Leveraging RSF and PET images for prognosis of Multiple Myeloma at diagnosis

Ludivine Morvan, Thomas Carlier, Bastien Jamet, Clément Bailly, Caroline M Bodet-Milin, Philippe Moreau, Francoise M Kraeber-Bodere, Diana Mateus

Leveraging RSF and PET images for prognosis of Multiple Myeloma at diagnosis

Ludivine Morvan[1,2] · **Thomas Carlier**[2,3] · **Bastien Jamet MD**[3] · **Clément Bailly MD**[2,3] · **Caroline Bodet-Milin MD**[2,3] · **Philippe Moreau MD** [3,4] · **Françoise Kraeber-Bodéré MD**[2,3] · **Diana Mateus**[1]

The final publication is available in IJCARS (International Journal of Computer Assisted Radiology and Surgery) at *https://doi.org/10.1007/s11548-019-02015-y*.
Received: 3 February 2019 / Accepted : 11 June 2019

Abstract Purpose Multiple myeloma (MM) is a bone marrow cancer that accounts for 10% of all hematological malignancies. It has been reported that FDG PET imaging provides prognostic information for both baseline and therapeutic follow-up of MM patients using visual analysis. In this study, we aim to develop a computer-assisted method based on PET quantitative image features to assist diagnoses and treatment decisions for MM patients.

Methods Our proposed model relies on a two-stage method with Random Survival Forest (RFS) and Variable importance (VIMP) for both feature selection and prediction. The targeted variable for prediction is the progression-free survival (PFS). We consider texture-based (radiomics), conventional (e.g. SUVmax) and clinical biomarkers. We evaluate PFS predictions in terms of C-index and final prognosis separation in two risk groups, from a database of 66 patients who were part of the prospective multi-centric french IMAJEM study.

Results Our method (VIMP+RSF) provides better results (1-C-index of 0.36) than conventional methods such as Lasso-Cox and Gradient-Boosting Cox (0.48 and 0.56 respectively). We experimentally proved the interest of using selection (0.61 for RSF without selection) and showed that VIMP selection is more stable and gives better results than Minimal-depth and Variable-Hunting (0.47 and 0.43). The approach gives better prognosis group separation (a p-value of 0.05 against 0.11 to 0.4 for others).

Conclusion

Our results confirm the predictive value of radiomics for MM patients, in particular, they demonstrate that quantitative/heterogeneity image-based features reduce the error of the predicted progression.

To our knowledge, this is the first work using RFS on PET images for the progression prediction of MM patients. Moreover, we provide an analysis of the feature selection process, which points towards the identification of clinically relevant biomarkers.

Keywords Random Survival Forest · Multiple Myeloma · Variable Selection · Radiomics · PET imaging

1 Introduction

Multiple Myeloma (MM) is characterized by the clonal proliferation of malignant plasma cells in the bone marrow (BM) and accounts for about 10-15% of hematological malignancies. 18F-FDG PET is useful for initial staging and therapeutic monitoring in MM [13]. An example of full-body FDG PET imaging is presented in Fig. 1. Image interpretation relies mostly on visual analysis for detecting the lesions, as well as on the extraction of simple quantitative measurements such as the standard uptake value of the maximum intensity voxel within the lesion, also named SUVmax. These measurements are then used together with clinical biomarkers to provide a prognosis. Our long term aim is to develop a computer-aided system capable of assisting physicians in the prognosis task.

Recently, more advanced image-based measurements have been suggested for PET image analysis, which

1. Ecole Centrale de Nantes, Laboratoire des Sciences Numériques de Nantes (LS2N), CNRS UMR 6004, Nantes, France. E-mail: ludivine.morvan@ls2n.fr
2. CRCINA, INSERM, CNRS, University of Angers, University of Nantes, Nantes, France.
3. University Hospital of Nantes, Nuclear Medicine Department, Nantes, France
4. University Hospital of Nantes, Haematology Department, Nantes, France

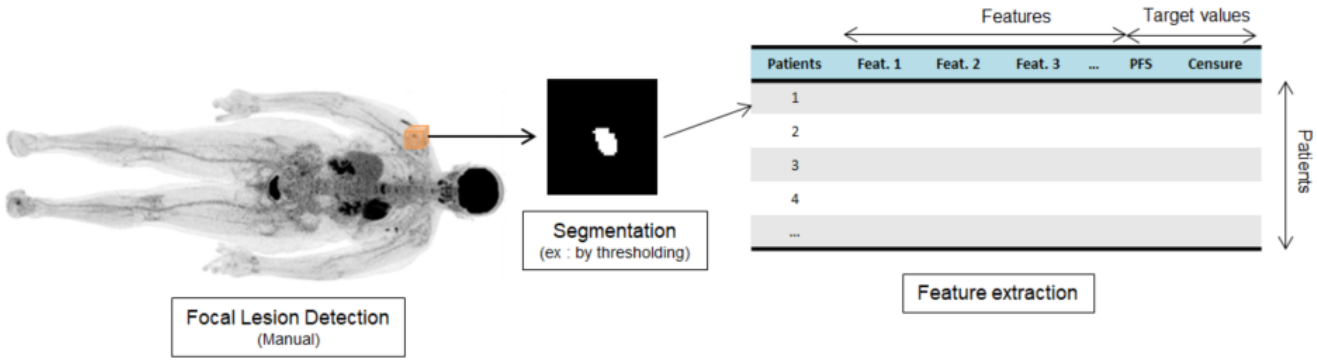


Fig. 1: Example of PET image of a multiple myeloma patient and processing of the radiomics features. Physicians localise lesions and segmentation is processed before extraction of radiomics features.

quantify the intra-tumoral heterogeneity of the lesions [22]. The consideration of such features for survival analysis is currently referred as *radiomics*. Their computation is largely inspired from standard image-texture analysis tools [7]. There is today an increasing interest in determining the value of such image-based quantitative features for predicting the progression and/or the outcome of patients.

