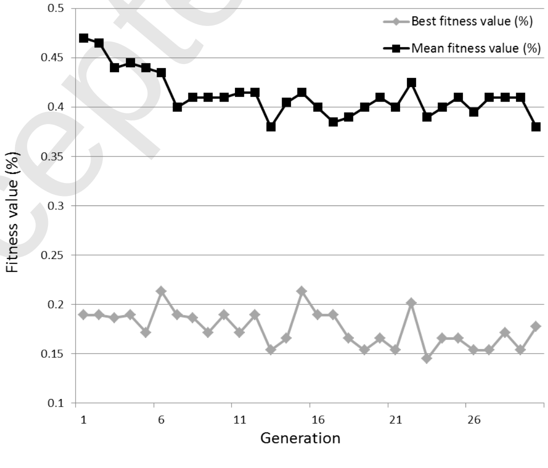


Figure 4: Predictive results of the GA fitness function according to the generation with the optimal α and β-value (8 and 5). The mean results for each generation are in black and the best results in grey.

Table 4: Evaluation of the step 1 impact on GARF for the predictive and the prognostic studies. Nb is the number of selected features and $\Delta\mathrm{RF}_{\mathrm{err}}$ is the misclassification rate difference between the methods.

	Method	Nb	$\mathrm{RF}_{\mathrm{err}}$ (%)	$\Delta\mathrm{RF}_{\mathrm{err}}$ (%)	Se (%)	Sp (%)
Predictive*	Without step 1	10	20	-2%	85	71
	With step 1	9	18		81	91
Prognostic**	Without step 1	9	25	-5%	85	60
	With step 1	8	20		88	72

*$\alpha = 8$ and $\beta = 5$

**$\alpha = 5$ and $\beta = 5$

Nine features are finally selected for the predictive study, it corresponds to patient usual weight (group 1), patient weight loss, disease histology, N stage, MTV (group 5), skewness, kurtosys, energy (group 4) and contrast from the GLD matrix. For the prognostic study, 8 features are selected, corresponding to patient age, disease location, stage, WHO stage, NRI (group 2), IDM from the GLC matrix (group 7), SZE from the GLSZ matrix and HGZE from the GLSZ matrix (group 6).

3.4. Comparison of Selection Methods

Firstly, to study the influence of Spearman's rank correlation analysis a comparison was done with and without the first step. Results of the evaluated criteria are given for the 2 methods in Table 4.

Four other feature selection methods are studied to be compared with GARF. These methods are LASSO [23], SFS [16], HFS [19] and RFE [22]. Both HFS and RFE approaches use a SVM classifier to select features. A radial basis function Gaussian kernel with $\sigma = 1$ and $C = 1$ is used. Like our method, the feature selection process started from the F_{nc} uncorrelated features. Furthermore, the same RF algorithm is used to evaluate these feature selection strategies.

Concerning the predictive study, 12 features are selected by the Lasso approach : the patient's age, the patient's gender, the patient's usual weight (group

1), the patient's weight loss, the location of the tumor, the T stage, the N stage, the WHO stage, the $V_{10\text{-}90}$ (group 3), the energy of the 1^{st} order (group 4), the IDM from the GLC matrix (group 7) and the contrast from the GLD matrix. Three features are selected by the HFS method: the MTV (group 5), the N stage and the contrast from the GLD matrix. Two features are selected by the SVM-SFS method: the MTV (group 5) and the IDM from the GLC matrix (group 7). Lastly, 3 features are selected by the RFE method: Kurtosis, Busyness from the GLD matrix (group 9) and GLNUz from the GLSZ matrix.

Concerning the prognostic study, 9 features are selected by the Lasso approach: the patient's age, the patient's weight loss, the location of the tumor, the N stage, the WHO stage, the NRI (group 2), the skewness, the energy of the 1^{st} order (group 4) and the IDM from the GLC matrix (group 7). Four features are selected by the HFS method: the MTV (group 5), the NRI (group 2), the IDM from the GLC matrix (group 7) and the contrast from the GLD matrix, 3 features selected by the SVM-SFS method: the patient gender, the MTV (group 5) and the GLNUz from the GLSZ matrix. Lastly, 3 features are selected by the RFE method: the MTV (group 5), the GLNUz from the GLSZ matrix and the HGZE from the GLSZ matrix (group 6).

In Table 5 are given results of the evaluation of the different subsets of features selected by the 4 studied methods.

Evaluation of the different subsets of features of size Nb obtained by our method GARF and the 4 others tested methods.

4. Discussion

Excellent performances are obtained by our GARF method with a classification accuracy and an AUC of 82% and 0.823 for the predictive study and 80% and 0.750 for the prognostic study (see Table 5). These results are consolidated by the comparison with the 4 other feature selection methods (Lasso, SFS, RFE, HFS). Our method always shows the most accurate results (see Table 5).

