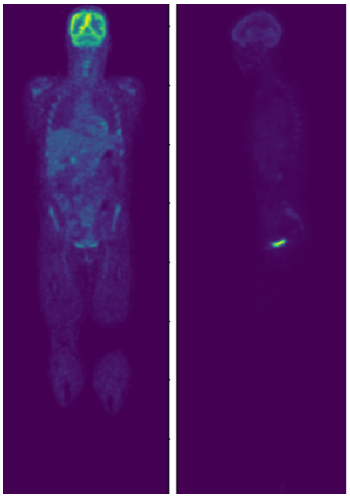
# Prédiction de survie par Random Survival Forest et imagerie TEP dans le contexte du myélome multiple



RAPPORT DE FIN D'ETUDE - 2018


Ludivine Morvan

Encadrants :

* Diana Mateus, Professeur des Universités, LS2N, Ecole Centrale Nantes

Dr. Thomas Carlier, physicien médical, CHU de Nantes

Hugues Talbot, Professeur, Centrale Supelec

# Remerciements

Tout d'abord, je tiens à remercier vivement Madame Diana Mateus, enseignant-chercheur dans l'équipe SIMS du Ls2n à l'Ecole Centrale de Nantes, ainsi que Monsieur Thomas Carlier physicien médical au CHU de Nantes, pour le sujet de ce stage, pour leur encadrement, leur expertise et le temps qu'ils m'ont consacré, mais aussi pour leur aide dans la rédaction du rapport. Je les remercie également de m'accorder leur confiance et de m'accueillir prochainement en thèse.

J'adresse également mes remerciements à Monsieur Hugues Talbot, nouvellement professeur à Centrale Supelec, qui m'a beaucoup aidé dans ma recherche et m'a permis de trouver ce stage. Je le remercie également de ses précieux conseils qui m'ont permis de découvrir le domaine qui me correspond.

Enfin, je remercie également toute l'équipe SIMS pour leur accueil et les étudiants de la salle de master pour l'entraide et la convivialité apportées lors de ce stage.

# Abréviations

TEP : Tomographie à Emission de Positons (anglais : PET)
CT-scan : computed tomography scan
PFS : Progression-Free Survival
RSF : Random Survival Forest (forêt de survie aléatoire)
VIMP : Variable IMPortance
FDG : fluorodésoxyglucose
IRM : Imagerie par Résonance Magnétique
SUV : Standardized Uptake Value
K-NN : K Nearest Neighbors
SVM : Support Vector Machine
RF : Random Forest (forêt aléatoire)
AUC : Area Under Curve
DICOM : Digital Imaging and Communications in Medicine
ROI : Region Of Interest
GLCM : Gray Level Co-occurrence Matrix
GLSZM : Gray Level Size Zone Matrix
GLRLM : Gray Level Run Length Matrix
0MAR : One Matrix Absolute Resampling
0MRR : One Matrix Relative Resampling
OOB : Out-Of-Bag
OS : Overall Survival

# Abstract

*The internship takes part of a 5-year-project called MILCOM (Multi-modal Imaging and Learning for Computational-based Medicine). Its purpose is to help physicians to diagnose, forecast and prescribe a personalized treatment for patients with multiple myeloma, thanks to more precise information about the disease, and the prospect of further development. The internship and the thesis that follows, belong to the first part of the project whose main goal is the research of biomarkers by developing machine learning algorithms that link quantitatively and reproducibly survival and PET (Positron Emission Tomography) and CT (Computed Tomography) images in the myeloma context, while minimizing changes that will need to be provided in the different stages of clinical works.*

*To fulfill this goal, even if it still needs to be improved, we developed a semi-automatic pipeline using python, which enables, from TEP images and clinical characteristics, to segment lesions, extract features and predict the PFS (Progression-Free Survival) for patients with myeloma, and classify them into two groups (good and bad prognosis). The pipeline is meant to help physicians to automate these tasks, save them time, and reduce possible errors. The main task was the features extraction and the survival part. The Random Survival Forest method (RSF) was used for the survival prediction. It is an ensemble trees method presented by Ishwaran [1] in 2008, which allows to take censoring and missing data into account. Despite having some errors in the segmentation and features extraction, the main algorithms are written.*

*About survival prediction, RSF algorithm was re-written using python instead of the algorithm available in R, to continue the python pipeline and to permit possible modifications and improvements in the future. Using the data base with all features, the prediction error was 0.41 (in R), but with a prior features selection, thanks to the variable importance (VIMP), we obtain error of 0.29. Features found like predictive were clinic, volumic and textural. Also, it permits to have two groups of patients (ones with good prognosis and others with a bad one), showed using the Kaplan-Meier method. The p-value was 0.005 and the logRank test value was 7.2, so we can consider that it is a good split.*

*To conclude, this internship allowed me to prepare the phd thesis and begin all branches of the project. It permitted finding new trails, as using a feature selection method before the survival prediction, the missing data algorithm to have more patients, or even deep learning.*

*We can say that the techniques used in this project, such as RSF, to study survival prediction of patients with myeloma, using image processing, are new to this domain, and the results are promising..*

# Table des matières


**1 Présentation du laboratoire** **7**

**2 Contexte** **9**

**3 Connaissances a-priori des méthodes pour l'analyse de la survie** **13**

3.1 Les méthodes d'analyse statistique . . . 15

3.1.1 Les méthodes non paramétriques . . . 15

3.1.1.1 Kaplan-Meier . . . 15

3.1.1.2 Estimateur de Nelson-Aelen . . . 16

3.1.1.3 Comparaison Kaplan-Meier et Nelson-Aalen . . . 17

3.1.2 Méthode semi-paramétrique : Cox . . . 17

3.2 Les méthodes d'apprentissage automatique . . . 18

3.2.1 Les SVM . . . 18

3.2.2 Les réseaux de neurones . . . 19

3.2.3 Les Random Forest (RF) . . . 19

**4 État de l'art** **20**

4.1 L'analyse de la survie . . . 20

4.2 L'utilisation d'images médicales pour l'étude de la survie . . . 22

4.3 Myélome multiple et survie . . . 23

4.4 L'utilisation de l'imagerie TEP pour la détection de lésion et la prédiction de survie en général 23

4.5 Conclusion . . . 24

**5 Méthodes** **25**

5.1 Détection et segmentation des lésions . . . 26

5.1.1 Les segmentations manuelles . . . 26

5.1.2 Le W-net et méthode des patchs . . . 26

5.1.3 Contribution . . . 28

5.2 Les caractéristiques radiomiques . . . 28

5.2.1 Les différentes caractéristiques texturales . . . 28

5.2.1.1 Caractéristiques de premier ordre . . . 29

5.2.1.2 Les caractéristiques de second ordre . . . 29

5.2.2 Contribution . . . 31

5.3 L'étude de la survie par random survival forest . . . 31

5.3.1 Les méthodes d'arbres . . . 31
5.3.2 La méthode de RSF . . . 34
5.3.3 Contribution . . . 37

**6 Détails d'implémentation** **39**
6.1 Les données . . . 39
6.2 Les logiciels utilisés . . . 40
6.2.1 Le langage python . . . 40
6.2.2 Le langage R . . . 41
6.3 Validation expérimentale . . . 41
6.3.1 La segmentation . . . 41
6.3.2 Les caractéristiques texturales . . . 41
6.3.3 L'analyse de la survie . . . 42
6.3.3.1 L'implémentation en python . . . 43
6.3.3.2 Validation de l'implémentation par comparaison avec R . . . 44
6.3.3.3 Optimisation du temps de calcul en python . . . 44
6.3.3.4 Valeur prédictive de notre base de données et détermination de biomarqueurs 44
6.3.3.5 Comparaison de la méthode RSF et autres méthodes de RF . . . 46

**7 Résultats** **47**
7.1 Segmentation . . . 47
7.2 Radiomics . . . 48
7.2.1 Validation par utilisation de l'ISBI . . . 48
7.2.2 Comparaison des valeurs calculées par pyradiomics avec la vérité terrain . . . 49
7.3 Validation de l'implémentation RSF en python . . . 49
7.4 Résultats sur la base du myélome multiple . . . 50
7.4.1 optimisation des hyperparamètres . . . 50
7.4.2 Valeur prédictive de notre base . . . 53
7.4.3 Comparaison avec les Random Forest . . . 54
7.4.4 interprétation des biomarqueurs . . . 56

**8 Discussion** **64**

**9 Conclusion du stage et Perspectives** **66**

**A Annexes** **68**
A.1 Annexe 1 : Les caractéristiques cliniques . . . 68
A.2 Annexe 2 : Les caractéristiques volumiques . . . 68
A.3 Annexe 3 : Les caractéristiques Texturales . . . 69

**Table des figures** **70**

**Liste des tableaux** **72**

**Bibliographie** **73**

*Chapitre 1*

# Présentation du laboratoire

Le stage s'est déroulé dans le laboratoire Ls2n, dans l'équipe SIMS. Le Laboratoire des Sciences du Numérique de Nantes (LS2N) est une nouvelle Unité Mixte de Recherche créée en janvier 2017 qui résulte de la fusion des UMR IRCCyN (Institut de Recherche en Communications et Cybernétique de Nantes), et LINA (Laboratoire d'Informatique de Nantes Atlantique). Elle est soutenue par 5 établissements publics d'Enseignement Supérieur et de Recherche et comprenant l'activité d'environ 450 personnes (Ecole Centrale de Nantes, Université de Nantes, CNRS, IMT Atlantique et INRIA). Il est organisé en 5 Pôles :

- SIEL (signaux, images, ergonomie et langues)
- CCS (conception et conduite systèmes)
- RPC (robotique, procédés, calculs)
- SDD (science des données et de la décision)
- SLS (science du logiciel et des systèmes distribués)

Et se positionne en 5 thèmes transversaux :

- EDF (Entreprise Du Futur)
- GEMIE (Gestion de l'Energie et Maîtrise des Impacts Environnementaux)
- SDV (Science du Vivant)
- VEM (Véhicule et Mobilité)
- CCSN (Création, Culture et Société Numérique)

Le stage s'est déroulé dans l'équipe SIMS (Signal, IMage, Son), dirigée par Saïd Moussaoui. C'est une équipe du pôle SIEL et est dirigée vers la santé. Elle a pour objectif de concevoir des outils méthodologiques et appliqués exploitant toute structure intrinsèque aux données à l'aide de techniques de traitement du signal et de l'image. Les thèmes traités par l'équipe sont :

- Modèles et méthodes pour la résolution de problèmes inverses
- Apprentissage machine pour la décision assistée par ordinateur
- Outils mathématiques et numériques pour le calcul en grande dimension
- Conception d'outils pour des applications dans des contextes multidisciplinaires

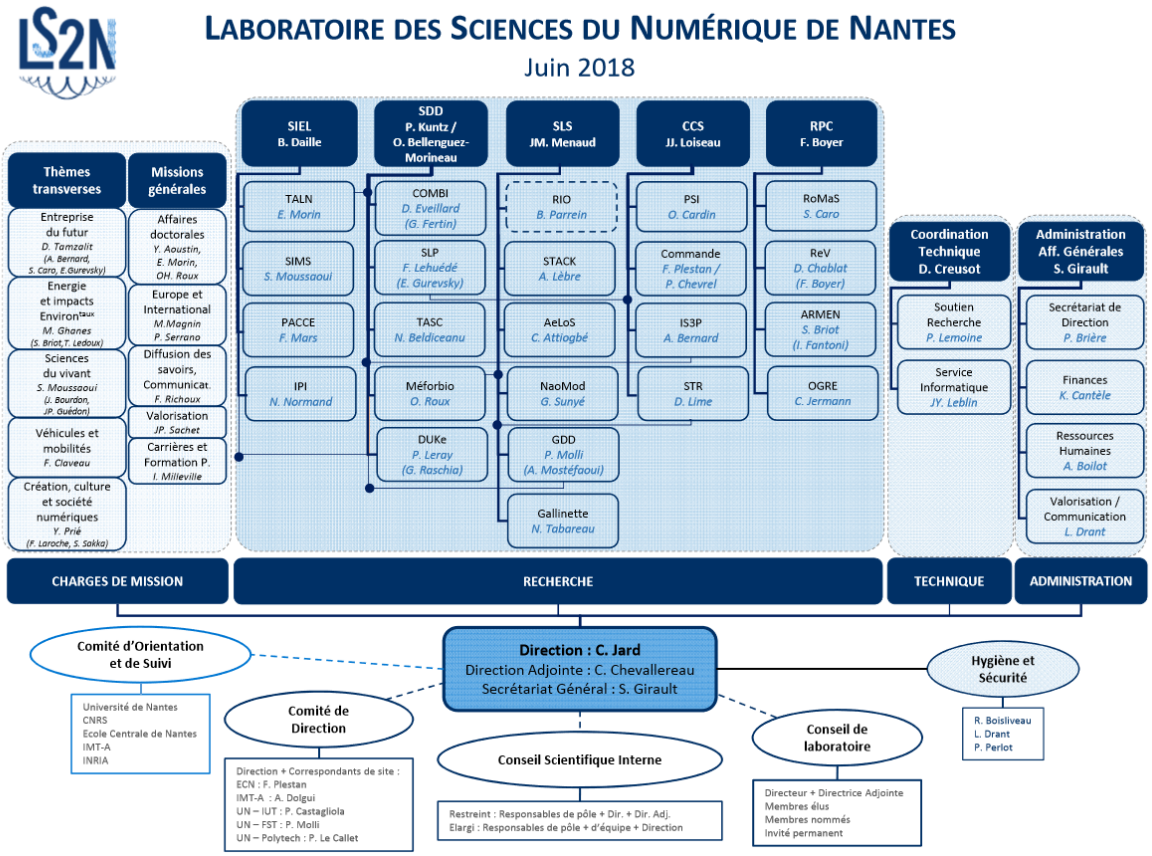



FIGURE 1.1 – Organigramme du laboratoire Ls2n

Le stage fut co-encadré par Diana Mateus, de l'équipe SIMS et Thomas Carlier, physicien médical du CHU de Nantes.

Le service de médecine nucléaire du CHU de Nantes se situe sur le site Hôtel-Dieu. Il dispose de trois gamma-caméras dont deux sont couplées avec un scanner (Spect-CT). Le service est co-utilisateur de deux TEP-CT (Tomographie à Emission de Positons - computed tomography) installées dans les locaux de l'Institut de cancérologie de l'Ouest (ICO). 2 500 examens TEP-CT sont réalisés par an. Il développe une activité de radiothérapie interne vectorisée et dispose à cet effet de deux chambres radio-protégées permettant l'isolement des personnes traitées. Le service est notamment expert en :

- Radiothérapie interne vectorisée (RIV)
- Pédiatrie (ostéoarticulaire et oncopédiatrie)
- Immuno-ciblage en oncologie : thérapeutique (radioimmunothérapie), diagnostique (immuno-TEP)
- Diagnostic et thérapie des tumeurs neuroendocrines : TEP au 68Ga, peptido-radiothérapie
- Évaluation thérapeutique : hémopathies et tumeurs solides, thérapies innovantes
- Nouveaux traceurs TEP en oncologie
- Recherche clinique en oncologie nucléaire
- Imagerie quantitative

*Chapitre 2*

# Contexte

Le stage rentre dans un projet sur 5 ans, le projet MILCOM (Multi-modal Imaging and Learning for Computational-based Medicine). Il s'inscrit dans un contexte d'écriture de la médecine du futur, prenant en considération les 4 P (prédictive, personnalisé, préventive et participative) et les défis liés aux images médicales. Le but de ce projet est d'aider les médecins à poser un diagnostic et un pronostic, et à prescrire un traitement personnalisé, grâce à des informations plus précises sur la maladie, le myélome multiple en premier lieu, et ses perspectives d'évolution. Le projet comporte 4 sous-projets répondant à quatre problématiques

1. Adoption par la communauté médicale : recherche de biomarqueurs liants survie et images, en minimisant le changement qui devrait être apporté dans les étapes d'un travail clinique.
2. Acquisition : amélioration de la qualité des images et étudie les effets des méthodes de reconstruction des images TEP
3. Annotation : adaptation des algorithmes aux défis médicaux
4. Données : représentation des images à nature multiple

Le stage et la thèse qui va suivre, s'inscrivent dans le premier sous-projet (Adoption par la communauté médicale) et ont pour objectif de développer des algorithmes d'apprentissage automatique pour lier de façon quantitative et reproductible les images TEP et CT (Computed Tomography), et la survie de patients atteints de myélome multiple.