The radiomics field is faced with several challenges. First, there are numerous possibilities on how to compute image-based features, so standardization is needed.[22] But it is not yet clear which of these computational approaches might be the most predictive. Moreover, it is necessary to combine image-based features with other clinical information (age, sex, calcemia etc.). Considering multiple feature-extraction implementations and the concatenation of image-based and clinical features leads to high-dimensional vectors and calls for the use of feature-selection methods or intelligent approaches to handle them.

A second important aspect in determining the value of radiomics for prognosis is the model used for prognosis prediction.

We opt for the Random Survival Forest (RSF) method, introduced by Ishwaran [8]. RSF is a machine learning approach for survival analysis based on decision-tree ensembles, which has demonstrated robustness to censoring data[1] and noisy variables. Moreover, RSF is suitable for high-dimensional and multi-modal input data and naturally avoids overfitting.

In this work, we propose a method to predict the PFS (Progression Free Survival) given both clinical variables and radiomics features extracted from PET images. The framework consists of a cascade of two RSF blocks: one for feature selection and the second for prognosis prediction The contributions of our work are:

- While feature selection and predictive models are usually considered separately, we propose here a unified model for variable selection and survival analysis.
- Our experimental results show that with the proposed method, it is possible to exploit radiomics features for the estimation of the PFS, confirming their predictive value.
- We illustrate how the RSF for feature selection finds relevant known clinical and conventional biomarkers and points to potentially new image-based ones.
- We are the first to investigate the use of RSF in the context of MM and PET imaging.

Our experimental validation on a clinical prospective database, demonstrates the improvement of the proposed method in terms of prediction error (1 - C-index [2]) on the predicted mortality (related on the PFS in our case) when compared to the Gradient-Boosting Cox [14] or Lasso-Cox approaches [16]. We also discuss a comparison of several RSF-based variable selection methods: Variable Importance (VIMP), Minimal Depth (MD) and Variable Hunting (VH). Then, we demonstrate the interest of using a combination of clinical features with imaging features for better prediction. Finally, we highlight the possibility to select

[1] Right censoring occurs when no event (death/progression) has taken place at the end of the evaluation period.

[2] The concordance index is the frequency of concordant pairs among all pairs of subjects.

the best features, and more precisely, the best computational approaches.

2 Related work

Classical statistical methods for survival analysis include Cox Proportional-Hazards Model [5] and Kaplan-Meier curves [10]. These methods have the advantage of being well understood and easy to interpret, but have some limitations. Regarding Kaplan-Meier models, there is no possibility to evaluate the individual survival of a single patient or to evaluate the impact of co-variables on the survival. Moreover, Cox models raise two problems: results can be biased when censoring is linked to the variables[21], and their statistical power is impacted by a high censoring rate. Such limitations motivate the use of more complex models and machine learning methods, such as tree ensembles.

In the past years, the use of RSF, a random forest method adapted to survival analysis [8], has increased to the point of becoming a new reference for survival analysis, due to its robustness to censoring data and noisy variables. Contrary to Cox, Survival trees are both able to detect a shift in non-linear relations (between variables and predicted target values), but also to detect variable interactions. Yan Zhou et al. [21] showed that Breiman's survival forest leads to better predictions than Cox regression, in particular for cases dealing simultaneously with a small sample size and a high number of variables. These important arguments motivate our use of RSF.

For most prior work, the input to a survival analysis method of choice has been mostly restricted to clinical variables. However, with the increase of image-data, radiomics features computed on CT, MRI, PET or other image modalities have brought additional valuable information [6,12], which is currently being investigated for a large variety of clinical applications, e.g. for lung [11] or head-neck cancer [17]. However, apart from Bühnemann et al. [3] who used confocal images of tissue microarrays, RSF and radiomics association has not been thoroughly investigated. We explore in this work their combination in the context of MM.

We promote here the use of multi-modal features, i.e. clinical and image-based, at the cost of increasing the number of variables that survival models need to consider. Moreover, for image-based features, several computation techniques are possible. With no manual pre-selection of features and with the typical size of prospective clinical databases, we are confronted with a high-dimensional problem with more features than patients. To tackle this issue, several classical feature selection methods exist like partial least squares or Cox-regression under lasso penalization [16]. More recently, the use of the concordance index for feature selection combined with a boosting and gradient boosting Cox models was shown to have a good performance under a comparative study of several machine learning and selection methods for survival [18]. However, at the core of RSF methods the randomized optimization performs a selection of features at every node. This fact was exploited by Ishwaran et al. [9], who proposed three variable selection methods: VIMP, MD, and VH. In the interest of keeping the method interpretable and to retain features that are coherent with the predictive model, we propose a framework where both tasks, the variable selection and the survival prediction model are based on RSF.

Regarding the survival analysis of MM patients using radiomics, a preliminary study [4] showed the interest of intra-tumoral heterogeneity for prognosis using FDG-PET images at diagnosis. Other recent studies have reported also the prognostic value of SUVmax, total metabolic tumor volume and whole body total glycolysis volume using PET imaging [20][23].