As step 1, a Spearman's rank correlation analysis is done in order to keep

Table 5: Means and standard deviations of the RF classifier misclassification rate, the sensitivity (Se), the specificity (Sp) and AUC evaluated by cross-validation. $\Delta\text{RF}_{\text{err}}$ is the misclassification rate difference between our method and the other ones and Nb the size of subset of studied features.

	Method	Nb	RF_{err} (%)	$\Delta\text{RF}_{\text{err}}$	Se (%)	Sp (%)	AUC
Predictive	**GARF**	**9**	**18±4**	/	**81±6**	**91±12**	**0.823±0.032**
	SFS	2	26±14	+8%	66±24	92±11	0.736±0.121
	HFS	3	26±8	+8%	71±6	87±13	0.723±0.118
	RFE	3	38±16	+20%	77±14	75±22	0.712±0.154
	Lasso	12	32±14	+14%	75±13	88±11	0.739±0.088
Prognostic	**GARF**	**8**	**20±7**	/	**88±15**	**72±23**	**0.750±0.108**
	SFS	3	32±6	+12%	73±24	68±30	0.635±0.133
	HFS	4	35±7	+15%	90±10	56±17	0.620±0.048
	RFE	3	46±12	+26%	71±19	74±19	0.635±0.091
	Lasso	9	31±12	+11%	88±18	76±16	0.760±0.162

uncorrelated features [11]. These correlations can explain different results found in the literature. For instance, energy and entropy from the GLC matrix which have similar outcomes in [2], are correlated in our study (group 7). An information complementarity between clinical and PET image features is shown by the fact that none of them are correlated (see Table 2). By comparing results of the RF classifications with and without step 1 (see Table 4), the importance of this step is demonstrated with an improvement for both predictive and prognostic studies with a $\Delta\text{RF}_{\text{err}}$ equal to -2% and -5%, respectively. Indeed, no additional information is provided by correlated features, moreover this redundancy could mislead the feature selection method. So, the results are improved by the removal of redundant features ($F_{nc} = 29/58$). Tixier et al. [7] have shown on an oesophageal cancer database that GLRL matrix features are highly correlated with GLSZ matrix ones justifying the fact that this matrix is not used in this study. In this study a threshold value of the Spearman's correlation coefficient of 0.8 is used according to Orlhac et al. [11]. It could be interesting to analyse the influence of this threshold value on GARF performances.

Concerning the GA fitness function, in both studies, α and β-values are much higher than 1. It means that their contribution is more important than the sparsity constraint. For the predictive study, α-value is higher than β-value (8 versus 5) showing that AUC measurement has an higher contribution than the RF misclassification rate. This result is different concerning the prognostic study because both α and β-values are equal to 5 meaning that these parameters have the same influence.

Concerning the most relevant features selected, the MTV (group 5) seems to be important because it is present in 3/5 predictive results (exept LASSO and RFE). This confirms the particular interest of this feature in the patient's treatment monitoring [1]. Instead, the $\mathrm{SUV}_{\mathrm{max}}$ (group 6), which is generally regarded as an important feature, is not relevant in predictive studies. Concerning the best prognostic features, three methods select the MTV (except our method and Lasso), while our method selects HGZE corresponding to the same group than $\mathrm{SUV}_{\mathrm{max}}$ (group 6).

Otherwise, we note the presence of an important part of clinical features in the GARF predictive (4/9) and prognostic (5/8) subsets. They play an important role in patient outcome. The fact of combining these clinical features with PET image features (1^{st} order and texture indices) can improve outcomes. These results show that these features brought additional information to clinical data. We can also assume that it can be difficult to predict long term events from the PET exams performed before the beginning of the treatment. It could be also interesting to look at the longitudinal analysis of image features evolution between the initial PET exam and another performed during the treatment as it has already been done in other studies [1], even if it requires performing a second examination.

GARF has been compared to 4 other standard selection feature strategies that are well known to achieve great success in feature selection and classification. Two of them are wrapper methods (SFS and HFS) and the two others are embedded (LASSO and RFE). Two of the compared methods are deeply associated with SVM. Indeed, HFS and RFE use intrinsic properties of the SVM in their feature selection process. Thus, it is not possible to distinguish between these methods and SVM, the one that is responsible for high misclassification rate, but whatever the selection feature strategy used, they were all evaluated using the same RF algorithm for comparison purpose. The comparison between GARF results with those obtained by other studied methods shows that our strategy is more efficient with a gain of at least 8% for the predictive study and 11% for the prognostic one (see Table 5). Concerning predictive study, the best results are clearly obtained by our method on all the evaluated criteria: AUC = 0.823, RF_{err}=18%, Se=81% and Sp=91%, whereas the second best results are found with LASSO: AUC = 0.739, RF_{err}=32%, Se=75% and Sp=88%. Similarly, in the prognostic study, our method shows the best results: AUC=0.750, RF_{err}=20%, Se=88% and Sp=72% whereas the others, except the LASSO approach, are far behind. Indeed, this last have a similar AUC (0.760), Se (88%) and Sp (76%), but a worse RF_{err} (31%). High standard deviation specificity values obtained in this study by all the methods can be explained by the fact that the number of patients in the surviving group is small (16/65), making machine learning more challenging than if the two outcome groups were similar.