Pour ce faire, l'apprentissage automatique et les images TEP (et CT) sont utilisés pour la détection des lésions, l'extraction de ses caractéristiques, et la prédiction de la survie des patients atteints de myélome multiple. Le stage porte principalement sur l'extraction des caractéristiques et la prédiction de la survie.

Le myélome multiple ou maladie de Kahler, est une maladie de la moelle osseuse caractérisée par la multiplication dans la moelle osseuse d'un plasmocyte anormal. Les plasmocytes anormaux envahissent la moelle osseuse, comme présenté dans la figure 2.1, avec différentes conséquences :

— Le système immunitaire est affaibli car le nombre de plasmocytes normaux diminue.
— La production des cellules sanguines au sein de la moelle osseuse peut être diminuée.
— Les plasmocytes anormaux activent des cellules qui détruisent l'os et perturbent ses mécanismes de construction. Cela peut aussi engendrer une hypercalcémie, ce qui peut provoquer des troubles cardiaques et cérébraux, faiblesse musculaire etc...
— L'immunoglobuline monoclonale produite par les plasmocytes anormaux circule dans le sang et lors de son passage dans les reins, peut y former des dépôts qui altèrent leur bon fonctionnement. L'insuffisance rénale est ainsi une complication fréquente du myélome multiple [2].

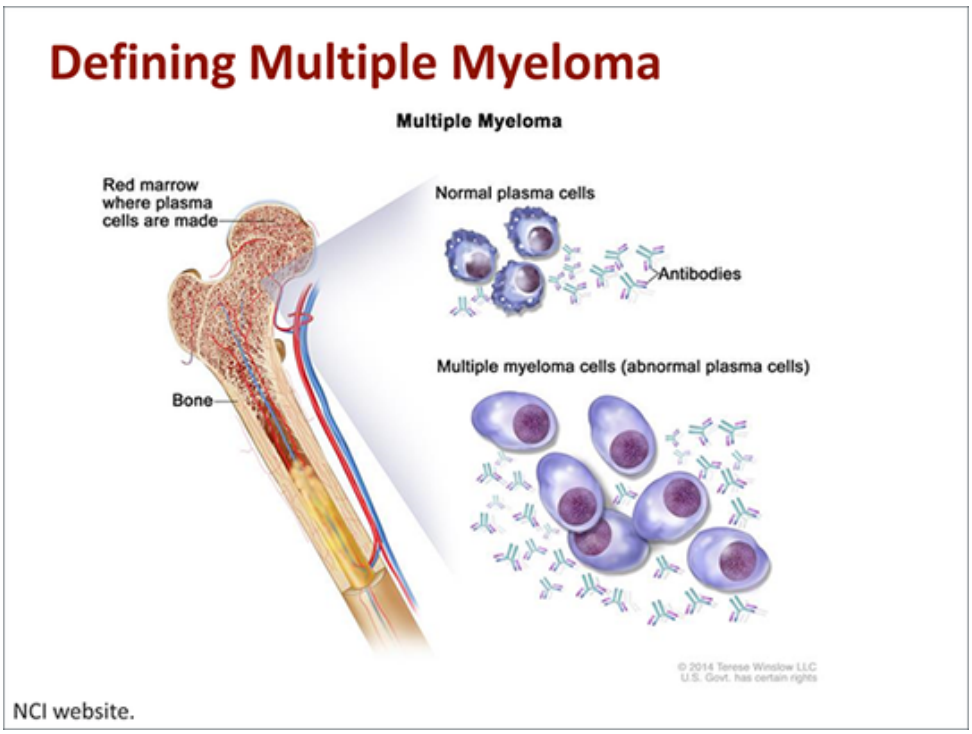



FIGURE 2.1 – Le myélome multiple est un cancer de la moelle osseuse se caractérisant par la prolifération de plasmocytes mutés

En France, environ 6000 à 7000 nouveaux cas de myélome multiple sont diagnostiqués chaque année en France et les risques de rechute sont très fréquents. La survie nette à 5 ans est en moyenne de 54% [3] [4].

Les images 3D obtenues avec la 18F-FDG (fluorodésoxyglucose) TEP permettent un diagnostic précoce des lésions osseuses focales du myélome multiple avec une sensibilité de 85 à 93% et une spécificité de 83 à 100%. Sa sensibilité est supérieure à celle des radiographies conventionnelles et peu différente de celle de l'IRM (Imagerie par Résonance Magnétique). Elle met en évidence 25 à 55% de nouvelles lésions par rapport aux autres techniques d'imagerie [5]. La FDG-TEP à l'avantage d'être quantitative (valeurs en SUV (Standardized Uptake Value)). Une absence de fixation (SUV faible), confirme une réémission, alors que la persistance d'une activité résiduelle après traitement est un mauvais pronostic et peut conduire au changement de traitement.

La méthode du 18F-FDG TEP utilise l'association d'un vecteur, le glucose et d'un émetteur, le fluor 18. Le glucose est consommé abondamment par les lésions cancéreuses mais aussi le cerveau et se retrouve dans la vessie par élimination. Le couple vecteur/émetteur (ou médicament radiopharmaceutique) se dirigera donc principalement vers ces zones. Le fluor 18 va se désintégrer dans 97% des cas en oxygène 18 par désintégration $\beta+$ en formant des positons (Et dans 3% des cas par capture électronique).

$$^{18}_{9}F \rightarrow ^{18}_{8}O + ^{0}_{1}e + neutrino$$

Une fois l'énergie des positons perdue (leur parcours est de 0.6 mm en moyenne) les positons vont s'annihiler avec des électrons ce qui produira des photons $\gamma$ émis dos à dos. Ce sont ces photons $\gamma$ qui seront détectés par des détecteurs opposés et donneront donc l'orientation. Puis en passant par des algorithmes de reconstruction on obtient la position de fixation du 18F-FDG. Les deux photons sont validés si leur énergie est proche de 511 keV et s'ils sont détectés dans un intervalle de temps très court, une fenêtre temporelle d'environ 4.1 ns. L'acquisition de données à des angles différents permet la reconstruction d'un plan tomographique, et à partir de ces coupes d'obtenir dans l'espace à trois dimensions la distribution du produit radiopharmaceutique dans le corps du patient.

FIGURE 2.2 – TEP-SCAN au CHU Nantes TEP/TDM (Tomographie par Emission de Positons/TomoDensitoMètre), Biogaph mCT (SIEMENS)

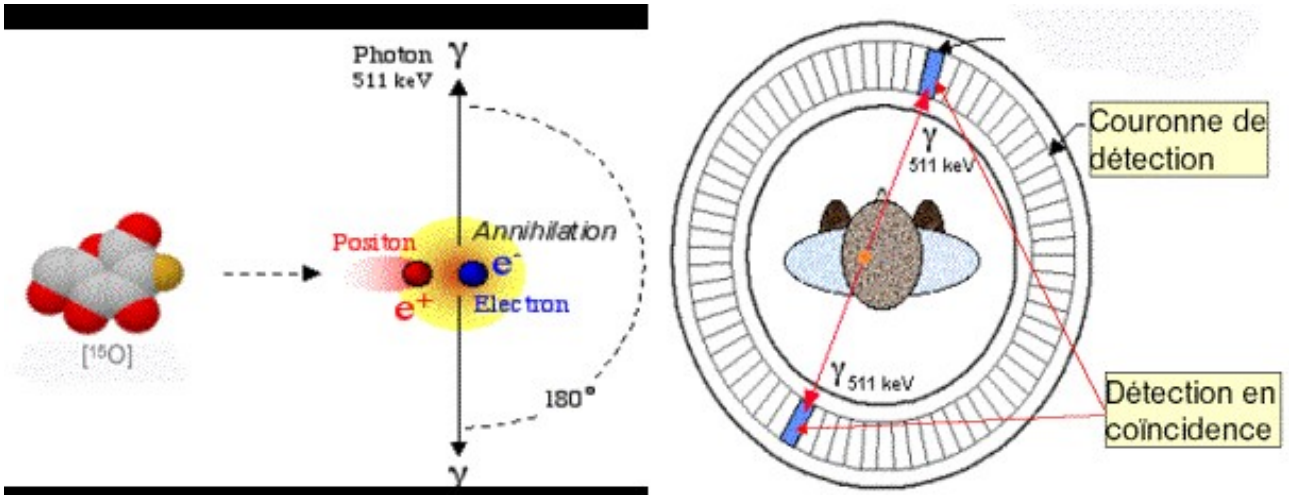



FIGURE 2.3 – Fonctionnement de la TEP. Les positons vont s'annihiler avec des électrons ce qui produira des photons $\gamma$ émis dos à dos. Ce sont ces photons $\gamma$ qui seront détectés et donneront l'orientation.

Le fluor 18 a pour avantage d'avoir une période radioactive raisonnablement longue (15 à 110 minutes). De plus, ses positons ont un parcours maximal relativement court (2,6 mm dans l'eau contre 4,1mm pour le C-11 et 8,2 pour l'oxygène-15) ce qui permet d'obtenir une carte de fixation radioactive plus proche de la réalité. L'association de la TEP à un scanner (comme c'est le cas au CHU de Nantes, présenté sur la figure 2.2), permet une identification plus aisée des lésions par les médecins car cela permet aussi de savoir si les zones à activité se trouvent sur une région osseuse ou non.

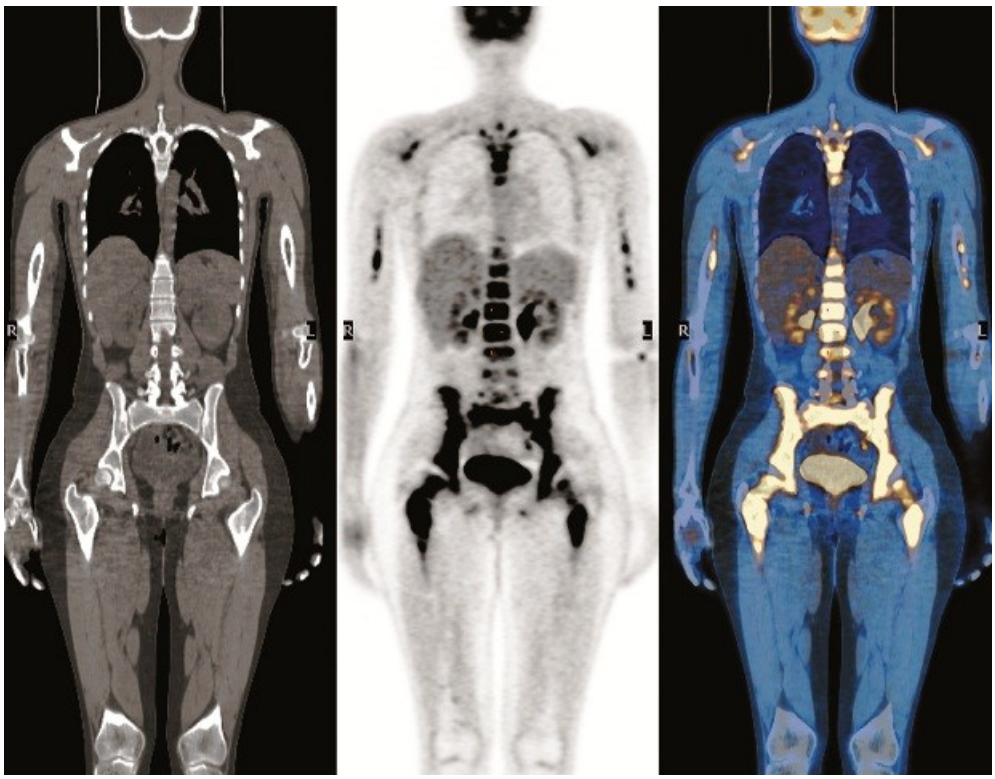


FIGURE 2.4 – Exemples d'images CT (gauche), TEP (centre), et TEP-CT fusionnée (droite)

Ces images TEP peuvent être utilisées pour le diagnostic et le suivi de la maladie mais aussi dans le cadre de la recherche.

Ce travail se place dans le contexte d'une étude clinique prospective, randomisée et multicentrique sur un traitement contre le myélome multiple [6]. Il a ainsi pour intérêt de dire quel traitement donne les meilleurs perspectives (rechute et décès) et quels sont les facteurs influençant la survie et la rechute. Cette étude clinique multicentrique utilise la base de données IMAJEM de 134 patients avec des images TEP. Nous utilisons pour notre projet, les images baseline (avant traitement). En plus des images TEP et CT sont disponibles les données cliniques des patients (âge, hémoglobine, traitement, nombre de lésions focales, etc.) et leur temps de survie et sans rechute depuis la détection de la maladie. Les données de temps sont disponibles sur 7 ans.

Ainsi, les objectifs du stage sont l'extraction de caractéristiques texturales et le développement d'algorithmes capables de prédire la survie à partir d'images FDG-TEP, et grâce à une méthode de Random Rurvival Rorest (RSF), une méthode de prédiction de survie à partir d'un ensemble d'arbres de décision, donnée par Ishwaran [1] et qui prend en compte la censure à droite et les données manquantes. Pour réaliser ces forêts d'arbres de décision, on donne en entrée des caractéristiques calculées sur des images TEP et des caractéristiques cliniques, et il en ressort une prédiction de la survie du patient mais aussi l'importance des caractéristiques dans le calcul de cette prédiction. Les techniques mises en œuvre sont relativement récentes et originales dans le contexte du myélome multiple.

*Chapitre 3*

# Connaissances a-priori des méthodes pour l'analyse de la survie

L'objectif du stage est de calculer la survie des patients atteints de myélome multiple. La méthode RSF (Random Survival Forest) est relativement récente et d'autres techniques étaient utilisées avant cela. Les résultats de la RSF seront d'ailleurs comparés avec les méthodes de Kaplan Meier et les classifieurs forêts aléatoires. Ainsi, seront présentés dans ce chapitre la définition de la survie et les méthodes préexistantes d'analyse de la survie comme les méthodes statistiques (Cox, Kaplan-Meier etc.) et les méthodes d'apprentissage automatique.

La survie correspond au temps écoulé jusqu'à la survenue d'un événement précis. Cela peut être le décès mais aussi la rechute, comme dans notre cas, ou la guérison par exemple. La fonction de survie S(t) est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire :

$$S(t) = P(X > t), \tag{3.1}$$

avec t<0 et X la durée de survie.
On peut aussi préciser la fonction de densité qui est définie par :

$$f(t) = \frac{dS(t)}{dt} \tag{3.2}$$

Cette fonction de survie peut être représentée par des courbes fonctions du temps.

On peut voir dans la figure 3.1 un exemple de courbes de survie (probabilité de ne pas avoir d'évènement au temps t en fonction de t).
De plus, en survie la notion de censure peut être présente, et plus particulièrement la censure à droite. La censure à droite est le fait de ne pas observer l'évènement d'intérêt chez un individu. Cela peut être dû au fait que le patient à été perdu de vue ou que l'étude s'est terminée avant que l'évènement ne se produise par exemple. La figure 3.2 nous présente les différents types de censure :

On appelle $\sigma_i$ la présence d'un évènement ou non. $\sigma_i$= 1 si l'évènement a eu lieu et indique qu'il n'y a pas de censure. Si $\sigma_i$ = 0, l'évènement n'a pas eu lieu et il y a censure.