With respect to the use of machine learning methods for MM survival, only a few recent studies exist, which focus on detection and segmentation of focal lesions [19] or the use of genic expression as features [1]. To our knowledge, we are the first to explore radiomics for MM, and to use RSF both as a biomarker selection and a prognostic tool for MM prognosis prediction.

3 Method

In this work, we propose to follow a machine learning approach to predict the PFS value for a new patient. Towards this goal, we build a unified framework to:

- deal with the large number of clinical and image-based features (in the order of one hundred),
- derive the most relevant features for the prediction,
- predict the progression of a patient given his/her personal data (clinical and image-based features).

The proposed framework consists of two stages, as illustrated in Fig. 2: i) an automatic variable selection, and ii) a survival analysis stage. Both steps rely on Random Survival Forests [8], a modified version of the more traditional Random Forests [2], which handles right censored data and forms clusters of similar survival curves in the leaves. In the first stage (green block in Fig. 2), the feature selection RSF is trained on all the provided clinical and image-based variables. A VIMP analysis of the split functions selected during the

randomized optimization of the RSF's nodes is then deployed to identify the most predictive features. In the second stage (pink block in in Fig. 2), the prognosis RSF is then trained on the features selected during the first stage. The resultant model is then used for personalized "mortality" prediction during testing. Since we are interested in PFS, the RSF mortality refers, for the remaining of this work, to the expected total number of a progression. Our method provides each new patient with a survival curve, a mortality rate and a prognosis group. Moreover, at the population level, the first stage (predictive feature selection) serves as a good starting point for the identification of potential biomarkers (variables allowing to predict the target value).

In the following we provide the method's details. Section 3.1 describes the extracted features. Section 3.2 recalls concepts from survival analysis and explains the RSF method. Section 3.3 describes the different compared feature selection methods and Section 3.4 details the prediction stage.

3.1 Feature collection and extraction

Two main types of features were extracted: *clinical* and *image-based*, both derived from the PET exam at diagnosis which is our baseline. Each lesion was delineated using a majority vote approach involving three simple segmentation methods: SUV 2.5, SUVmax 40% and K-means with 2 classes. Intra-tumoral heterogeneity/texture features were computed on the most intense focal lesion (FL). Features extracted from PET images were further categorized into two sub-categories: *conventional* (quantitative measurement performed on PET imaging but not based on intra-tumoral heterogeneity) and *textural features*.

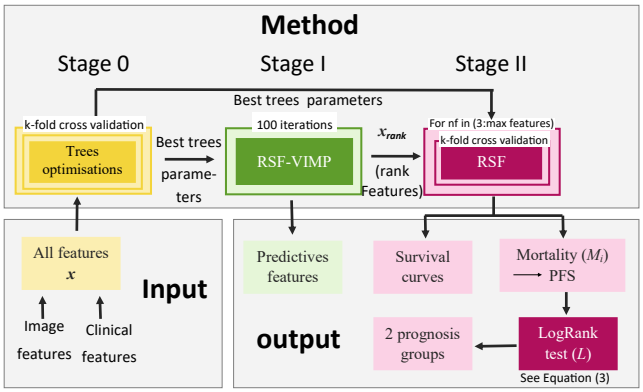


Fig. 2: Proposed approach for the prognosis prediction of multiple myeloma from clinical, conventional and image-based features. The yellow block corresponds to the stage 0 (optimisation), the green block to the stage I (feature selection) and the pink block to the stage II (prognosis prediction).

We follow the standard radiomics approach in order to extract agnostic quantitative descriptors of the textural/heterogeneity information [22] of the segmented lesions. These descriptors can be of: **1st order**, describing the value distribution of individual voxels without taking account their spatial relationships, such as the SUVmax or MTV (metabolic total volume); or of **2nd order**, describing the texture or relationship between voxels, and calculated from different matrices such as the co-occurrence matrix, e.g. GLCM (Gray Level Co-Ocurence Matrix).

A summary of all the features can be found in Table 1. In total, we have thirteen clinical, six conventional and 110 textural features. Note that the computation of textural features include different implementations of the same concepts, in the aim to study if a given implementation increases the prediction value of a feature. These implementations consider options for normalization (absolute, relative or histogram equalization), and voxel equalisation (equal size or not) [22]. The correlation among different implementations of the same feature will be dealt with a 100 VIMP-RSF stage run.

Table 1 List of clinical and conventional features (first row) and the textural features (second and third rows). All features have been calculated with 6 different implementations. ** : image-based features*

Conventional*	**Clinical features**	**Clinical features***
FL SUVmax	Age	Number of FL(PET baseline)
Bone Marrow Involvement SUVmax	Sex	Presence of extramedullary disease
Total MTV	Hemoglobin	FL Deauville score
MTV FL	Calcemia	Bone Marrow Involvement Deauville score
Total TLG (Total lesion glycolysis)	Creatinine	Global Deauville score
TLG FL	R-ISS	Number of FL (MRI)
	Treatment arm	
GLCM* (Gray Level Co-Ocurence Matrix)	**GLRLM*** (Gray Level Run-Length Matrix)	**GLSZM*** (Gray Level Size-Zone Matrix)
Homogeneity	HGRE (High Gray Level Run Emphasis)	HGZE (High Gray Level Zone Emphasis)
Entropy	LGRE (Low Gray Level Run Emphasis)	ZLNU (Size-Zone Non-Uniformity)
Energy	SRE (Short Run Emphasis)	SZHGE (Small Area High Gray Level Emphasis)
Correlation	LRE (Long Run Emphasis)	LZLGE (Low Gray Level Zone Emphasis)
Contrast		SZE (Small Area Emphasis)
Dissimilarity		ZP (Zone percentage)
		RP (Run percentage)
First order*		
Maximum		

All features are concatenated in a feature vector $\mathbf{x} \in \mathbb{R}^D$ where D is the total number of features.