5. Conclusion

To conclude, we have shown that GARF, our feature selection method, improves the outcome prediction compared to other tested methods by at least 8% for predictive study and 11% for the prognostic one. These good results are confirmed by other evaluation criteria (AUC, sensitivity and specificity). Machine learning techniques, and particularly RF, provide a useful expertise in the selection of subsets of multimodal features.

To further evaluate GARF, it remains necessary to test it on a larger patients cohort to assess its robustness and also to study the influence of the threshold value of the Spearman's correlation coefficient. Moreover, in the future it might be interesting to test this selection on other types of cancer, such as lung cancer and lymphoma.

Conflict of interest statement

This work was partly funded by ANRT (French National Association for Research and Technology) under the contract ANRT n2013/1153. No potential conflict of interest relevant to this article is reported.

References

[1] C. Van De Wiele, V. Kruse, P. Smeets, M. Sathekge, A. Maes, Predictive and prognostic value of metabolic tumour volume and total lesion glycolysis in solid tumours, European Journal of Nuclear Medicine and Molecular Imaging 40 (2) (2013) 290–301. doi:10.1007/s00259-012-2280-z.

[2] I. El Naqa, P. W. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W. L. Thorstad, J. O. Deasy, Exploring feature-based approaches in PET images for predicting cancer treatment outcomes, Pattern Recognition 42 (6) (2009) 1162–1171. doi:10.1016/j.patcog.2008.08.011.

[3] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification (1973). doi:10.1109/TSMC.1973.4309314.

[4] M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Transactions on Systems, Man and Cybernetics 19 (5) (1989) 1264–1273. doi:10.1109/21.44046.

[5] M. M. Galloway, Texture analysis using gray level run lengths, Computer Graphics and Image Processing 4 (2) (1975) 172–179. doi:http://dx.doi.org/10.1016/S0146-664X(75)80008-6.

[6] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, J.-l. Mari, Texture Indexes and Gray Level Size Zone Matrix Application to

Cell Nuclei Classification, Pattern Recognition and Information Processing (2009) 140–145doi:10.1142/S0218001413570024.

[7] F. Tixier, C. Cheze-Le Rest, M. Hatt, N. M. Albarghach, O. Pradier, J.-P. Metges, L. Corcos, D. Visvikis, Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer, Journal of Nuclear Medicine 52 (3) (2011) 369–378. doi:10.2967/jnumed.110.082404.

[8] M. Hatt, D. Visvikis, N. M. Albarghach, F. Tixier, O. Pradier, C. Cheze-Le Rest, Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology, European Journal of Nuclear Medicine and Molecular Imaging 38 (7) (2011) 1191–1202. doi:10.1007/s00259-011-1755-7.

[9] S. Tan, H. Zhang, Y. Zhang, W. Chen, W. D. D'Souza, W. Lu, Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in 18F-FDG uptake patterns., Medical physics 40 (10) (2013) 101707. doi:10.1118/1.4820445.

[10] P. Lambin, E. Rios-Velazquez, R. T. H. Leijenaar, S. Carvalho, R. G. P. M. Van Stiphout, P. Granton, C. M. L. Zegers, R. J. Gillies, R. Boellaard, A. Dekker, H. J. W. L. Aerts, Radiomics: Extracting more information from medical images using advanced feature analysis, European Journal of Cancer 48 (4) (2012) 441–446. doi:10.1016/j.ejca.2011.11.036.

[11] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, 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 55 (3) (2014) 414–422. doi:10.2967/jnumed.113.129858.

[12] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers and Electrical Engineering 40 (1) (2014) 16–28. doi:10.1016/j.compeleceng.2013.11.024.

[13] K. Kira, L. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: AAAI-92 Proceedings, 1992, pp. 129 – 134. doi:10.1016/S0031-3203(01)00046-2.

[14] R. Gilad-Bachrach, A. Navot, N. Tishby, Margin based feature selection - theory and algorithms, in: Proceedings of the 21st International Conference on Machine Learning, 2004.

[15] X.-w. Chen, M. Wasikowski, FAST: a roc-based feature selection metric for small samples and imbalanced data classification problems, Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08 (2008) 124—-132doi:10.1145/1401890.1401910.