L'analyse de la survie peut être utile par exemple pour prédire l'efficacité d'un traitement ou pour déterminer la gravité d'une maladie en réalisant des sous-groupes de patients. Edward B. Garon et al.[9] utilise par exemple la méthode de Kaplan-Meier [voir section 3.1.1.1.] pour évaluer l'efficacité du traitement au Pembrolizumab chez des patients atteints d'un cancer des poumons (Non–Small-Cell Lung Cancer).

Un concept important est le risque instantané. Il est définit comme la probabilité d'un évènement dans l'intervalle de temps $[t + \Delta t[$ en sachant que l'évènement ne s'était pas réalisé avant t. Ceci permet de définir la

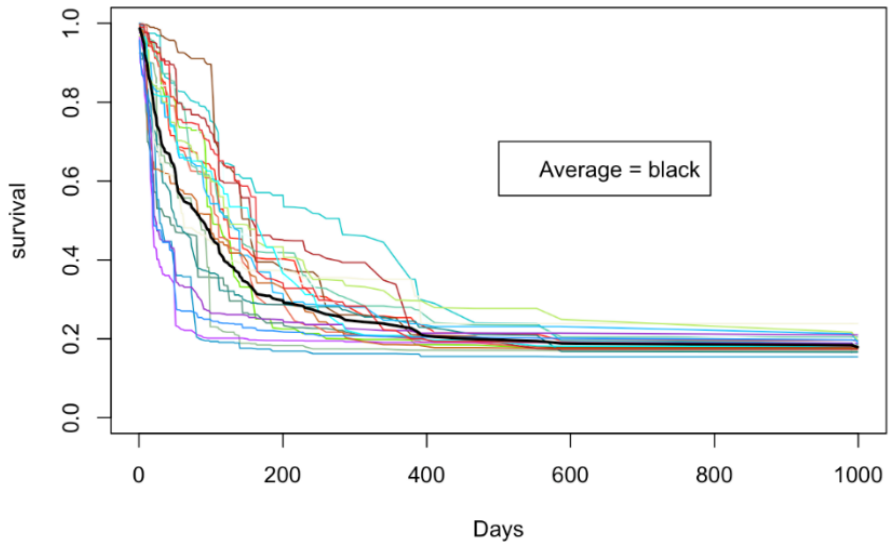



FIGURE 3.1 – Exemple de courbes de survie. Graphique représentant la probabilité de ne pas avoir eu d'évènements jusqu'au temps t, en fonction de t en jours. La courbe noir représente la courbe moyenne.[7]

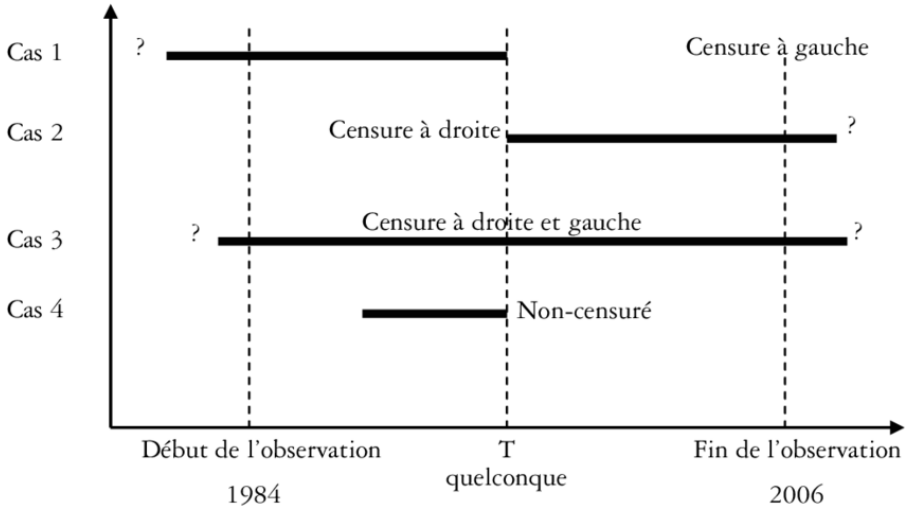



FIGURE 3.2 – Représentation des types de censure. Le cas 1 représente la censure à gauche (exemple : ne pas connaître la date d'apparition de la maladie). Le cas 2 représente la censure à droite (exemple : perdre de vue un patient).Cas 3 : censure droite et gauche. Cas 4 : pas de censure (exemple : La date d'apparition de la maladie est connue et un évènement à été enregistré).[8]

fonction de risque (ou fonction de hasard) ($\lambda$) qui apparaît comme une mesure du risque instantané :

$$h(t) = \lim_{\Delta t \to 0} \frac{Prob[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{3.3}$$

Il est aussi possible de définir le risque cumulé H(t) défini par :

$$H(t) = \int_0^t h(s)ds \tag{3.4}$$

Or, d'après le théorème des probabilités conditionnelles,

$$h(t) = \frac{f(t)}{S(t)} \tag{3.5}$$

On obtient donc à partir des équations 3.4 à 3.5 :

$$H(t) = log[S(t)] \tag{3.6}$$

Ainsi, la connaissance de l'un de ces fonctions permet d'estimer les autres. Il existe différentes méthodes pour estimer ces fonctions et en particulier la fonction de survie : Des non paramétriques comme le Kaplan-Meier, utilisée lorsque qu'aucune hypothèse ne peut être faite sur la distribution des temps de survie. L'estimateur de Nelson Aalen estime lui la fonction de risque cumulé. Il existe aussi l'approche paramétrique comme les modèles à risque proportionnel (PH) qui demande une hypothèse sur la distribution des temps de survie mais qui ne seront pas présentés dans ce chapitre [plus de détails disponibles sur le cours de G. Colletaz [10]], et les modèles semi-paramétriques comme le modèle de Cox.

Outre ces modèles d'analyse statistiques, il existe aussi des méthodes d'apprentissage automatique qui sont applicables à l'étude de la survie.

## 3.1 Les méthodes d'analyse statistique

Pour cette analyse de la survie il est possible d'utiliser l'analyse statistique et l'apprentissage automatique. L'analyse statistique est la méthode la plus courante et la plus utilisée par les médecins. Elle peut être non paramétrique, paramétrique ou semi-paramétrique. Les méthodes paramétrique et semi-paramétrique induisent une prise en compte des variables explicatives pour créer le modèle. Voici quelques exemples de méthodes les plus communes.

### 3.1.1 Les méthodes non paramétriques

#### 3.1.1.1 Kaplan-Meier

L'estimateur de la fonction de survie le plus utilisé et le plus simple lorsqu'aucune hypothèse ne veut être faite sur la distribution des temps de survie est l'estimateur de Kaplan-Meier [10].

Elle permet la description de la survie d'une population et d'estimer la survie médiane et le taux de survie à un temps donné mais aussi de comparer la survie de différentes populations, souvent par le test du logrank [voir le calcul du logRank dans la section 5.3.2].

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t (On considérera l'évènement comme étant le décès dans cette partie, bien qu'il puisse s'agir par exemple de rechute ou gérison). Soit t" < t' < t,

$$P(X > t) = P(X > t|X > t') * P(X > t'|X > t'') * P(X > t'') \tag{3.7}$$

L'équation 3.7 présente la probabilité de survie au temps t en faisant le produit de :

- la probabilité de survie au temps t sachant que le décès n'a pas eu lieu au temps t',
- la probabilité de survie au temps t' sachant que le décès n'a pas eu lieu au temps t",
- et la probabilité de survie au temps t"

En considérant les temps d'événements (décès et censure) distincts $T_i$ (i = 1,...,n) rangés par ordre croissant, on obtient la probabilité de survie au temps $T_i$ en faisant le produit de la probabilité de survie à chaque temps $T_{k \leq i}$ sachant que le décès n'a pas eu lieu au temps $T_{k-1}$ :

$$P(X > T_i) = \prod_{k=1}^{i} P(X > T_k|X > T_{k-1}), \tag{3.8}$$

avec $T_0 = 0$

Considérons les notations suivantes :

- $Y_i$ le nombre d'individus à risque de subir l'événement juste avant le temps $T_i$ ,

— $d_i$ le nombre d'évènements en $T_i$ .

Alors la probabilité $\hat{p_i}$ d'avoir un évènement dans l'intervalle $]T_{i-1},T_i]$ sachant que l'on était vivant en $T_{i-1}$, i.e. $\hat{p_i} = P(X \leq T_i | X > T_{i-1})$, peut être estimée par :

$$\hat{p_i} = \frac{d_i}{Y_i} \tag{3.9}$$

Comme les temps sont supposés distincts, on a :

— $d_i = 0$ en cas de censure en $T_i$, soit quand $\sigma_i = 0$,

— $d_i = 1$ en cas de censure en $T_i$, soit quand $\sigma_i = 1$

L'équation 3.10 donne ainsi l'estimateur de Kaplan-Meier :

$$\hat{S(t)} = \prod_{i=1,...,n,T_i \leq t} (1 - \frac{\sigma_i}{Y_i}) \tag{3.10}$$

Les courbes de survie peuvent ainsi être tracées sur un graphique de la façon suivante : avec généralement un intervalle de confiance à 95%. [Voir calcul de l'intervalle de confiance dans la section 6.3.3.4]. Ces graphiques permettent notamment de voir de façon claire la différence entre les courbes de survie de deux groupes.

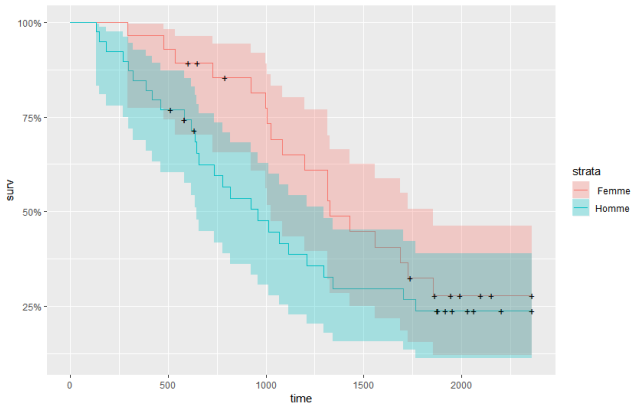



FIGURE 3.3 – Exemple de courbes de Kaplan Meier présentant la survie de deux groupes. En rose, les femmes et en bleu les hommes.

Les courbes de survie de Kaplan-Meier sont représentées par un graphique en marche d'escalier de hauteurs inégales, où la survenue d'un ou plusieurs évènements à une même date représente la verticale d'une marche (la hauteur de la marche proportionnelle au nombre d'événements survenus). La figure 3.3 permet de l'illustrer. Une méthode équivalente de Kaplan-Meier est la méthode actuarielle. Elle se différencie par le fait que les calculs ne se font pas au temps ou des évènements interviennent mais à des temps fixés réguliers, ce qui donne une courbe en segment de droites.

#### 3.1.1.2 Estimateur de Nelson-Aelen

L'estimateur de Nelson-Aalen est une méthode alternative pour estimer la fonction de survie S(t) en temps continu [11]. Soit H(t) la fonction de hasard cumulée. Dans le cas continu :

$$H(t) = \int_{t_0}^{t} h(s)ds = -lnS(t) \tag{3.11}$$

D'où :

$$S(t) = e^{-H(t)} \tag{3.12}$$

L'idée est alors d'estimer S(t) à partir d'un estimateur de H(t). En considérant tous les instants $T_i$ où des événements surviennent jusqu'à l'instant t, nous avons :

$$\hat{H}(t) = \sum_{T_i \leq t} \frac{d_i}{Y_i} \Rightarrow \hat{S}(t) = e^{-\hat{H}(t)}, \tag{3.13}$$

) avec $d_i$ le nombre d'évènements survenant en $T_i$, et $Y_i$ le nombre d'individus à risque de subir l'événement juste avant le temps $T_i$

#### 3.1.1.3 Comparaison Kaplan-Meier et Nelson-Aalen

Asymptotiquement, Kaplan-Meier et Nelson-Aalen sont équivalents. Cependant il est préférable d'utiliser Kaplan-Meier lorsque le hasard diminue au fil du temps, sur de petits échantillons, et Nelson-Aalen lorsque le hasard augmente au fil du temps.
D'autres estimateurs existent tel que Estimateur de Breslow du risque cumulé, Estimateur de Harrington et Fleming de la survie. [Pour plus de détails voir le cours de A. Berchtold [12] ou celui de P. Saint-Pierre [11], desquels sont extraites les définitions].

### 3.1.2 Méthode semi-paramétrique : Cox

Le modèle de Cox se retrouve beaucoup dans la littérature. Primrose et al. [13] l'utilisent par exemple pour étudier l'effet de l'adjuvant capecitabine sur la survie des patients après une chirurgie de résection de cholangiocarcinome ou d'un cancer de la vésicule biliaire. Il est employé lorsque l'objectif est d'évaluer l'effet de covariables sur la durée de vie. Il permet d'expliquer la survenue d'un évènement qualitatif au cours du temps par une ou plusieurs variables explicatives (respectivement analyse univariée et multivariée). qui peuvent être qualitatives ou quantitatives. Pour chacune des variables présentes dans le modèle final, on obtient une estimation du risque relatif (hazard ratio) de survenue du décès en présence de la variable, et de son intervalle de confiance. Le hazard ratio est égal au risque relatif instantané de décès ajusté sur l'ensemble des variables explicatives introduites dans le modèle. Cela implique l'hypothèse que le risque de décès dans les différents groupes d'étude est constant dans le temps et similaire dans tous les sous-groupes.

On introduit une fonction de hasard de base qui donne la forme générale du hasard et qui est commune à tous les individus. Les modèles à hasards proportionnels se caractérisent par la relation 3.14.
Soit un individu i, pour tout t > 0 :

$$\lambda(t|Z_i) = \lambda_0(t) * h(\beta, Z_i), \tag{3.14}$$

avec :

- $Z_i = (Z_{i1}, ...Z_{ip})$ est un vecteur de covariables de dimensions p de l'individu i (les covariables étant des variables propres aux individus pouvant influer la sortie du modèle),
- $\beta$ le paramètre d'intérêt (ne dépend pas du temps t) de dimension p, qui représente l'effet des covariables sur le risque instantané
- h une fonction positive de hasard.

Ce modèle est dit à risques proportionnels car, quels que soient deux individus i et j qui ont pour covariables les vecteurs $Z_i$ et $Z_j$, le rapport des fonctions de hasard ne varie pas au cours du temps,

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \frac{h(\beta * Z_i)}{h(\beta * Z_j)} \tag{3.15}$$

Un cas particulier très important est le modèle de Cox, qui suppose que la fonction h est la fonction exponentielle, c'est-à-dire,en développant les vecteurs , l'équation 3.16 :

$$\lambda(t|(Z_{i1}, ...Z_{ip}) = \lambda_0(t) e^{\beta_1 * Z_{i1} + ... + \beta_p * Z_{ip}} \tag{3.16}$$

D'autres choix de fonctions h sont possibles, néanmoins la fonction exponentielle est très souvent utilisée dans la littérature car ses valeurs sont toujours positives et $e^0 = 1$.
$e^{\beta_0 Z_i}$ est appelé risque relatif.