3.2 Random Forests for Survival Analysis (RSF)

A *Random Forest* is a collection of N_{trees} decision trees $\mathcal{T} = \{f_1, \ldots, f_{N_{\text{trees}}}\}$ trained to predict a target value y given an input feature vector. The target y can be a

categorical or continuous variable. One decision tree f consists of a series of nodes h, each characterized by a binary split decision function ϕ and a threshold th. A set of data points $\mathcal{S}$ reaching a node is divided in two subsets, and points are pushed either into the left $\mathcal{S}_l$ and right $\mathcal{S}_r$ children according to the split function ϕ. Often, ϕ is axis-aligned, selecting one among the different dimensions of the feature space R^D, such that, $\mathcal{S}_l = \{\mathbf{x} \in \mathcal{S} | \phi(\mathbf{x}) \leq th\}$ and $\mathcal{S}_r = \{\mathbf{x} \in \mathcal{S} | \phi(\mathbf{x}) > th\}$. Given a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^{N_{\text{train}}}$, randomized node optimization is used during training to find the best $\{\phi, th\}$ pairs for each node, according to certain criteria, e.g. the Gini index. Nodes are grown until a termination criteria is reached, such as the maximal depth or a minimum number of data points per node. A terminal node, called a *leaf*, stores the estimated target value from the training points falling into it. During testing, a new feature vector is conducted down the nodes of each of the trees, and assigned a target value per tree. The final prediction results from the aggregation of the tree-wise predictions.

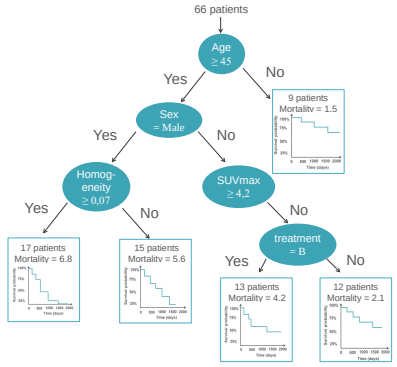


Fig. 3: Example of survival tree. Each leaf node stores i) the mortality (a high mortality means a bad prognosis), and ii) a survival curve (the survival probability as a function of time).

We now recall important concepts of Survival Analysis in order to later introduce the adaption of RFs to survival. Survival Analysis is caracterised by a dataset $\{\mathbf{x}_i, \theta_i, \tau_i\}_{i=1}^{N_{\text{train}}}$,where τ_i stands for the the time to the event of interest and θ_i is a binary variable indicating censorship (which is equal to 1 if an event occurred during the study period and 0 if not). The aim of the analysis is usually to find out variables that allow splitting the population into groups related to $\{\theta_i, \tau_i\}$ that have different mortality with events occurring at similar times. The target value y_i is the expected mortality, which in the application of interest is related to the PFS. The mortality M can be interpreted as the the expected total number of deaths (here the number of patients with a relapse or disease progression). In practice, the mortality is derived from the cumulative hazard function (CHF): a function of time collecting the time to event data τ_i for the individuals i in the training set. Formally, the Nelson-Aalen estimator $\hat{H}(t)$ for the CHF is:

$$\hat{H}(t) = \sum_{\tau_k \leq t} \frac{d_k}{Y_k}, \tag{1}$$

where k is the index of the event time between 1 and the total number of distinct event times m, d_k stands for the number of events up to time τ_k and Y_k is the number of individuals at risk at time τ_k. $\hat{H}$ can be interpreted as the sum of death rates at each event time. The mortality M_i of an individual i is the sum of the CHF over each unique time:

$$M_i = \sum_{k=1}^{m} \hat{H}(\tau_k | X_i). \tag{2}$$

In other words, M_i measures the number of deaths expected under a null hypothesis of similar survival behavior.

The *Random Survival Forest* is an ensemble-tree method, introduced by Ishwaran in 2008 [8] to adapt the Random Forests to right-censored data and survival analysis. As before, tree growing is done through randomized node optimization, generating several candidates for feature (axis aligned split) and threshold $\{\phi, th\}$. However,the input now includes censorship, and instead of an information theoretic criteria, the split function and threshold are chosen to maximize the survival difference between the individuals going to the two daughter nodes. In particular, the criteria to evaluate the best population split at each node is the *log-rank* L, defined as:

$$\max_{\phi, th} L = \max_{\phi, th} \frac{\sum_{k=1}^{m} (d_{k,\phi \leqslant th} - Y_{k,\phi \leqslant th}) \frac{d_k}{Y_k}}{\sqrt{\sum_{k=1}^{m} \frac{Y_{k,\phi \leqslant th}}{Y_k} (1 - \frac{Y_{k,\phi \leqslant th}}{Y_k}) \frac{Y_k - d_k}{Y_k - 1}) d_k}}, \tag{3}$$

where $\tau_1 < ... < \tau_m$ are distinct event times reaching node h; $d_{k,\phi \leqslant th}$ and $Y_{k,\phi \leqslant th}$ are respectively the number of events and the number of individuals at risk which have had no event before τ_k and fall into the left daughter node according to the threshold th over feature ϕ (Analogously, $d_{k,\phi > th}$ and $Y_{k,\phi > th}$ for the right daughter node); and finally, $Y_k = Y_{k,\phi \leqslant th} + Y_{k,\phi > th}$ and $d_k = d_{k,\phi \leqslant th} + d_{k,\phi > th}$.