[16] A. W. Whitney, A Direct Method of Nonparametric Measurement Selection, IEEE Transactions on Computers C-20 (9) (1971) 1100–1103. doi:10.1109/T-C.1971.223410.

[17] S. Theodoridis, K. Koutroumbas, Introduction to Pattern Recognition: A Matlab Approach, 2010.

[18] P. Pudil, J. Novovieova, J. Kittler, Pattern Recognition Letters, Pattern Recognition Letters 15 (June 1993) (1994) 1119–1125. doi:10.1016/j.patrec.2010.09.010.

[19] H. Mi, C. Petitjean, B. Dubray, Robust Feature Selection to Predict Tumor Treatment Outcome., Artif Intell Med 64 (3) (2015) 195–204. doi:10.1016/j.artmed.2015.07.002.

[20] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning 20 (3) (1995) 273–297. arXiv:arXiv:1011.1669v3, doi:10.1023/A:1022627411411.

[21] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, Monterey, CA, 1984.

[22] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines., Machine Learning 46 (4) (2002) 389–422. doi:10.1109/5254.708428.

[23] R. Tibshirani, Regression shrinkage and selection via the lasso., Journal of the Royal Statistical Society. Series B (Methodological 58 (1) (1996) 267–288.

[24] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H. J. W. L. Aerts, Machine Learning methods for Quantitative Radiomic Biomarkers (Supplement), Scientific reports 5 (2015) 13087. doi:10.1038/srep13087.

[25] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[26] I. Huertas-Fernandez, F. J. Garcia-Gomez, D. Garcia-Solis, S. Benitez-Rivero, V. A. Marin-Oyaga, S. Jesus, M. T. Caceres-Redondo, J. A. Lojo, J. F. Martin-Rodriguez, F. Carrillo, P. Mir, Machine learning models for the differential diagnosis of vascular parkinsonism and Parkinson???s disease using [123I]FP-CIT SPECT, European Journal of Nuclear Medicine and Molecular Imaging 42 (1) (2014) 112–119. doi:10.1007/s00259-014-2882-8.

[27] C. Spearman, The Proof and Measurement of Association between Two Things, The American Journal of Psychology 15 (1) (1904) 72–101.

[28] J. Holland, Adaptation in natural and artificial systems, MIT Press, Cambridge, MA, USA.

[29] S. Vauclin, K. Doyeux, S. Hapdey, A. Edet-Sanson, P. Vera, 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 54 (22) (2009) 6901–6916. doi:10.1088/0031-9155/54/22/010.

[30] R. A. Bundschuh, J. Dinges, L. Neumann, M. Seyfried, N. Zsótér, L. Papp, R. Rosenberg, K. Becker, S. T. Astner, M. Henninger, K. Herrmann, S. I. Ziegler, M. Schwaiger, M. Essler, Textural Parameters of Tumor Heterogeneity in 18F-FDG PET/CT for Therapy Response Assessment and Prognosis in Patients with Locally Advanced Rectal Cancer, Journal of Nuclear Medicine 55 (6) (2014) 891–897. doi:10.2967/jnumed.113.127340.

[31] F. Hofheinz, A. Lougovski, K. Zöphel, M. Hentschel, I. G. Steffen, I. Apostolova, F. Wedel, R. Buchert, M. Baumann, W. Brenner, J. Kotzerke, J. Van Den Hoff, Increased evidence for the prognostic value of primary tumor asphericity in pretherapeutic FDG PET for risk stratification in patients with head and neck cancer, European Journal of Nuclear Medicine and Molecular Imaging 42 (3) (2014) 429–437. doi:10.1007/s00259-014-2953-x.

[32] M. Hatt, M. Majdoub, M. Vallières, F. Tixier, C. Cheze-Le Rest, D. Groheux, E. Hindié, A. Martineau, O. Pradier, R. Hustinx, R. Perdrisot, R. Guillevin, I. El Naqa, D. Visvikis, 18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a MultiCancer Site Patient Cohort, Journal of Nuclear Medicine 56 (1) (2015) 38–44. doi:10.2967/jnumed.114.144055.

[33] F. Orlhac, M. Soussan, K. Chouahnia, E. Martinod, I. Buvat, 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer, PLoS ONE 10 (12) (2015) 1–16. doi:10.1371/journal.pone.0145063.

[34] R. Genuer, J.-m. J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognition Letters 31 (14) (2010) 2225–2236. doi:10.1016/j.patrec.2010.03.014.

[35] C. Gini, Measurement of Inequality of Incomes, The Economic Journal 31 (121) (1921) 124–126. doi:10.1017/CBO9781107415324.004.

[36] B. Bhanu, Y. Lin, Genetic algorithm based feature selection for target detection in SAR images, Image and Vision Computing 21 (7 SPEC.) (2003) 591–608. doi:10.1016/S0262-8856(03)00057-X.