Le principe de la méthode est d'estimer uniquement le coefficient de régression $\beta$, qui correspond à l'effet des covariables sur le risque instantané. Ils sont obtenus par la méthode de la vraisemblance maximale, et plus particulièrement la méthode de vraisemblance partielle qui comporte l'information sur les coefficients $\beta_i$ [pour plus de détails voir le cours de P. Saint-Pierre [11]]. La fonction $\lambda_0$ est considérée comme un risque de base. L'idée de Cox est qu'aucune information ne peut être donnée sur $\beta$ par les intervalles pendant lesquels aucun événement n'a eu lieu, car on peut concevoir que $\lambda_0$ soit nulle dans ces intervalles (On suppose que les moments où se produisent les censures n'apportent peu ou pas d'information sur $\beta$)[11]. Le modèle de Cox est adapté aux données dont le délai de suivi est variable selon les sujets et aux données censurées. Si la période de suivi est fixe et qu'il n'y a pas de données censurées, le modèle de régression logistique convient aussi bien que le modèle de Cox.

## 3.2 Les méthodes d'apprentissage automatique

Outre ces méthodes, l'apprentissage automatique est de plus en plus utilisé pour étudier la survie. On peut par exemple trouver dans la littérature la méthode K-NN (K Nearest Neighbors), Baysienne [14] [15] , ou encore les méthodes basées sur les arbres. La majorité de ces méthodes, comme le Support Vector Machine (SVM) sont des méthodes de classification et de régression qui ne sont pas forcément adaptées pour l'analyse de la survie et notamment pour les données censurées.

### 3.2.1 Les SVM

Les SVM possèdent une structure flexible et sont régulièrement retrouvés dans les articles comme dans l'article de Fayçal Ben Bouallègue et al [16]. Ils permettent la régression et la classification. On part d'un ensemble $(x_i, y_i) i = 1, \ldots, n$ , avec $x_i$ un vecteur de p caractéristiques pour l'individu i et $y_i$ la classe de l'individu. Dans le problème de classement en deux classes (par exemple bon et mauvais pronostic), le but est de construire une fonction qui permet de prédire si notre nouvel exemple appartiendra à la classe 1 ou 2. On cherche alors une surface de séparation. C'est-à-dire qu'il faut trouver la fonction f tel que f(x)=0 sépare les deux classes avec le moins d'erreurs possibles comme dans la figure 3.4. [Plus de détails dans le cours de M. Crucianu [17].

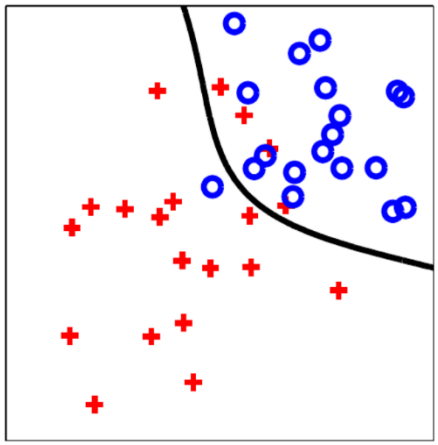


FIGURE 3.4 – Exemple de problème de séparation à deux classes. Il faut trouver une surface de séparation f qui permette de classer les éléments [17].

### 3.2.2 Les réseaux de neurones

Les réseaux de neurones permettent la régression et la classification. Ils sont définis comme une succession de couches de neurones, chacune prenant ses entrées sur les sorties précédentes [voir figure 3.5]. Chaque couche est composée de neurones $N_i$ . A chaque liaison entre deux neurones est associé un poids $W_{ij}$, de sorte que les neurones $N_{i-1}$ sont multipliés par ce poids, puis additionnés par les neurones $N_i$. A chaque couche est associée une fonction d'activation $\varphi$ qui permet d'avoir la sortie (ex : si l'âge est supérieur à 40 ans, sortie = 1 et 0 sinon). [Plus de détails sont disponibles dans le cours de M. Parizeau [18]]

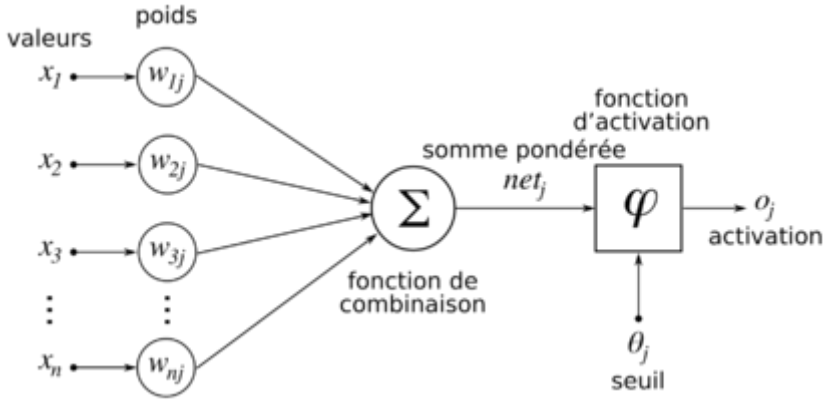



FIGURE 3.5 – schéma d'un réseau de neurones

### 3.2.3 Les Random Forest (RF)

Les Random forest (RF) ou forêts aléatoires sont une méthode d'apprentissage automatique basée sur les arbres et permettent de traiter la régression, de la classification bi-classe ou multi-classe. Les forêt et le caractère aléatoire permettent aux RF de palier au problème d'instabilité des arbres tels que les arbres CART (Classification And Regression Trees) de Leo Breiman. De nombreux algorithmes furent proposés comme le bagging ([19], 1996) et Addaboost([20], 1997). Mais ce sont les forêts aléatoires de Breiman ( [21], 2001) qui seront les plus performantes sur le plan expérimental, et qui sont de plus en plus utilisées. Les forêts aléatoires sont une collection d'arbres qui sont entraînés sur des sous-ensembles aléatoires et indépendants d'échantillons. La partition à chaque noeud, se peut aussi se faire sur un sous-ensemble aléatoire de variables (ou caractéristiques). Les méthodes d'arbres et de forêts aléatoires seront expliquées en détails dans la section 5.3.1.

*Chapitre 4*

# État de l'art

Le chapitre état de l'art à pour objectif d'expliquer ce qui existe dans la littérature, en commençant tout d'abord au niveau de l'analyse de la survie. Nous parlerons ensuite de l'utilisation de l'imagerie médicale pour l'étude de la survie, dans le domaine médical en général. Après cela, nous parlerons de l'étude de la survie dans le cas de patients atteints du myélome multiple, ainsi que l'imagerie médicale qui est associée à cette maladie, et enfin nous discuterons de l'utilisation de l'imagerie TEP dans la détection des lésions et l'étude de la survie, sans se limiter au myélome multiple. Nous conclurons cet état de l'art en expliquant nos choix.

## 4.1 L'analyse de la survie

Dans la littérature, différentes méthodes sont présentées pour l'étude de la survie. Pour estimer la survie d'un groupe ou comparer la survie entre deux groupes, la méthode de Kaplan-Meier est souvent utilisée comme par exemple dans l'article de Vallières [22]. En effet, cette méthode a pour avantage d'être simple d'utilisation et d'interprétation, et ne nécessite pas d'émettre des hypothèses sur les distributions de survie. Elle est ainsi, régulièrement utilisée pour comparer la survie de deux groupes de patients par exemple.

Cependant, celle-ci ne suffit plus lorsque l'on veut évaluer la survie de patients, individuellement en fonction de leurs variables, ou évaluer l'impact des covariables sur la survie. Auparavant, l'analyse de la survie se faisait principalement à l'aide du modèle de Cox ([23] 1972). Cependant, ces estimateurs sont de plus en plus remplacés par des méthodes d'arbres et d'apprentissage automatique. L'article de Yan Zhou et John J. McArdle [24], sur l'application des arbres et ensembles de survie nous présente les intérêts et inconvénients qu'apportent les différentes méthodes d'arbres. Ils montrent que la régression de Cox pose un problème dans deux cas : les résultats peuvent être biaisés lorsque la censure est reliée aux variables d'exploration, et le pouvoir statistique est affecté par un haut taux de censure. Les arbres de survie peuvent être utiles pour détecter un décalage dans des relations non linéaires mais aussi détecter les interactions entre les variables.

Yan Zhou et John J. McArdle [24] nous indiquent ainsi que les forêts de survie de Breiman (2002) sont plus performantes que la régression de Cox (spécialement lorsque la régression de Cox ignore les variables qui sont prédictives sur une période de temps et non tout le temps). De plus il est expliqué qu'il y a des conditions où les forêts de survie sont plus intéressantes que la régression de Cox, notamment dans le cas où il y a grand nombre de prédicteurs et une petite taille d'échantillons. Parmi les algorithmes d'arbres de survie, le « conditional inference survival tree » développés par Hothorn en 2006 [25] semblent plus fiables et moins enclins à « l'overfitting » (ou surapprentissage. Se dit lorsqu'un modèle correspond trop étroitement ou exactement à un ensemble particulier de données et peut donc ne pas correspondre à des données supplémentaires ou ne pas prévoir de manière fiable les observations futures) que d'autres méthodes d'arbres comme les "bagging survival trees" (Hothorn et al.[26],

2004), et les "random survival forests" (Ishwaran et al. [1], 2008) mais ce problème est compensé dans ces deux dernières méthodes par l'utilisation d'un grand nombre d'arbres. Il a de plus été montré par Iswharan [1] que, bien que la méthode d'Horthorn [25] soit bonne dans les cas où le taux de censure soit faible, elle dépend grandement de celui-ci. Ainsi Iswharan [1] propose en 2008, un algorithme de forêts aléatoires adaptées à la survie et à la censure à droite : les "Random Survival Forest" ou RSF. La méthode est robuste aux données de censure et au bruit dû aux variables.

L'article de Ingrisch et al. [27], utilise ainsi les RSF pour prédire la survie chez des patients porteurs de tumeurs intra-hépatiques traitées par radioembolisation à l'$^{90}Y$ et identifier les variables prédictives. Il compare de plus, les résultats à ceux obtenus par le modèle de Cox et le score de concordance (c-index) est équivalent dans les deux cas. Ceci s'explique par le fait que peu de variables sont testées (une quinzaine) pour un grand nombre de patients (366) et seulement 38% de censure.

Les RSF sont maintenant grandement utilisés et sont la méthode de référence pour l'étude de la survie (Autres exemples d'utilisation : [28], [29], [30]). Il est apparu depuis peu des améliorations de la méthode : par exemple, Fen Miao et al. [31] proposent en 2018 une nouvelle méthode basée sur les RSF en modifiant le critère d'arrêt et le critère de séparation, ce qui permet d'avoir de meilleurs résultats lorsqu'une variable est très prédictive mais peu présente. Arabin Kumar Dey et al. (2018) [32] proposent une méthode basée sur RSF mais qui utilise « Extremly randomize trees » et « Adaboost » permettant d'augmenter la rapidité et les performances pour des bases de données de grandes dimensions par rapport à RSF.

Outre ces méthodes de survie, sont aussi utilisées des méthodes de classification. On peut citer l'utilisation du SVM (par exemple dans l'article de Fayçal Ben Bouallègue et al [16]) où des Neural Network (exemple dans l'article de Edward Choi, et al. [33]), mais surtout les classifieurs forêts aléatoires comme (l'étude de Vallières et al. sur les cancers tête-cou en 2017 [22]). L'article de Parmar et al. [14] compare 12 méthodes différentes de classification et 14 méthodes différentes de sélection de paramètres sur une base de données du cancer du poumon.

Comme on peut le voir dans la figure 4.1, ils démontrent que la méthode la plus efficace parmi toutes celles testées et pour la majorité des méthodes de sélection de paramètres, est la méthode RSF, selon l'AUC.

Cependant ces méthodes présentées dans l'article de Parmar et al. [14] ne permettent pas de prédire une survie, seulement une classe. Enfin, de plus en plus d'articles parlent de l'utilisation de l'apprentissage profond pour l'étude de la survie. En 2016, Linxia Liao et Hyung-il Ahn [34] proposent un modèle à 3 couches d'apprentissage profond qui donne de bien meilleurs résultats que le modèle de Cox mais n'est pas comparé aux RSF. Un autre article, de Jared L.Katzman et al de 2017 [35], compare leur modèle « Cox proportional hazards deep neural network » (DeepSurv) avec le modèle de Cox et les RSF dans le cadre de la recommandation d'un traitement personnalisé. Il semble donner de meilleurs résultats que les deux derniers en termes de précision dans la prédiction sur les données de survie et pour déterminer le traitement recommandé.

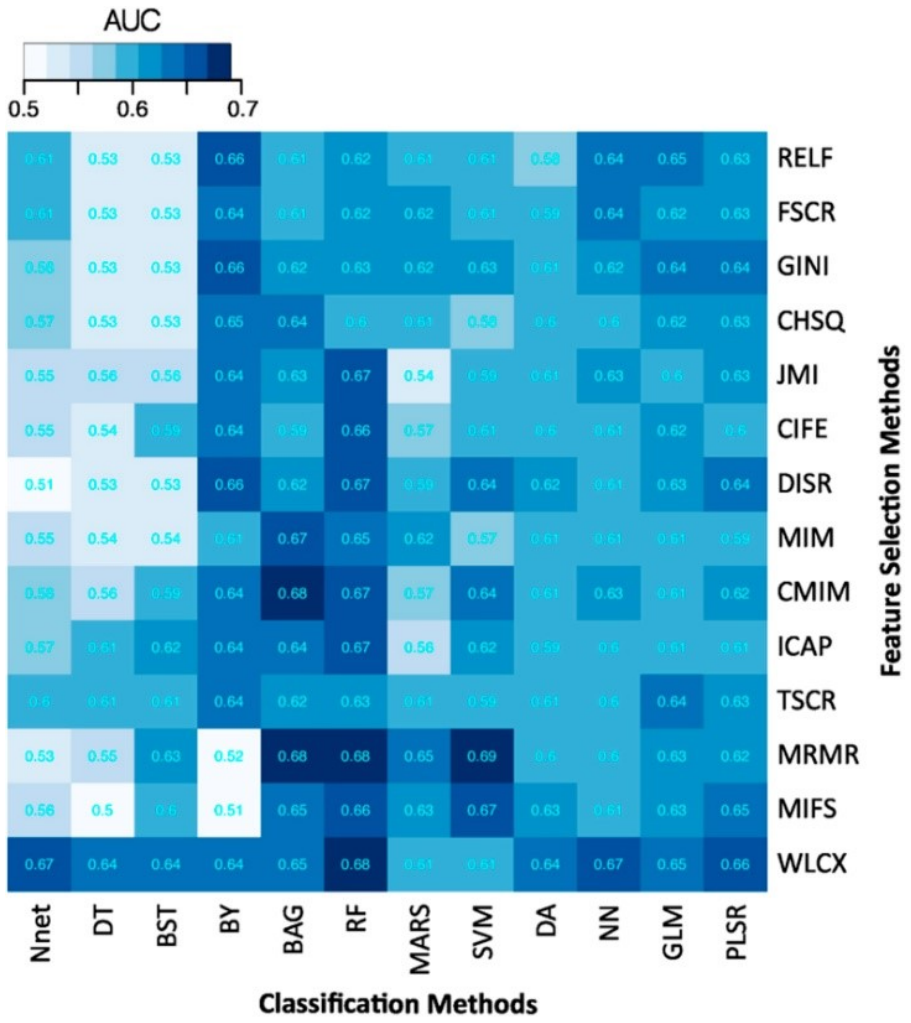



FIGURE 4.1 – Comparaison de 12 méthodes de classification ( Nnet : Neural network, DT : Decision Tree, BST : Boosting, BY : Bayesian, BAG : Bagging, RF : Random Forest, MARS : Multi adaptive regression splines, SVM : Support vector machines, DA : Discriminant analysis, NN : Neirest neighbour, GLM : Generalized linear models, PLSR : Partial least squares and prinicipal componenet regression) et 14 méthodes de sélections de paramètres. La comparaison se fait sur la valeur de l'AUC (Area Under Curve) réalisée dans l'article de Parmar et al. [14]

## 4.2 L'utilisation d'images médicales pour l'étude de la survie

La plupart des études de survie se basent sur des variables cliniques. Cependant, de plus en plus d'articles louent les mérites de l'utilisation de la radiomics comme paramètre en entrée des modèles ( [36], [15], [37]). La radiomics est définie par C. Bourgier et al. [38] par un outil qui « permet une analyse qualitative et quantitative ultra performante, consistant en l'extraction à haut débit de données numériques d'imagerie médicale afin d'obtenir des informations prédictives et/ou pronostiques concernant les patients pris en charge pour une pathologie cancéreuse ».