A last detail of RSF is that, unlike class histogram or regression value in RF, each leaf in RSF stores the mortality and survival curve for the patients falling into that leaf. An example of a survival tree is presented in Fig. 3. In the case of RSF, the CHF is computed for each leaf $\hat{H}_h(t)$ with h the leaf index, and only from the patients falling within h.The mortality equation 2 is now the sum of the CHF in the leaf at each time $\hat{H}_h(t)$. The ensemble mortality is obtained averaging the mortalities over all the trees in the forest.

A conventional performance evaluation measurement for survival methods is the concordance index CI. CIrefers to the probability that, for a pair of randomly chosen samples, the sample with the higher risk prediction does experience an event (e.g. a progression) after one with a lesser risk. We consider the prediction error to be $1 - CI$.

3.3 Feature selection (stage I)

The input to our framework is a concatenated vector of all the clinical and image-based features. As opposed to common practice in the MM literature where a small number of features are studied, our objective is to automatically select the most relevant features in agreement to our prediction model. Therefore, we deploy a first RSF for feature selection using three strategies:

- The *variable importance* (VIMP) for one variable, starts by finding all the nodes where the chosen variable was selected as the optimal split direction and then assigning a daughter node randomly for each patient in the node. VIMP measures then for each variable how much the error increases after the replacement [8].
- The *minimal depth* assesses the predictiveness of a variable assuming variables selected close to the root are more important [9].
- *Variable-Hunting* [9], was defined for ultra-high dimensional problems. An RSF is fit to a random subset of data points, and a first group of variables is selected using minimal depth thresholding. Additional variables are then appended to the initial model in increasing order of minimal depth until the VIMP criterium stabilizes. The process is repeated several times, to finally retain the variables that appear most frequently over the trials.

The VIMP and minimal depth methods were run $nrun = 100$ times. The repetition is performed with the purpose to handle the variability of the features selection methods and the features were ranked according to their sum of importance/minimal depth value over the 100 runs, in a vector $\mathbf{x}_{\text{rank}} \in \mathbb{R}^D$.

3.4 Progression prediction (stage II)

In the second stage of the proposed framework, we take as input only the ranked feature vector $\mathbf{x}_{\text{rank}}$ obtained in the first stage. A second RSF is then trained for the PFS prediction task. The optimal number of features $N_{\text{feat}} \leq D$ is determined at this stage by a grid search and final results come from the RSF trained on N_{feat} features

During testing, the 2nd RSF provides a mortality prediction for each test subject and thus allows the computation of the prediction error against the recorded times of first progression. In addition, we further analyze the mortality rates with a long-rank test to separate the test subjects into two groups (bad and good prognosis) according to their predicted mortality. After the RSF, several thresholds on the ensemble mortality $\hat{M}_i$ are evaluated using the log rank test of Equation 3, where now ϕ is the mortality. Thus $\phi \leqslant th$ and $\phi > th$ represent respectively the good prognosis group and the bad prognosis groups.

4 Experimental validation

Our experiments were designed using the French trial IMAJEM aimed at evaluating the potential benefit of PET/CT in MM at diagnosis [13]. 134 patients were included in the study but only 66 patients were eligible for textural feature computation on the most intense FL. As mention before, survival time of interest is the progression-free survival (PFS). An important aspect of this dataset is that it is based on a prospective, multicentric trial with a long-term follow-up (median: 5.3 years).

In order to demonstrate the overall performance of the proposed framework, including both the feature selection and prediction phases, we compare our results against two common baseline algorithms, Gradient-Boosting Cox and Lasso-Cox in section 4.2. We further analyze in section 4.4 the contribution of the feature selection step as well as the effect of the different types of features in section 4.4, in particular, towards understanding the predictive value of image-based features for MM.

4.1 Implementation details

First, a k-fold cross validation was performed to find the combination of hyper-parameters allowing the best average prediction error over the k-folds. This combination was then used in all the following steps to train the forests (stages I and II). For the feature selection, the VIMP and MD selection methods were run 100 times.

As a result, we obtained for each method a list $\mathbf{x}_{\text{rank}}$ of features ranked according to the sum of the variable importance or minimal depth over the 100 iterations. The variable-hunting method was performed in parallel. For the performance validation, we performed a k-fold cross validation of the RSF prediction stage with the best features.

The performance was reported as the PFS prediction error ($1 - CI$), as well as the separation into two prognosis groups with the associated p-value (calculated on the difference between the two survival curves thanks to the log-rank test (see equation 3). In all our experiments 10-fold cross validation was used, except for the evaluation of the prognosis group separation were it is a 5-fold cross validation, given the small number of samples for the curves.

In order to make the framework as automatic as possible, we conduct a hyper-parameter grid-search optimization of the RSFs, prior to the training of the feature selection and prediction models RSF, over the following parameters and values:

- number of trees: {20,50,100,500,1000}. Retained: 50
- split mode: logrank and random logrank. Retained: random
- minimal number of samples in node: between 5 and 12. Retained: 7
- max features parameter in each node: 50% of the number of features, 70% of the number of features and 100% of the number of features. Retained: 100%

4.2 Overall performance in comparison with baseline methods

The purpose of this experiment was to evaluate the overall performance of the proposed framework, receiving as input the table of all features and the survival data (PFS and censorship). The results are presented in figure 4a. and table 4c., where we also report the performance of three other methods: the common Cox-regression under lasso penalization [16], an RSF trained directly on all the features, and the gradient-boosting Cox method [15], recently reported as competitive in [18]. All methods were run with a 10-fold cross validation repeated 10 times to observe the RSF variability. Figure 4a. illustrates the prediction error suggesting that our method performs better than Lasso and gradient-boosting Cox. We observe in table 4c. that the mean prediction error is very promising (0.36) and the lowest among the compared methods. Moreover, all methods were run with a 5-fold cross validation to compare the p-value of the prognosis group separation (one time). Table 5b. shows that only the prognosis curves obtained with a long-rank test on the PFS values predicted by our model are statistically meaningful according to the reported p-value (0.05).