La radiomics peut être réalisée à partir d'images tomographique provenant de CT, IRM et TEP, ou n'importe quelle autre modalité. Ces données images sont souvent accompagnées par des données cliniques ou génomiques. L'article de Parmar et al. [14] présente différents modèles d'apprentissage automatique avec différentes méthodes de sélection de caractéristiques en utilisant des caractéristiques (texturales, de forme, d'intensité et basées sur des ondelettes) extraites à partir d'images CT. L'article de Hugo J. W. L. Aerts [39] montre qu'un grand nombre de caractéristiques ont un pouvoir pronostic dans des bases de données indépendantes de cancer du poumon

et tête-cou. Ils indiquent que c'est une méthode rapide, peu chère et non invasive pour étudier l'information phénotypique, et que la signature radiomique est significativement associé à des motifs d'expression de gènes sous-jacents. L'article de F. Ben Boullègue [16] montre que la combinaison de facteurs pronostiques habituels avec des paramètres de texture et de forme appropriés permettent d'améliorer la prédiction d'une réponse métabolique précoce dans plusieurs types de lymphome. Ils montrent de plus l'impact de la segmentation sur les résultats. Les images sont de plus en plus utilisées comme facteur pronostique, comme par exemple pour le cancer du poumon ( [40], [41]), le lymphome [42], le cancer tête-cou ( [22]), le cancer de l'œsophage [43] ou encore le carcinome [44]. L'utilisation de l'imagerie pour la prédiction de la survie sur le myélome multiple existe aussi dans la littérature, comme par exemple dans l'article de Lapa et al. [45]. C. Bhnemann et al. [30] utilisent des RSF avec en entrée des caractéristiques provenant d'images confocalee de « tissue microarrays ». Cependant, outre C. Bhnemann et al. et nous, l'application des RSF aux caractéristiques images reste encore peu étudiée.

## 4.3 Myélome multiple et survie

Peu d'articles tentent de prédire la survie des patients atteints de myélome multiple. La majorité des papiers mettant en relation apprentissage automatique et myélome multiple s'attellent à la segmentation et la détection des lésions. C'est le cas de Xu et al. [46] par exemple.

D'autres tentent de prédire la survie des patients atteints de myélome multiple à partir de l'expression génique ( [47], [48]). Comme la majorité des articles médicaux, O. Decaux et al. [47] utilisent les méthodes de Cox et de Kaplan Meier pour réaliser l'étude. Amin et al. [48](2014) testent plusieurs méthodes de apprentissage automatique (Compound covariate predictor, Linear discriminant analysis, K-nearest neighbor, Nearest centroid, SVM : Support vector machine) pour prédire une réponse complète en fonction du profil d'expression génique.

Outre l'article de H. Pang, M. Hauser et S. Minvielle [49] qui utilise une base de données de myélome multiple pour montrer la corrélation entre la survie et les polymorphismes du nucléotide simple, et nous, aucun papier ne présente pour l'instant l'utilisation de RSF pour l'étude de la survie des patients atteints de myélome multiple.

La majorité des études reliant myélome et imagerie, ont pour but de déterminer la méthode la plus efficace de détection des lésions. La méthode traditionnelle de détection des lésions est la radiologie conventionnelle planaire. Cependant, d'autres méthodes sont utilisées et ont de meilleurs résultats. Le CT permet de détecter des lésions osseuses plus petites qui ne sont pas détectables avec la radiographie conventionnelle [50]. L'IRM est aussi beaucoup utilisée car il est plus sensible et peut détecter infiltration de la moelle osseuse diffuse avec une bonne différentiation des tissus mous [51], [52]. L'utilisation de la 18-FDG-TEP combiné au CT permet aussi une bonne sensibilité [53] [54] [55] [56]. Plus récemment, de nouveaux traceurs font leur apparition, comme le 68Ga-Pentixafor qui permet une haute sensibilité de détection des lésions du myélome multiple. [46] [57].

## 4.4 L'utilisation de l'imagerie TEP pour la détection de lésion et la prédiction de survie en général

Des articles tels que ceux de [58], Desseroit MC et al. [59], M. Hatt et al. [60] et Bailly et al. [61] ont investigué pour déterminer quels sont les caractéristiques les plus intéressantes, celles qui dépendent le moins de la segmentation et celles qui sont liées. M. Vallières et al. [28] s'intéresse à la recherche de nouvelles textures composites entre TEP et IRM pour mieux identifier les tumeurs agressives, et montre que les caractéristiques extraites des images FDG-TEP scan sont généralement plus prédictives que celles extraites de MRI, dans le cas

des métastases pulmonaires d'un sarcome mais la valeur prédictive est fortement augmentée lors de l'association des deux imageries. T. Carlier et al. [62] montrent l'intérêt de l'hétérogénéité déterminée sur FDG-TEP au diagnostic chez des patients atteints de myélome multiples. Ils rapportent aussi que des études prospectives ont prouvé la valeur pronostique de plus de 3 lésions focales, de la SUV Max, des lésions extramédullaires ( [63], [64]) et du volume métabolique total et de la glycolyse totale [65]. Tixier et al. [43] ont démontré que l'analyse texturale d'images FDG-TEP scans peut prédire la réponse à un traitement contre le cancer de l'œsophage.

Certains articles utilisent la images TEP (associées à des images CT) pour l'étude de la survie comme l'article de Vallières et al. [22], celui de F. Ben Bouallègue et al. [16], et celui de L. Bi et al. [42] avec des classifieurs forêts aléatoires, SVM et Deep CNN respectivement. Des TEP scans ont été utilisés pour montrer la stabilité des caractéristiques radiomics dans un groupe de patients atteint de NSCLC (Non-Small Cell Lung Cancer) dans l'article de Larue et al. [37]. Cependant aucun d'eux n'associent RSF et imagerie TEP. Seul un article récent, de Steigner et al. [44] associe des images FDG-TEP et des RSF dans le cadre de la détermination de la mortalité pour des patients atteints de carcinomes bronchique mais s'intéresse principalement a un modèle "survival tree" et le modèle de Cox.

Etant donné les bons résultats, la FDG-TEP reste une des méthode d'imagerie les plus utilisés dans l'exploration clinique du myélome multiple (généralement couplée au CT). C'est aussi la modalité utilisée au CHU de Nantes (le 68Ga-Pentixafor étant encore un traceur récent), ce qui nous amène donc à son utilisation dans le cadre de l'étude du myélome multiple, non seulement pour la prédiction de survie mais pour la détection des lésions. L'article de C. Bodet-Milin et al. [66] confirme l'intérêt d'utiliser la TEP. En effet, la FDG-TEP de corps entier permet de détecter les lésions myélomateuses avec une sensibilité de 90% contre 70% avec IRM.

## 4.5 Conclusion

Dans l'étude de la survie, un grand nombre de papiers prouvent l'intérêt de l'utilisation de RSF par rapport aux méthodes plus conventionnelles. De plus, la méthode RSF reste relativement récente et malgré l'intérêt grandissant pour la radiomics et son efficacité prouvée, l'utilisation de caractéristiques provenant d'images médicales avec des RSF reste peu commun. Ceci nous amène à s'intéresser à l'utilisation des RSF pour la prédiction de la rechute chez les patients atteints de myélome multiple, en y associant de la radiomics, qui, d'après la littérature a un intérêt certain. L'imagerie choisie correspond à l'imagerie FDG-TEP. En effet, pour l'étude du myélome multiple, le choix peut se porter sur l'IRM, l'imagerie TEP ou le CT. Or, comme l'indique P. Moreau et al. [6], les images TEP sont équivalents en terme de détection de lésions mais les images TEP permettent une meilleur prédiction de la PFS ou de l'OS, ce qui fait de cette méthode notre premier choix.

*Chapitre 5*

# Méthodes

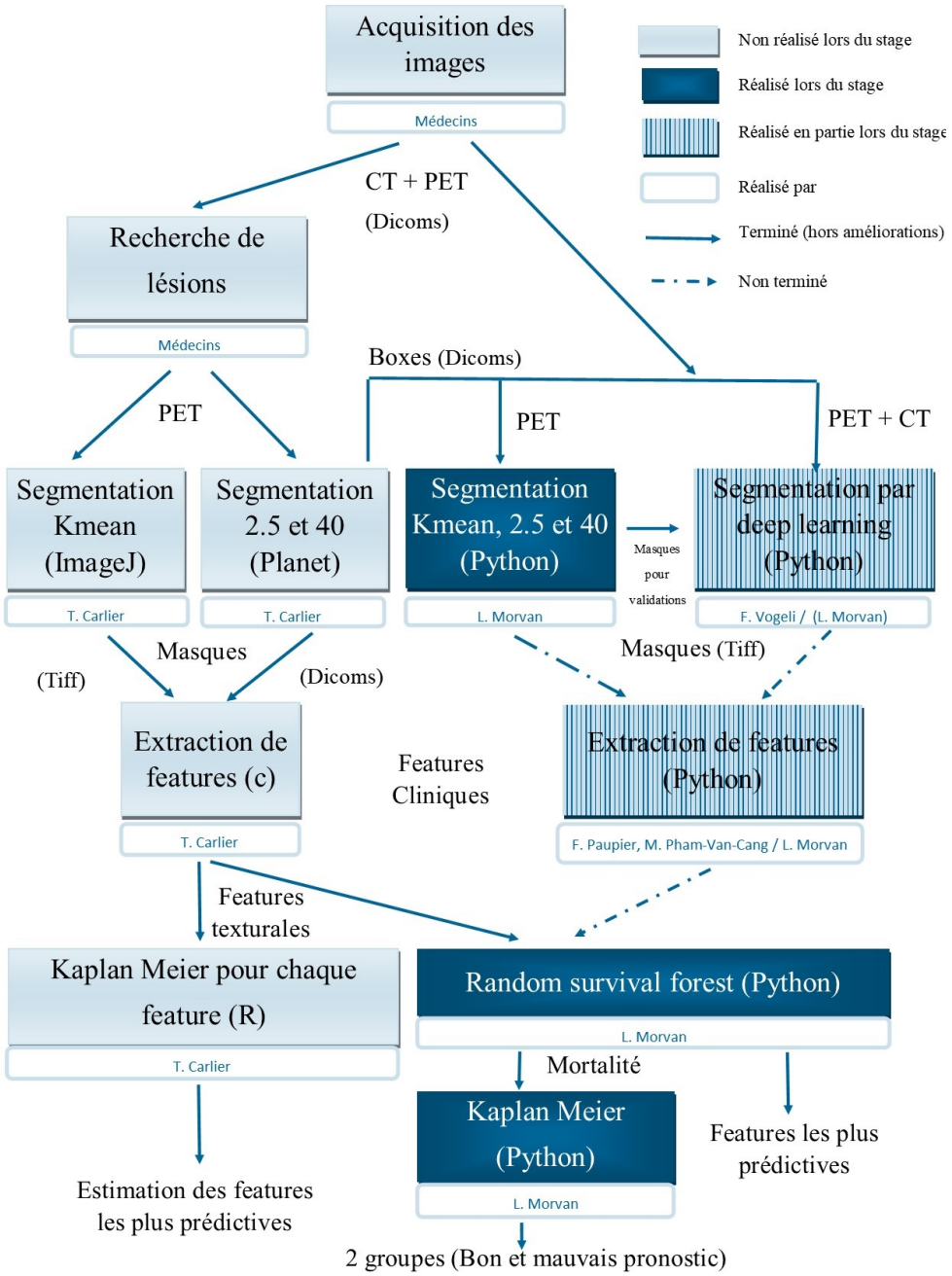



FIGURE 5.1 – Diagramme UML du pipeline du projet

Ce chapitre présente les différentes méthodes utilisées correspondant aux différentes étapes du pipeline. L'objectif est ici de réaliser toutes ces étapes dans un seul langage. La première étape concerne la détection et la segmentation des lésions. Elles sont faites manuellement ou semi-automatiquement à l'hôpital pour le moment et demande de nombreuses étapes. Elles ont donc été segmentées en python par les méthodes 2.5, 40% et k-means, et un vote majoritaire y a été appliqué. Il est aussi introduit la méthode d'apprentissage profond V-net qui permettrait à long terme de non seulement segmenter mais aussi détecter les lésions.

La deuxième partie concerne l'extraction des caractéristiques texturales qui est pour le moment fait en c.

La troisième partie concerne l'étude de la survie. Elle introduit les méthodes d'arbres et plus particulièrement les RSF, la méthode utilisée par la suite, ainsi que l'implémentation réalisée en python . Les différentes étapes du pipeline sont résumés dans la figure 5.1.

## 5.1 Détection et segmentation des lésions

### 5.1.1 Les segmentations manuelles

**Les segmentation 40% et 2.5** La segmentation à 40% correspond à considérer dans la lésions tous les voxels qui ont une valeur de SUV supérieur à 40% du Suv max :

*Si* $value \geq 0.4 * max(masque)$ *: Le pixel est dans la lésion,*
*Sinon : Le pixel n'est pas dans la lésion*

La segmentation à 2.5 correspond à considérer dans la lésions tous les voxels qui ont une valeur de SUV supérieur à 2.5 SUV.

**La segmentation par k-means** La méthode des k-means est un outil de classification qui permet de répartir un ensemble de données en k classes homogènes mais permet aussi d'apporter une solution à la segmentation d'images. On cherche ainsi à diviser l'espace en classes de groupes de voxels. L'algorithme des k-means vise à minimiser la variance intra-classe, qui se traduit par la minimisation de l'énergie suivante :

$$E = \frac{1}{2} \sum_{c \in C} \sum_{x \in c} ||x - m_c||_2 \tag{5.1}$$

avec C l'ensemble des clusters et pour chaque cluster c, x un élément du cluster et $m_c$ son centre (appelé noyau). La minimisation de cette énergie peut se réaliser par une descente de gradient sur les noyaux. Elle peut se traduire par les étapes suivantes :

1. Initialisation des noyaux.
2. Mise à jour des clusters.
3. Réévaluation des noyaux.
4. Itérer les étapes 2 et 3 jusqu'à stabilisation des noyaux.

On peut donc appliquer cet algorithme afin de réaliser 2 classes (Lésion/ fond).

### 5.1.2 Le W-net et méthode des patchs

Xu et al. [46] présentent, une méthode d'apprentissage profond pour la détection et la segmentation des lésions chez les patients atteints de myélome multiple : le W-net.

Xu et al. partent du V-net. Le V-net est une adaptation 3D de l'U-net, une structure 2D pour la segmentation des images médicales en utilisant un chemin de contraction pour extraire l'environnement des lésions et un chemin symétrique pour la localisation des lésions. Le w-net correspond ainsi à l'association de deux V-net

améliorés en cascade permettant de connaître les caractéristiques volumétriques du squelette et de ces lésions, des détails les plus grossiers aux plus fins. Il ne nécessite que peu de pré-traitement et pas de post-traitement. Pour le premier V-net, seules les données CT sont utilisées afin de connaître l'anatomie osseuse. On en ressort un masque binaire du squelette. Le second V-net utilise les images CT et les images TEP, et fournit des caractéristiques additionnelles pour la détection des lésions.

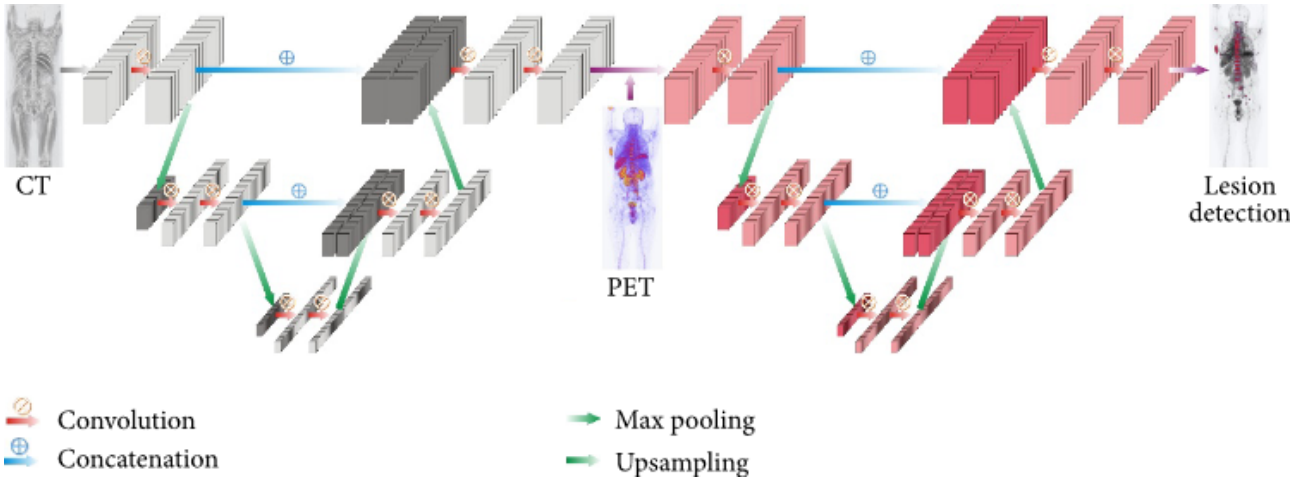



FIGURE 5.2 – Présentation d'un W-net simplifié. En gris le premier V-net pour la détection du squelette, et en rose le second V-net pour la détection des lésions [46].

Dans le but de détecter les petites lésions et équilibrer la taille des différents os, une stratégie de contrepoids est utilisée. Ainsi, pour les images CT, de volume V, les labels sont 0 pour les régions sans os et 1 sinon. Pour les images PET, 0 définit une région non myélomateuse et 1 sinon. Soit $p(x_i = l \vee V)$ la probabilité que le voxel i ait le label l dans le volume V avec $l \in [0, 1]$. La fonction de perte de cross-entropie s'écrit alors :

$$Lc = c\frac{-1}{N}\sum_{i=1}^{N}\omega_i^{label}[\hat{p}_i log p_i + (1-\hat{p}_i)log(1-p_i)] \tag{5.2}$$

avec $\hat{p}_i$ la vérité terrain, $p_i$ la probabilité assignée au voxel i quand il appartient à l'os, $\omega_i^{label}$ une valeur inversement proportionnel au nombre de voxels qui appartient à ce label.

A côté de ça est adopté une autre stratégie de poids, qui est la stratégie des patchs équilibrés. Le set d'entraînement est sous-échantillonné pour réalisé un écart entre les labels os, lésion et fond. Pour chaque patient, le réseau est pré-entraîné en extrayant des patchs (64x64x64) sur le corps entier du patient puis 30 patchs sont sélectionnés sur la base du ratio entre le label et le fond dans chaque patch, ce qui améliore le ratio Label/fond.

Les résultats du V-net concernant la segmentation avec soit CT, soit PET donne un score Dice entre 26 et 29 % lorsque l'association des deux modalités (PET et CT) donne 69,49% et le W-net avec les deux modalités un score Dice de 72,98%. En ce qui concerne la détection des lésions, les meilleurs spécificité, sensibilité et précisions sont toutes obtenues avec le W-net.

Les résultats de classification en lésion/non lésion du V-net (avec les deux modalités) sont aussi comparés aux classifieurs forêts aléatoires, k-Nearest Neighbor (k-NN) classifier, et au support vector machine (SVM) . Bien qu'il ait une sensibilité légèrement plus basse, le V-net a une spécificité plus élevée (99,68%), et une précision et un score Dice beaucoup plus élevés (respectivement 88,82% et 89,26% contre moins de 16% et moins de 27% pour les autres méthodes).

### 5.1.3 Contribution

Les segmentations utilisées pour le moment sont les segmentations à 40% du SUV max, à 2.5 SUV et par K-mean. Les segmentations 2.5 et 40 sont réalisées à l'aide du logiciel Planet (Dosisoft, Cachan, France) par les médecins. Les masques récupérés sont sous un format DICOM (Digital Imaging and Communications in Medicine) (une pile de DICOM par masque). La segmentation k-means est réalisée pour le moment à l'aide du logiciel ImageJ et les masques sont récupérés au format Tif.

Enfin, un vote majoritaire est réalisé sur les trois masques pour chaque lésion, afin d'obtenir un masque final de la lésion. La méthode de vote majoritaire sur des images consiste, pour chaque pixel de l'image, à choisir s'il appartient ou non à la lésion en moyennant les pixels des 3 masques. Si un pixel a une valeur supérieure à 0.66, alors il appartiendra à la lésion.

Comme le montre la figure 5.1, pour le moment, la détection et la segmentation des lésions ne sont pas automatiques, demandent beaucoup de temps et sont réalisées sur différents logiciels par différentes personnes, ce qui peut conduire à des erreurs. En effet, à partir des images CT et TEP les médecins déterminent la position et le nombre de lésions présentes.

Lors du stage seules les méthodes 2.5 et 40% ont été réalisées (le k-means n'a pas encore été implémenté). Le calcul se fait à l'aide de boîtes récupérées à l'aide du logiciel Planet. En effet, par la suite il serait intéressant que les segmentations ne soient plus réalisées par les médecins, mais cela implique tout de même qu'elles soient détectées. Pour ce faire, le médecin sélectionne les lésions dans le logiciel Planet. Ceci va créer pour chaque lésion une boîte amorphe et grossière englobant la lésion. C'est dans ces masques que sont réalisés les segmentations 2.5 et 40% (puis k-means). Le vote majoritaire a aussi été implémenté.

L'objectif futur pourra être la segmentation des lésions par l'apprentissage profond afin de gagner du temps mais aussi parce que l'apprentissage profond, avec assistance des médecins, pourrait diminuer le nombre d'erreurs. Pour l'instant la méthode implémentée est le V-net. Ainsi, la première contribution du stage est la standardisation d'un pipeline de traitement des images, ce qui inclus la segmentation. La segmentation va permettre d'obtenir des masques de chaque lésions dans lesquelles seront calculées les caractéristiques radiomiques.

## 5.2 Les caractéristiques radiomiques

Afin d'intégrer les images en entrée des méthodes de prédiction de survie, il faut réaliser différents calculs permettant de les caractériser. Ces calculs donnent les caractéristiques radiomiques. Elles peuvent être texturales ou volumiques.

### 5.2.1 Les différentes caractéristiques texturales

Des caractéristiques quantitatives peuvent être extraites d'images tomographiques et notamment des images TEP afin de décrire les lésions. Pour chaque caractéristique, les calculs sont appliqués sur les voxels appartenant au masque de la lésion. Il y a deux principales catégories de caractéristiques extraites ; les agnostiques et les sémantiques. Les sémantiques sont celles qui sont généralement utilisées en radiologie pour décrire la région d'intérêt, et les agnostiques pour évaluer l'hétérogénéité à travers des descripteurs quantitatifs. [15] Les caractéristiques agnostiques peuvent être séparées en caractéristiques de premier, second ordre, voir plus.

— Premier ordre : description de la distribution des valeurs des voxels individuels sans prendre en compte les relations dans l'espace (moyenne, max, min, uniformité, entropy, des intensités, asymétrie, kurtosis de l'histogramme des valeurs).

| Sémantiques | Agnostiques |
|---|---|
| Taille | Histogramme (asymétrie, kurtosis) |
| Forme | Textures de Haralick |
| Localisation | Dimensions fractales |
| Vascularisation | Ondelettes |
| Nécrose | Transformations Laplaciennes |

TABLE 5.1 – Exemples de caractéristiques sémantiques et agnostiques

— Second ordre : description de la texture (relation entre les voxels d'intensité similaire) et sont calculés en utilisant matrice de co-occurrence. (GLCM, GLSZM, GLRLM, ...)

Le tableau 5.1 donne des exemples de caractéristiques agnostiques et sémantiques.

Les définitions peuvent différer. Les définitions données ici sont celles de l'ISBI [67].

#### 5.2.1.1 Caractéristiques de premier ordre

Voici la définition de quelques exemples de caractéristiques du premier ordre que l'on retrouve régulièrement (Cependant lors de ce projet, seul le maximum est gardé).

Soit :

1. X, un ensemble de Np voxels inclus dans la région d'intérêt (ROI)
2. $H_i$ l'histogramme de premier ordre avec Ng niveaux d'intensités, avec Ng le nombre de bins non nuls, également répartis entre 0 et le binWidth choisi.
3. $h_i$, l'histogramme normalise égal à $H_i$*Np

**L'énergie :**

$$Energie = \sum_{i=1}^{Np}(X(i)+c)^2 \tag{5.3}$$

Avec c, une valeur optionnelle qui change les intensités pour prévenir des valeurs négatives de X. L'énergie traduit la magnitude des valeurs de voxels de l'image.

**L'entropie :**

$$Entropie = -\sum_{i=1}^{Ng} p(i) * log_2(p(i)) \tag{5.4}$$

L'entropie traduit le caractère aléatoire des valeurs dans l'image.

**Asymétrie :**

$$Asymtrie = \frac{\frac{1}{Np}\sum_{i=1}^{Ng}(X(i)-\bar{X})^3}{(\sqrt{(\frac{1}{Np}\sum_{i=1}^{Ng}(X(i)-\bar{X})^2})^3} \tag{5.5}$$

L'asymétrie est celle de la distribution des valeurs par rapport à la moyenne $\bar{X}$. L'asymétrie est celle de la distribution des valeurs par rapport à la moyenne.

#### 5.2.1.2 Les caractéristiques de second ordre

Les caractéristiques de second-ordre sont basées sur des matrices de co-occurrence, dont voici quelques exemples.

**Gray Level Co-occurrence Matrix (GLCM)**

La matrice est de taille NgxNg GLCM traduit la probabilité jointe P(i,j|$\sigma$,$\theta$) de la région contenue dans le masque. La position (i,j) représente le nombre de fois que la combinaison des niveaux i et j apparaît dans deux pixels de l'image et sont séparés par une distance de $\sigma$ pixels et d'un angle $\theta$.
Exemple :
Soit une matrice d'intensité I. Pour une distance de 1 et un angle de 0° (horizontal plan, de gauche à droite) la matrice de co-occurrence P sera :

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 5 | 2 | 3 |
| 3 | 2 | 1 | 3 | 1 |
| 1 | 3 | 5 | 5 | 2 |
| 1 | 1 | 1 | 1 | 2 |
| 1 | 2 | 4 | 3 | 5 |

(a) Matrice I

| | | | | |
|---|---|---|---|---|
| 6 | 4 | 3 | 0 | 0 |
| 4 | 0 | 2 | 1 | 3 |
| 3 | 2 | 0 | 1 | 2 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 3 | 2 | 0 | 2 |

(b) Matrice GLCM

FIGURE 5.3 – Construction de la matrice GLCM

En effet, si on prend l'exemple de la valeur (1,2), la valeur sera 4 car de façon horizontale le couple (1,2) apparaît deux fois dans la matrice I (en bleu dans la matrice I). Le couple (4,1) n'apparaît pas, donc la valeur de P(i,j|1,0)sera de 0 : Sur cette matrice P seront ensuite faits différents calculs. Les caractéristiques calculées sur cette matrice sont : l'entropie [voir l'équation **??**], la corrélation, le contraste, l'énergie [voir équation ?] et la dissimilarité.

**Gray Level Size Zone Matrix (GLSZM)** :

La matrice GLSZM quantifie les niveaux de gris dans l'image. Une zone de niveau gris est définie par le nombre de voxels connexes qui partagent la même intensité de niveau de gris. Ainsi, l'élément (i,j) est égal au nombre de zones de niveau de gris i et de taille j qui apparaissent dans l'image. Une seule matrice est calculée pour toutes les directions contrairement aux matrices GLRLM et GLCM.
Exemple :
Soit une matrice I à 5 niveaux de gris :

| | | | | |
|---|---|---|---|---|
| 5 | 2 | 5 | 4 | 4 |
| 3 | 3 | 3 | 1 | 3 |
| 2 | 1 | 1 | 1 | 3 |
| 4 | 2 | 2 | 2 | 3 |
| 3 | 5 | 3 | 3 | 2 |

(a) Matrice I

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |

(b) Matrice GLSZM

FIGURE 5.4 – Construction de la matrice GLSZM

En effet, il y a par exemple 1 zone de taille 4 à valeur 1 (en rose).
Une des caractéristiques calculées sur cette matrice est par exemple la Small zone high grey level emphasis (SZHGE ou SAHGLE) :

$$SZHGE = \frac{\sum_{i=1}^{Ng} \sum_{j=1}^{N_s} \frac{i^2 P(i,j)}{j_2}}{Nz} \tag{5.6}$$

Avec Ns le nombre de tailles de zones dans l'image et Nz le nombre de zones dans la ROI. Cette caractéristique mesure la proportion dans l'image de la distribution jointe de plus petites zones avec de plus grandes valeurs de niveau de gris.

### 5.2.2 Contribution

Le calcul de ces matrices et caractéristiques peut se faire de diverses façons. En effet, les matrices peuvent être calculées en 2D ou 3D. Dans le cas du stage seul le calcul en 3D est gardé –> (0M : One matrix). De plus, il existe plusieurs méthodes de rééchantillonnage. Deux ont été testées ici :

— AR : Absolute Resampling. La taille du binwidth reste fixe pour tous les patients (ici 0.3).

— RR : Relative Resampling. Le nombre de binwidth qui reste fixe (ici 64).

De plus, on peut égaliser l'histogramme (on appellera Heq dans la suite du rapport lorsque l'histogramme est égalisé) ou non pour construire la matrice. Enfin, on peut utiliser une taille de voxel qui varie ou non (on appellera equalsize dans la suite du rapport lorsque la taille de voxel ne varie pas). Au total 6 méthodes sont testées, ce qui donne 114 caractéristiques :

— OMAR

— OMRR

— OMRR + Heq

— OMAR + equalsize

— OMRR + equalsize

— OMRR + equalsize + Heq

Dans l'optique de ne faire qu'un seul et unique pipeline python, les caractéristiques ont été recalculées grâce à pyradiomics [68], un package python qui permet d'extraire de façon rapide toutes les caractéristiques. Cette partie a été réalisée dans un premier temps par deux étudiants en cycle ingénieur de l'école Centrale Nantes, puis repris lors du stage car les résultats obtenus était très différents de la vérité terrain.