Figure 5a. shows an example of the Kaplan-Meier curve obtained after a separation with a log-rank test, on predicted mortalities.

4.3 Comparison of feature selection methods

Here we evaluate the RSF-VIMP feature selection against other methods. Results can be found in tables 4c., 5b. and figure 4a. Methods that benefit from feature selection performed better than without. Moreover, the VIMP-RSF combination provides better results, both in terms of prediction error and p-value of the log-rank test over the predictions. Figure 4b. and table 4c. show that among RFS-based feature selection methods, minimal depth and variable-hunting tend to have a larger variability in terms of optimal number of features.

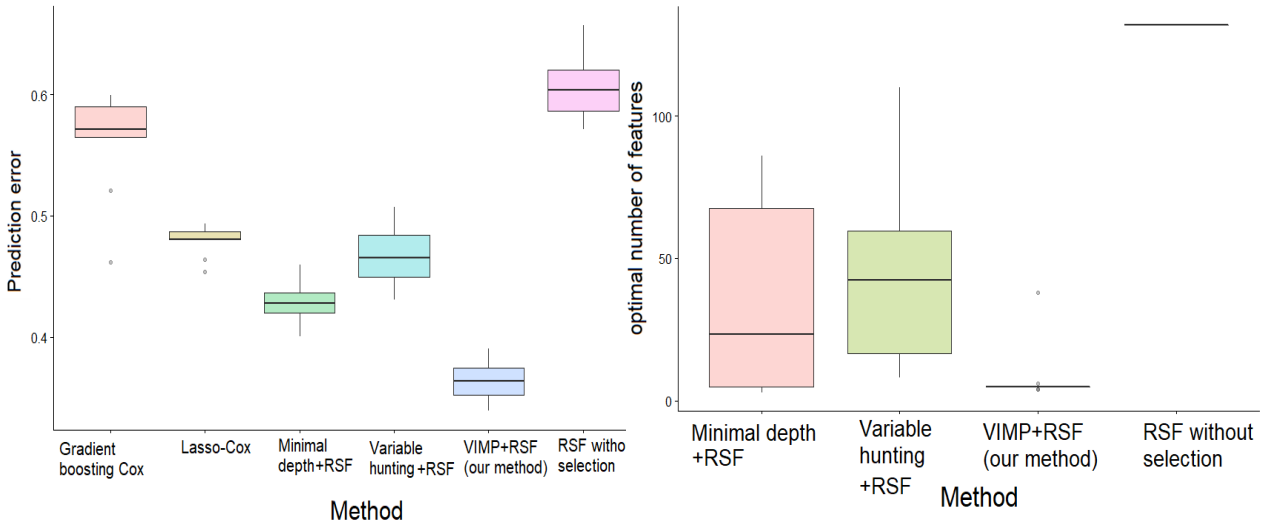
4.4 Experiments demonstrating the predictive value of images

The potential predictive added-value of image-based features was analyzed using 3 sub-databases: one with all features, one with clinical-only features, and one with imaging-only features. All sub-databases have the same number of patients (66). The prediction error was calculated to analyze the contribution of each sub-database to prognosis prediction as presented in table 2. The variables retained in each sub-analysis were also reported.

It can be seen that when considering all the available features as input, a majority of the retained ones are image-based (see Fig. 6). Moreover, as presented in table 2, the use of image features provides better results than clinical features alone for all methods with selection. In terms of error prediction, the use of both clinical and image features (including conventional and textural features) gives better results than using clinical features alone, and slightly better results than using image features alone. Finally, looking at the best ranked features, we note that among all features almost all of the thirty best (see Fig. 6) were obtained with a relative resampling instead of absolute resampling (see section 3.1).

5 Discussion

We highlight the benefit of using image-based features (including textural features) for progression prediction,



a)Average prediction error (1 - C-index) for each method over 10 folds repeated 10 times.

b) Average optimal number of features N_{feat} over 10 folds repeated 10 times.

Method	Average best number of features	Average prediction error
Our method (RSF+VIMP)	8.2 ± 10	0.36 ± 0.015
RSF Without selection	132	0.61 ± 0.027
RSF + Minimal depth	45 ± 33	0.47 ± 0.025
RSF + Variable-Hunting	8.7 ± 34	0.43 ± 0.016
Gradient-Boosting Cox		0.56 ± 0.011
Lasso-Cox		0.48 ± 0.042

c)Average error for the PFS prediction, and average best number of features over 10 runs according to the different compared methods, using all features as input.

Fig. 4: Comparison of PFS prediction methods including different feature selection strategies

Table 2 Prediction error according to the type of feature provided, and for different variable selection methods associated to the RSF.

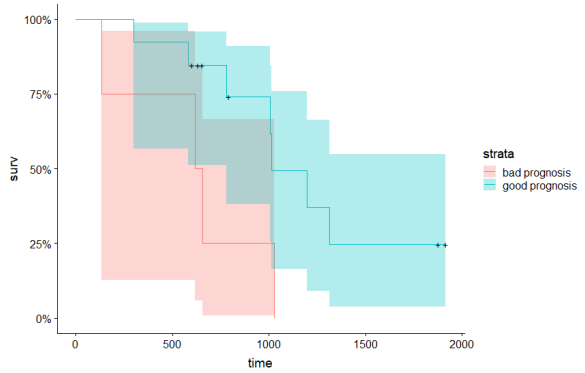
	All features	textural and conventional features	clinical features
Our method	0.34	0.36	0.45
Minimal depth method	0.45	0.45	0.48
Variable-Hunting method	0.40	0.41	0.45
Without selection	0.51	0.67	0.52

while all experiments show that the best feature-sets contain both imaging and clinical features.

The results of variable selection need to be further studied. In particular, the exact order of the best features is not always relevant. Indeed, when there are few retained variables, the importance values or the minimal depth values are close among features and the order is not always stable over different runs. For a larger number of retained features, the features ranked at the top are often the same but not always in the same order. The fact that the number of image-based features is larger than the number of clinical ones may induce a bias towards having more of the former within the top ranked ones. It would be interesting to further study the behavior under balancing sampling strategies [9].

6 Conclusion

In this work we presented an automatic framework for the progression prediction in Multiple Myeloma based on clinical and image-based features. The framework consists of a hyper-parameters optimization step, and a two RSF model for variable selection and prediction. The model provides several outcomes including a prediction of the mortality and a prognosis class for each patient. A further analysis of the feature selection process also gives hints on the most predictive features



a)An example of survival curve for the best model (Error prediction 0.39), separated according to the log-rank rule over the estimated mortality (p-value 0.04). (See table 4.c)

Method	Average p-value
Our method	0.05
Gradient-Boosting Cox	0.27
Lasso-Cox	0.4
Without selection	0.40
Minimal depth	0.24
Variable-Hunting	0.11

b) Average p-value over the 5-fold cross-validation (1 run) according to the method using all features.

Patient	1	2	3	4	5	6	7	8	9
Prognosis group	bad	bad	good	bad	good	good	good	bad	good
Mortality	38.906	39.165	6.489	39.165	31.448	23.185	31.034	39.623	33.499
patient	10	11	12	13	14	15	16	17	
Prognosis group	good	good	good	good	good	good	good	good	
Mortality	10.042	33.486	23.508	6.569	26.998	36.075	22.780	31.530	

c) An example of our method's output: each patient is assigned a prognosis group and a predicted mortality. The table corresponds to the survival curves in figure b).

Fig. 5: Evaluation of the prognosis group separation.

that could be further studied as biomarkers. The experimental results show a promising prediction error of 0.34 and a meaningful separation between prognosis classes compared to standard methods (gradient-boosting Cox, Lasso Cox) and other feature selection approaches (minimal depth, variable hunting). Our experiments further demonstrate the interest of feature selection and the importance of considering both clinical and image-based features. To conclude, this is the first method combining PET radiomics and RSF for progression prediction in multiple myeloma. The proposed framework may also serve for the automatic survival analysis in other clinical contexts, in particular when large number of clinical and imaging features are to be considered.

References

1. Amin S.B., Minvielle S., Hanlon B., Shah P.K., Li C., Li Y., Swanson D., Moreau P., Magrangeas F., Anderson K.C., Avet-Loiseau H., Munshi N.C.: Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. In: Leukemia, vol. 28, pp. 2229–2234 (2014)
2. Breiman L.: Bagging predictors. Machine Learning **42 2**, 123–140 (1996)
3. Bhnemann C., Li S., Yu H., White H.B., Schfer K., Llombart-Bosch, A., Machado, I., Picci, P., Hogendoorn, P., Athanasou, N., Noble, J., Hassa, A.: Quantification of the heterogeneity of prognostic cellular biomarkers in ewing sarcoma using automated image and random survival forest analysis. Plos one (2014)
4. Carlier, T., Bailly, C., Leforestier, R., Touzeau, C., Moreau, P., Bodere, F., Bodet-Milin, C.: Prognostic added value of PET textural features at diagnosis in multiple myeloma. Journal of Nuclear Medicine **58**(supplement 1), 111 (2017)
5. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**(2), 187–220 (1972)
6. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. In: Radiology, vol. 278, pp. 563–577 (2016)
7. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**(6), 610–621 (1973)
8. Ishwaran, H., Kogalur, U., Blackstone, E., Lauer, M.: Random surival forest. The annals of applied statistics **2**(3), 841–860 (2018)
9. Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S.: High-dimensional variable selection for sur-