## 5.3 L'étude de la survie par random survival forest

Les caractéristiques images ainsi obtenues vont être utilisées en association avec des caractéristiques cliniques pour déterminer la survie des patients, en entrée d'un modèle statistique. Le modèle choisi dans ce projet est celui des Random Survival Forest, une méthode d'arbres relativement récente. Les objectifs du projet sont de réécrire son algorithme en python afin de réaliser un pipeline automatique, ainsi que déterminer la valeur prédictive de notre base de données contenant ces caractéristiques images

### 5.3.1 Les méthodes d'arbres

Un arbre de décision permet de traiter la régression, de la classification bi-classe ou multi-classe ou encore de mélanger des variables explicatives quantitatives et qualitatives. La méthode des arbres de décision est connue depuis les années 60 mais ont connut leur apogée dans les années 80,avec les arbres CART (Classification And Regression Trees) de Leo Breiman qui permettent une large applicabilité, une facilité d'interprétation et des garanties théoriques. Les arbres CART ont cependant un problème d'instabilité. En effet, de petites modifications dans l'échantillon d'apprentissage peuvent avoir des effets importants sur la prédiction. La solution fut d'utiliser des forêts et la perturbation aléatoire des arbres. De nombreux algorithmes furent proposés comme le bagging

( [19], 1996) et Addaboost( [20], 1997). Mais ce sont les forêts aléatoires (Random Forest ou RF) de Breiman ( [21], 2001) qui se montrent encore aujourd'hui les plus performantes sur le plan expérimental, et qui sont de plus en plus utilisées.

La méthode des arbres font partie de la catégorie des méthodes d'apprentissage automatique dites supervisées. C'est à dire qu'il faut au préalable entraîner le modèle avec des échantillons étiquetés, afin de pouvoir réaliser le test sur des échantillons non étiquetés et prédire leur sortie.

**Les arbres CART** :

Le principe général de CART est de partitionner récursivement l'espace d'entrée X de façon binaire (X étant une matrice de dimensions (Np*Nc) avec le Np le nombre d'individus et Nc le nombre de variables), puis de déterminer une sous-partition optimale pour la prédiction. Bâtir un arbre CART se fait en deux étapes. Une première phase est la construction d'un arbre maximal, qui permet de définir la famille de modèles à l'intérieur de laquelle on cherchera à sélectionner le meilleur. L'arbre se construit en commençant par partitionner dans deux noeuds fils, l'entrée X, en fonction d'une variable et d'une valeur choisies. Le choix de la variable et de la valeur de la séparation est faite soit dans le but de diminuer la variance des nœuds obtenus pour la régression, soit en cherchant à diminuer la fonction de pureté de Gini, et donc à augmenter l'homogénéité des nœuds obtenus, pour la classification. La seconde phase, dite d'élagage, construit une suite de sous-arbres optimaux élagués de l'arbre maximal et qui comprend la racine. CART permet une bonne gestion des données manquantes et une bonne interprétabilité. Un autre avantage est la résistance naturelle aux valeurs aberrantes, la méthode étant purement non paramétrique, la présence d'une donnée aberrante dans l'ensemble d'apprentissage va contaminer essentiellement la feuille qui la contient, avec un faible impact pour les autres [voir l'article de R. Genuer et J. M. Poggi [69] pour plus de détails].

**Le bagging** :

Le Bagging est une méthode introduite par Breiman ( [19], 1996) pour les arbres, et directement issue de la remarque selon laquelle les arbres CART sont instables et sensibles aux fluctuations de l'ensemble des données de l'échantillon d'apprentissage Ln. Le principe du Bagging est de tirer un grand nombre d'échantillons, indépendamment les uns des autres, et de construire, en appliquant à chacun d'eux la règle de base, un grand nombre de prédicteurs. La collection de prédicteurs est alors agrégée en faisant simplement une moyenne ou un vote majoritaire (agrégation) comme présenté dans la figure 5.5 [69].

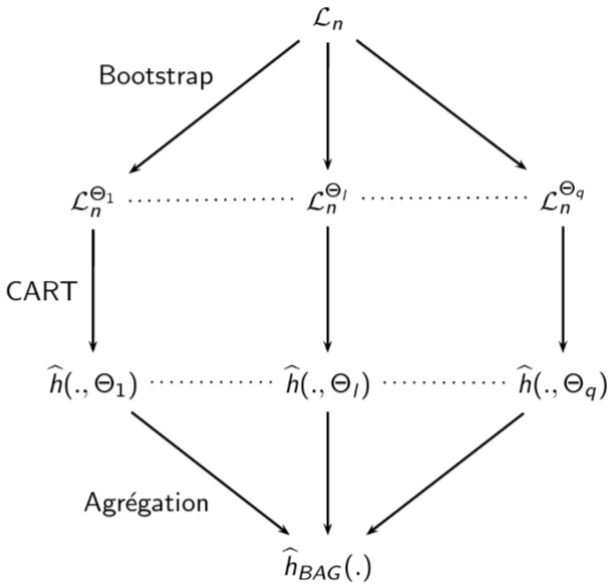



FIGURE 5.5 – Schéma du bagging avec pour règle de base un arbre CART [69]

**Le boosting** :

Le principe du Boosting est de tirer un premier échantillon bootstrap, où chaque observation a une probabilité 1/Np d'être tirée (avec Np le nombre d'échantillons), puis d'appliquer la règle de base pour obtenir un premier prédicteur. Ensuite, l'erreur de ce prédicteur sur l'échantillon d'apprentissage est calculée. Un deuxième échantillon boostrap est alors tiré mais la loi du tirage des observations n'est maintenant plus uniforme. La probabilité pour une observation d'être tirée dépend de la prédiction du premier estimateur sur cette observation. Le principe est, par le biais d'une mise à jour exponentielle bien choisie, d'augmenter la probabilité de tirer une observation mal prédite et de diminuer celle de tirer une observation bien prédite. Une fois le nouvel échantillon obtenu, on applique à nouveau la règle de base du deuxième prédicteur. On tire alors un troisième échantillon, qui dépend des prédictions du prédicteur 2 sur le set d'entraînement et ainsi de suite. La collection de prédicteurs obtenus est alors agrégée en faisant une moyenne pondérée, là encore via des poids exponentiels bien choisis. L'idée du Boosting est de se concentrer de plus en plus sur les observations mal prédites par la règle de base, pour essayer d'apprendre au mieux cette partie difficile de l'échantillon en vue d'améliorer les performances globales.

**Les forêts aléatoires** :

Les forêts aléatoires ont été introduites par Breiman ( [21], 2001) par la définition très générale suivante : Soit X une matrice de dimension (Np*Nc), contenant pour chaque patient les valeurs associées aux variables $\theta$, $(T(.,\theta_1),...,T(.,\theta_q))$ une collection de q prédicteurs, avec $\theta_1,...,\theta_q$ des vecteurs de variables aléatoires (indépendantes de l'échantillon d'apprentissage Ln), et Y un vecteur d'entrée (par exemple, la classe à laquelle appartiennent les individus). Le prédicteur des forêts aléatoires $T_{rf}$ est obtenu en agrégeant cette collection d'arbres aléatoires de la façon suivante :

- $T_{rf}(Y) = \frac{1}{q}\sum_{l=1}^{q} T(Y,\theta_l)$ (Moyenne des prédictions individuelles des arbres en régression
- $T_{rf}(Y) = argmax_{1\leq k\leq K}\sum_{l=1}^{q} 1_{T(Y,\theta_l)=k}$ (Vote majoritaire parmi les prédictions individuelles des arbres) en classification.

Breiman, dans son article de 2001, définit les forêts aléatoires comme ci-dessus et sont donc pour lui une famille de méthodes. Or, dans le même article, il présente un cas particulier de forêts aléatoires, appelées Random Forests-RI (random input), qu'il a implémentées. Par la suite, ce sont ces Random Forests-RI qui ont été quasi-systématiquement utilisées dans d'innombrables applications. Finalement, la dénomination "forêts aléatoires" désigne maintenant très souvent les Random Forests-RI.

Le principe de leur construction est tout d'abord de générer plusieurs échantillons bootstrap. Ensuite, sur chaque échantillon, une variante de CART est appliquée. Plus précisément, un arbre est, ici, construit de la façon suivante :

- Pour chaque arbre T :
  - pour chaque noeud k :
    1. construire un vecteur $\theta$ contenant m variables sélectionnées aléatoirement
    2. choisir la meilleur coupure parmi les variables de $\theta$
    3. réaliser une partition de l'échantillon en deux noeuds fils suivant la meilleur coupure.
    4. pour chaque noeud fils :
       - reprendre les étapes 1 à 4.
       - Arrêter lorsque le critère d'arrêt n'est pas plus respecté (par exemple, il peut-être de nécessiter un nombre minimal d'échantillons dans le noeud pour pouvoir réaliser la séparation, la profondeur maximale etc ...). Lorsqu'un noeud ne présente pas de noeuds fils, il est appelé feuille.

— Agrégation des arbres T (moyenne en régression et vote majoritaire en classification) pour donner le prédicteur Random Forests-RI.

Ainsi, les Random Forests-RI peuvent être vues comme une variante du Bagging, où la différence intervient dans la construction des arbres individuels. Le tirage, à chaque nœud, des m variables se fait, et uniformément parmi toutes les variables. L'algorithme est résumé dans la figure 5.6.

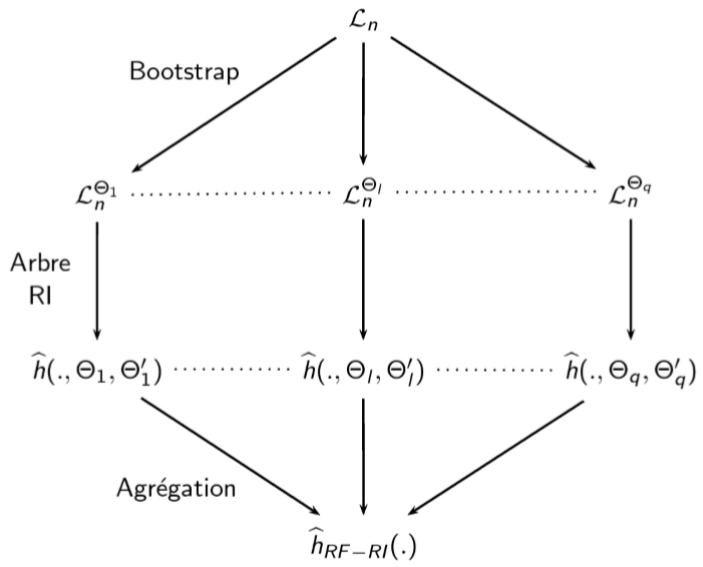



FIGURE 5.6 – Schéma des forêts aléatoires RF-RI. Parmi les échantillons Ln, des sous échantillons sont sélectionnés aléatoirement pour construire des prédicteurs $\hat{h}(.,\Theta_l,\Theta_l')$. Les prédicteurs sont ensuite agrégés pour donner$\hat{h}_{RF-RI}(.)$

En pratique, les Random Forests-RI améliorent les performances du Bagging. Cette méthode a été reprise pour être appliquée à différents domaines. Pour l'analyse des données de survie ce sont principalement Hothorn et al. ( [26], 2006) et Ishwaran et al. ( [1], 2008).

**Survival trees and ensembles méthodes** :

La première adaptation des arbres CART aux données censurées fut développée par Gordon and Olshen (1985). Depuis, une dizaine d'algorithmes d'arbres de survie ont été proposée et seulement quelques-uns ont été implémentés dans les logiciels publiques.

### 5.3.2 La méthode de RSF

La méthode de Random Survival Forest est une méthode de forêts aléatoires pour l'analyse de la survie avec des données censurées à droite, présentée par Ishwaran en 2008 [1].

Une forêt aléatoire est une méthode d'apprentissage sur de multiples arbres de décision qui modélisent une hiérarchie de tests sur les valeurs d'un ensemble de variables. La différence entre la méthode des forêts aléatoires standard et les RSF est dans le fait que chaque aspect de la construction des RSF prend en compte la sortie (survie, censure).

On peut décomposer la méthode en deux parties : L'entraînement et le test.

— **L'entraînement**

L'entraînement correspond à la création des arbres. Soit une base de donnée X qui correspond à un tableau de données qui comprend pour chaque individu i un vecteur de caractéristiques $\theta_i$ de dimension $N_c$ et égal à $(\theta_{i1}, ..., \theta_{iN_c}$. A cette base de donnée est associée pour chaque individu i un couple de valeurs $(\tau_i, \sigma_i)$ qu'on

appellera $Y_i$ ($\tau_i$ correspond au temps jusqu'à l'évènement et $\sigma_i$ la censure qui est égale à 0 si l'évènement n'a pas eu lieu et 1 sinon).

Pour chaque arbre, un échantillon d'individus est pris aléatoirement dans la base de donnée X et constitue l'in-bag $X_{in}$ (Le reste étant placé dans l'Out-Of-Bag (OOB) $X_{oob}$ et pouvant être utilisé pour le calcul de l'erreur de prédiction et de l'importance des variables).

A chaque nœud h, un sous-ensemble des caractéristiques est sélectionné aléatoirement parmi toutes les caractéristiques $\theta$. La meilleure caractéristique est choisie dans ce sous-ensemble. Elle correspond à celle qui maximise la différence de survie entre les noeuds fils. Ishwaran présente 4 méthodes de sélection de ces caractéristiques mais deux sont gardées pour ce travail :

— Le log Rank : pour chaque caractéristique $\theta$ du sous-ensemble $X_{in}$, et pour différentes valeurs $c$ de cette caractéristique , on calcule la valeur du test du logRank pour la séparation du groupe de patients en deux par la caractéristique donnée avec le seuil à la valeur donnée. Le log rank se calcule sur les données de survie des patients. La meilleure caractéristique sera celle qui aura le résultat de log rank le plus haut et la séparation dans les nœuds fils se fera au seuil qui donne le plus haut résultat de logrank. Définissons que lorsque $\theta_{ij} \leq c$ l'individu i est inscrit dans le noeud fils gauche, et dans le noeud fils droit sinon.
Soit $t_1 < ... < t_m$ les temps distincts des évènements dans le noeud h, $d_{k,l}$ et $Y_{k,l}$ respectivement le nombre d'évènements et le nombre d'individus à risque au temps $t_k$ dans le noeud fils gauche ($d_{k,r}$ et $Y_{k,r}$ pour le fils droit), $Y_{k,s}$ le nombre d'individus dans les noeuds fils (avec $s \in \{l, r\}$) qui n'avait pas eu d'évènement à pas d'évènement à $t_{k-1}$. Définissons $Y_k$ = $Y_{k,l}$ + $Y_{k,r}$ et $d_k$ = $d_{k,l}$ + $d_{k,r}$. Appelons $n_s$ le nombre total d'individus dans le noeud fils s, et $n = n_l + n_r$. La valeur du test logRank pour la variable $\tau$ et la valeur c est :

$$L(\theta, c) = \frac{\sum_{k=1}^{m}(d_{k,l} - Y_{k,l}\frac{d_k}{Y_k}}{\sqrt{\sum_{k=1}^{m}\frac{Y_{k,l}}{Y_k}(1 - \frac{Y_{k,l}}{Y_k})\frac{Y_k - d_k}{Y_k - 1})d_k}} \tag{5.7}$$

La valeur absolue de absolue $L(\theta, c)$ mesure la séparation. Plus elle est élevée et meilleure est la différence entre les deux groupes

— Le logrank random : cette méthode est équivalente au logrank mais il n'y a, ici, pas de test sur différentes valeurs. Seule une valeur aléatoire est prise par caractéristique $\tau$.

La construction de l'arbre se construit jusqu'à ce que le critère arrêt ne soit plus respecté [voir Section 5.3.1 les forêts aléatoires]. Dans les feuilles seront calculés les valeurs désirées.
Cela peut être comme ici la mortalité, ou la probabilité de survie, des courbes de survie, un temps ou une classe. Dans son article, Ishwaran décrit la mortalité comme :

Chaque feuille doit normalement contenir un groupe de patients homogène du point de vue de la survie. Elle est calculée grâce à l'estimateur du CHF correspondant à l'estimateur de Nelson-Aalen [Equation 5.8].

$$\hat{H}_h(t) = \sum_{t_{k,h} \leq t} \frac{d_{k,h}}{n_{k,h}} \tag{5.8}$$

Avec $\hat{H}_h(t)$ le CHF au nœud h et temps t, $d_{k,h}$ le nombre de morts au temps $t_{k,h}$ et $Y_{k,h}$ le nombre d'individus à risque au temps $t_{k,h}$. Dans chaque nœud h, tous les individus i ont le même CHF. La mortalité $M_i$ d'un individu i dans le noeud h, correspond à la somme des CHF sur chaque temps unique des données.

$$M_i = \sum_{k=1}^{m} \hat{H}_h(t_k | X_i) \tag{5.9}$$

Plus la mortalité est élevée par rapport aux autres individus, plus le risque d'évènement est grand (l'évènement peut-être un décès, une rechute ou les deux dans le cas de la PFS (Progression-Free Survival) par exemple).