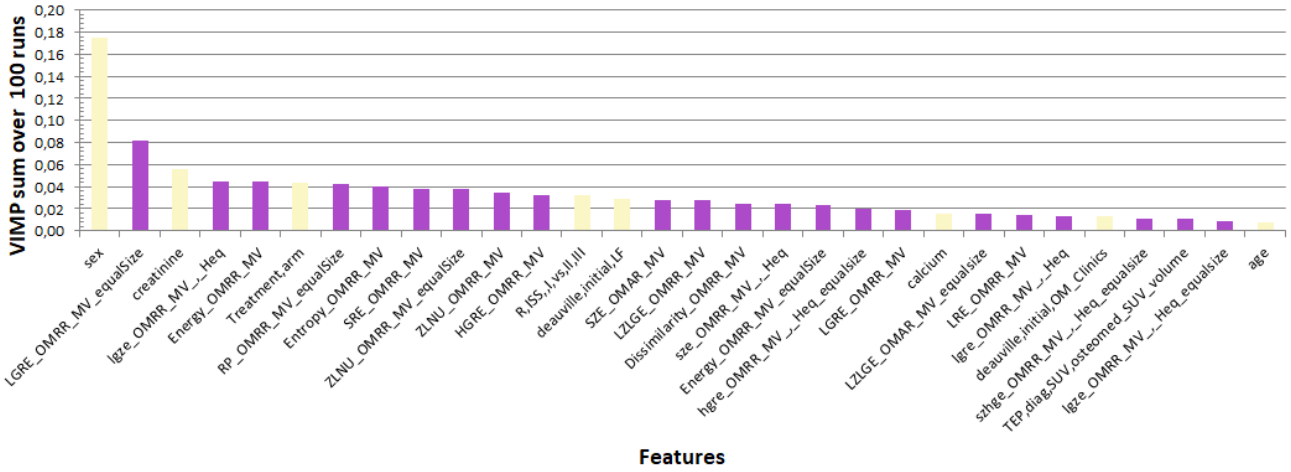


Fig. 6: Histogram of the 30 best features according to our VIMP-RSF method. Yellow: clinical features, purple: image-based features.The different implementations are denoted as OMRR (One Matrix relative resampling), OMAR (One Matrix absolute resampling), Heq (histogram equalization), and equalsize (equal size of voxels.)

vival data. Journal of the American Statistical Association **105**(489), 205–217 (2010)
10. Kaplan, E.L., Meier, P.: Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association **53**(282), 457–481 (1958)
11. Lartizien, C., Rogez, M., Niaf, E., Ricard, F.: Computer-aided staging of lymphoma patients with FDG PET/CT imaging based on textural information. IEEE Journal of Biomedical and Health Informatics **18**(3), 946–955 (2014)
12. Larue, R.T.H.M., Defraene, G., Ruysscher, D.K.M.D., Lambin, P., van Elmpt, W.J.C.: Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. The British journal of radiology (2017)
13. P.Moreau, F.Caillon, C.bodet-Milin: Prospective evaluation of magnetic resonance imaging and [18F]Fluorodeoxyglucose positron emission tomography-computed tomography at diagnosis and before maintenance therapy in symptomatic patients with multiple myeloma included in the IFM/DFCI 2009 trial: Results of the IMAJEM study. Journal of Clinical Oncology **35**(25), 2911–2918 (2017)
14. Ridgeway, G.: The state of boosting. Computing Science and Statistics (1999). DOI citeulike-article-id:7678637
15. Ridgeway, G.: Generalized Boosted Models : A guide to the gbm package (2007)
16. Tibshirani, R.: The lasso method for variable selection in the cox model. Statistics in Medicine (1997)
17. Vallires, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J.W.L., Khaouam, N., Nguyen-Tan, P.F., Wang, C.S., Sultanem, K., Seuntjens, J., Naqa, I.E.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Reports **7**(10117), 2911–2918 (2017)
18. Wenzheng, S., Jiang, M., Dang, J., Chang, P., Yin, F.F.: Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiation Oncology **13**(1), 197 (2018)
19. Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P., Piraud, M., Buck, A., Shi, K., Menze, B.H.: Automated whole-body bone lesion detection for multiple myeloma on 68Ga-Pentixafor PET/CT imaging using deep learning methods. Contrast Media & Molecular Imaging **2018**(2391925) (2018)
20. Zamagni, E., Patriarca, F., Nanni, C., Zannetti, B., Englaro, E., Pezzi, A., Tacchetti, P., Buttignol, S., Perrone, G., Brioli, A., Pantani, L., Terragna, C., Carobolante, F., Baccarani, M., Fanin, R., Fanti, S., Cavo, M.: Prognostic relevance of 18-F FDG PET/CT in newly diagnosed multiple myeloma patients treated with up-front autologous transplantation. Blood **118 23**, 5989–95 (2011)
21. Zhou, Y., Mcardle, J.J.: Rationale and applications of survival tree and survival ensemble methods. Psychometrika **80 3**, 811–33 (2015)
22. Zwanenburg, A., Leger, S., Vallières, M., Lck, S.: Image biomarker standardisation initiative. arXiv1612.07003 (2016)
23. McDonald, J.E., Kessler, M.M., Gardner, M.W.; Buros, A.F., Ntambi, J.A., Waheed, S., van Rhee, F., Zangari, M., Heuck, C.J., Petty, N., Schinke, C., Thanendrarajan, S., Mitchell, A., Hoering, A., Barlogie, B., Morgan, G.J. Davies, F.E.: Assessment of Total Lesion Glycolysis by ^{18}F FDG PET/CT significantly improves prognostic value of GEP and ISS in myeloma Clin Cancer Res **23 8**, 1981–7 (2017)

Funding: This work has been supported in part by the European Regional Development Fund, the Pays de la Loire region on the Connect Talent scheme MILCOM (Multi-modal Imaging and Learning for Computational-based Medicine), Nantes Métropole (Convention 2017-10470), the French National Agency for Research called "Investissements d'Avenir" IRON Labex n° ANR-11-LABX-0018-01 and INCa-DGOS-Inserm_12558 (SIRIC ILIAD)

Conflict of Interest: The authors declare that they have no conflict of interest. **Ethical approval:** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki dec-

laration and its later amendments or comparable ethical standards.This article does not contain any studies with animals performed by any of the authors. **Informed consent:** Informed consent was obtained from all individual participants included in the study.