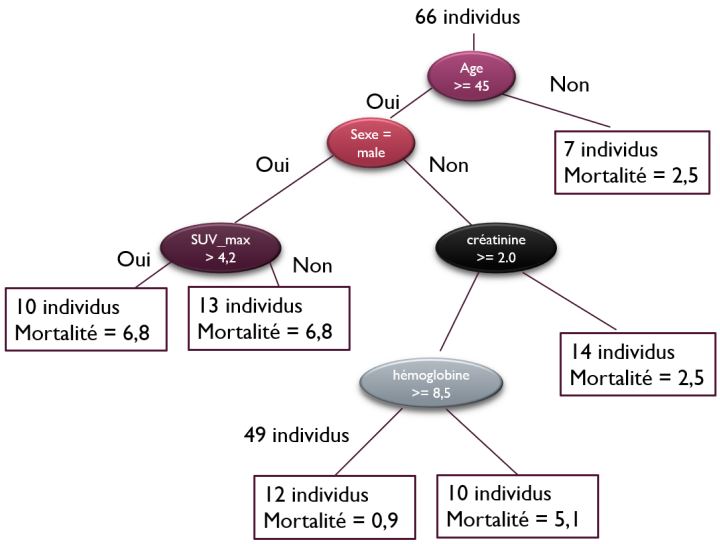



FIGURE 5.7 – Exemple d'arbre, où le critère d'arrêt est d'avoir un minimum de 15 individus dans un nœud pour réaliser une séparation. Profondeur de l'arbre : 4.

Une fois tous les arbres créés, une forêt est créée.

— **Le test**

Pour tester la forêt, pour chaque arbre, on prédit la mortalité de chaque individu de l'OOB $X_{oob}$. Si ce qui est recherché est la mortalité, la mortalité associée avec la feuille dans laquelle se trouvera finalement l'individu, sera la mortalité prédite de cet individu. Il aura ainsi, une mortalité prédite par arbre. La mortalité finale correspond à la moyenne des mortalités de chaque arbre.

**L'évaluation du modèle**

Pour évaluer l'arbre on utilise l'erreur de prédiction qui correspond à 1 moins l'indice de concordance (c-index). Le c-index indique le taux de paires possibles qui sont bien ordonnancées et est calculé comme suit :

1. Former toutes les paires possibles avec les données
2. Ne pas prendre en compte les paires dont le temps de survie (T) le plus petit est censuré. Ne pas prendre en compte les paires i et j si $T_i = T_j$, sauf si au moins l'un a un évènement. Appelons $N_p$ le nombre de paires admissibles.
3. Pour chaque paire admissible, où $T_i \neq T_i$, compte 1 si le temps de survie le plus court a la pire prédiction. Compte 0,5 si les prédictions sont équivalentes. Pour chaque paire admissible, où $T_i = T_j$ et les deux ont un évènement, compter 1 si les prédictions sont équivalentes ; sinon compte 0,5. Pour chaque paire admissible, où $T_i = T_j$ mais les deux n'ont pas d'évènements, compte 1 si celui ayant un évènement a une pire prédiction, sinon compte 0,5.
4. La concordance correspond à la somme sur toutes les paires admissibles.
5. Le C-index, C, est définit par C = $\frac{Concordance}{N_p}$

Plus l'erreur de prédiction $err_{oob}$ sera faible et meilleur sera le résultat

**L'importance des variables**

L'article de Ishwaran et al. [1] donne aussi une méthode de calcul de l'importance des variables (VIMP). Son calcul consiste à :

1. pour chaque caractéristique $\theta$ :

a) Reconstruire les arbres utilisant la caractéristique $\theta$ à l'aide de l'OOB. Lorsque un noeud h utilise $\theta$ assigner les individus du noeud h aléatoirement dans les noeuds fils. Recalculer l'erreur de prédiction $err_{vimp}$ associée à la nouvelle forêt.

b) VIMP = $err_{vimp}$ - $err_{oob}$

Plus la valeur du VIMP est grande, plus la caractéristique a une valeur prédictive.

**L'algorithme des données manquantes**

Il est aussi présenté un algorithme des données manquantes « adaptative tree imputation ». En effet, il existe une stratégie qui permet d'enlever grossièrement les données manquantes : les valeurs manquantes pour les données continues sont remplacées par la médiane des valeurs non manquantes, et les variables catégorielles sont remplacées par la valeur la plus fréquente. Ensuite, après utilisation de ces données dans un RF, elles sont remises à jour pour être réutilisées dans un RF suivant, et ceci plusieurs fois de suite. Cependant il réside des problèmes comme un biais sur l'erreur de prédiction de OOB et sur le VIMP. De plus, la forêt ne peut pas être réutilisée sur un set de tests avec des valeurs manquantes. La solution présentée dans l'article est l'« adaptative tree imputation ». L'idée est de remplacer les valeurs manquantes au fur et à mesure de la construction de l'arbre. Les étapes sont les suivantes :

1. On récupère aléatoirement un set de données de l'in-bag n'ayant pas de valeurs manquantes.

2. Pour chaque cas de h avec une valeur manquante de la variable k, on remplace en utilisant une valeur aléatoire de la fonction de distribution des valeurs non manquantes pour la variable k dans le sac au nœud h, et ceci pour chaque k.

3. A chaque nœud fils, les valeurs qui était manquantes sont remises à jour en appliquant l'étape 2.

4. Afin d'éviter que la précision ne soit affectée lorsque le nombre de données manquantes augmente, on réalise des itérations sur l'algorithme des données manquantes.

### 5.3.3 Contribution

Après avoir récupéré et nettoyé la base de données, en ne gardant que les patients aux données complètes, la méthode de Random Survival Forest a été écrite en langage python en partant du papier de Ishwaran et al. [1]. Le but est de pouvoir avoir un pipeline entièrement en python mais aussi de pouvoir, plus tard appliquer les modification désirées afin d'améliorer l'algorithme. Certaines méthodes n'ont pas encore été écrites par manque de temps tel que l'algorithme des données manquantes qui permet de compléter les données et prendre en compte tous les patients, ou les méthodes de séparation logRank score rules et logRank splitting rules. La méthode de calcul de l'importance des variables VIMP a elle été implémentée.

Le modèle donne finalement pour chaque patient testé une mortalité prédite et une courbe de survie (La figure 5.8 donne un exemple de la mortalité prédite de 3 patients en fonction du temps obtenue en python)

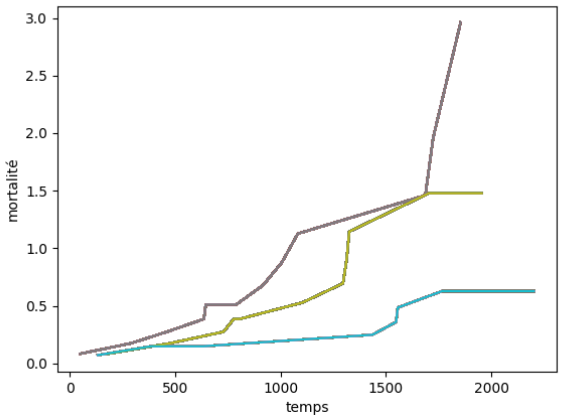



FIGURE 5.8 – Exemple de courbes de mortalité prédites. Une courbe correspond à un patient.

Une fois le modèle optimisé, les mortalités obtenues sont utilisées pour séparer les patients en deux groupes : Bon pronostic et mauvais pronostic. En effet, à partir de ces mortalités prédites est déterminée la meilleure des séparations par log rank. Cela correspondra à la séparation sur la mortalité prédite présentant la plus petite valeur de p-value. Cette p-value permet d'évaluer la séparation. La p-value est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée. D'après Ronald Fisher on ne peut jamais accepter l'hypothèse nulle mais on peut la rejeter. On considère généralement les valeurs suivantes :

— $p \leq 0.01$ : très forte présomption contre l'hypothèse nulle
— $0{,}01 < p \leq 0{,}05$ : forte présomption contre l'hypothèse nulle
— $0{,}05 < p \leq 0{,}1$ : faible présomption contre l'hypothèse nulle
— $p > 0{,}1$ : pas de présomption contre l'hypothèse nulle

Ainsi, plus la valeur est faible et plus la séparation semble non aléatoire et une valeur trop forte de p-value indique une séparation qui semble aléatoire.
A partir de cela on peut réaliser un Kaplan-Meier afin d'obtenir deux courbes [voir section 3.1.1.1.1.] grâce aux valeurs réelles des temps et évènements.
L'importance des variables est aussi calculée à partir de ce modèle. (ajouter ex ou illustration)

*Chapitre 6*

# Détails d'implémentation

Ce chapitre décrit en détails ce qui a été utilisé (Données, logiciels etc ...) et les expériences réalisées, mais surtout comment cela a été validé, aussi bien du point de vue du pre-traitement, que de l'extraction des caractéristiques texturales ou de la RSF.

## 6.1 Les données

Pour l'entraînement et le test des Random Survival Foret, sont disponibles 70 patients avec des caractéristiques texturales issues d'images TEP, ainsi que des caractéristiques cliniques (au nombre de 13) et volumiques (au nombre de 11) et les données de survie et d'évènement.

Parmi les caractéristiques texturales :

— une caractéristique de premier ordre
— 6 caractéristiques provenant de la matrice GLCM (Gray Level Coocurrence Matrix)
— 5 provenant de la matrice GLRLM (Gray Level Run Length Matrix)
— 7 provenant de la matrice GLSZM (Gray Level Size Zone Matrix)

La liste détaillée de ces caractéristiques texturales se trouve en annexe [Annexe A.3], au même titre que celle des caractéristiques cliniques [Annexe A.1]
Des caractéristiques volumiques sont aussi disponibles [Annexe A.2]. Elles correspondent par exemple, au volume de la lésion la plus fixante ou au volume total de toutes les lésions identifiées.
Cela fait un total de 135 caractéristiques pour chaque patient qui correspondent aux données du début de la maladie.

Outre ces caractéristiques sont utilisées les données de survie des patients. Ces données sont disponibles pour quatre évènements :

— L'OS ou Overall Survival correspondant au décès
— La PFS ou Progression-free survival correspondant au temps avant la rechute ou le décès si le patient n'a pas rechuté avant son décès.
— La PFS à 1 an
— La PFS à 2 ans

Pour chaque évènement est disponible un index indiquant si l'évènement a eu lieu et le temps jusqu'à l'évènement si celui a eu lieu, ou jusqu'au dernier suivi du patient sinon (on parle à ce moment-là de censure à droite). L'évènement qui nous intéresse ici est la PSF. L'OS pourra aussi être étudiée dans un second temps.
Les données sont enregistrées sur 7 ans (au maximum) et sont utilisée dans un essai clinique sur un traitement

contre le myélome multiple [6].
Cependant parmi les 70 patients dont les calculs sur les images ont été réalisés, certains ont des données manquantes. Certaines données peuvent être complétées mais d'autres non. Ainsi, 5 patients ne seront pas pris en compte car leurs données sont incomplètes. Il pourra être intéressant dans un second temps d'essayer de résoudre ce problème.
Il y a aussi 64 patients dont nous possédons seulement les données cliniques qui pourraient peut-être être inclus grâce à un algorithme des données manquantes.
De plus, un modèle sera aussi réalisé avec seulement les données cliniques et volumiques, ce qui permettra d'obtenir 89 patients.
Une autre base de données est aussi utilisée, la base Vétéran, qui correspond aux données d'une étude clinique randomisée sur deux traitements contre le cancer du poumon chez des vétérans. C'est une base de données standard dans l'analyse de la survie et notamment dans l'article de Iswharan et al. [1] pour montrer l'efficacité des RSF.
Les taux de censure des bases de données sont :

— Myélome multiple : 32,8% de censure
— Vétéran : 6,6% de censure

Pour la segmentation et l'extraction des caractéristiques 134 images TEP en baseline sont disponibles sous format DICOM (ainsi que les images CT correspondantes), ainsi que des boîtes au format DICOM correspondant à une première segmentation grossière de la zone où se trouve la lésion par le logiciel Dosisoft.

## 6.2 Les logiciels utilisés

### 6.2.1 Le langage python

Le but final du projet est d'avoir un pipeline entier en partant des images et en arrivant aux résultats de survie. Il est donc préférable de réaliser chaque étape dans un seul et même langage. Le choix s'est porté sur le langage Python car il possède de bons packages pour l'apprentissage automatique, l'extraction de caractéristiques et l'analyse d'image. Il est aussi le langage d'outils informatiques d'apprentissage automatique tels que TensorFlow ou Pytorch qui peuvent être utilisés pour l'apprentissage profond. En effet, pour le moment, chaque étape pour arriver des images aux caractéristiques, en passant par la segmentation des lésions, est réalisée sur différents logiciels, par différentes personnes, à différents moments. Cela est chronophage et une source d'erreur plus grande.
En plus des packages usuels tels que *numpy* ou *pandas*, les packages utilisés sont :

— *Lifelines* pour le calcul du test du logRank et la réalisation du Kaplan Meier
— *Scikit learn* pour le cross test validation et les classifieurs forêts aléatoires
— *Ezodf* pour l'ouverture du fichier Excel

Pour la segmentation par vote majoritaire et l'extraction des caractéristiques sont utilisés :

— *Pyradiomics* pour l'extraction des caractéristiques
— *Pydicom* et *dicom numpy* pour la manipulation des images médicales
— *Simple ITK* pour la manipulation des images médicales
— *Scikit image* pour la manipulation des images
— *Scilit learn* pour le calcul de kmean et de la labellisation

### 6.2.2 Le langage R

La méthode des RSF de Ishwaran n'est pour le moment trouvée qu'en R, dans le package *randomForestSRC*. Dans le but de réaliser tout le pipeline dans un unique langage, la méthode RSF a été réécrite en python. La méthode RSF en R a tout de même été utilisée afin de comparer les résultats des deux algorithmes.
Le package *survival* est aussi utilisé et comprend notamment la base de données *veteran*.
Les packages *ggfortify* et *ggplot2* ont été utilisés afin de réaliser des courbes de Kaplan Meier.

## 6.3 Validation expérimentale

### 6.3.1 La segmentation

Pour le moment seul les semgentations 40% et 2.5 on été réalisées (Le k-means n'a pas encore été réalisé). Le vote majoritaire a aussi été réalisé mais afin de valider nos résultats sans prendre en compte le k-means nous comparons d'abord les segmentations 2.5 et 40% avec la vérité terrain (Planet, Dosisoft, Cachan, France).

### 6.3.2 Les caractéristiques texturales

La validation des caractéristiques s'est faite grâce à l'ISBI. L'ISBI ou Image biomarker standardisation initiative [67], est une collaboration internationale qui travaille à la standardisation des biomarqueurs extraits d'images. Ainsi ils utilisent une base de données de référence et fournissent les valeurs de biomarqueurs associés afin de vérifier la validité des caractéristiques calculées. La vérification peut se faire avec des matrices fantômes ou des images provenant de leur base de données avec différents paramètres de calculs. La matrice fantôme est une matrice créée de façon artificielle pour simuler une image médicale (voir Figure 6.1 pour exemple). C'est une matrice 3D qui a ici été utilisée afin de calculer les biomarqueurs désirés :

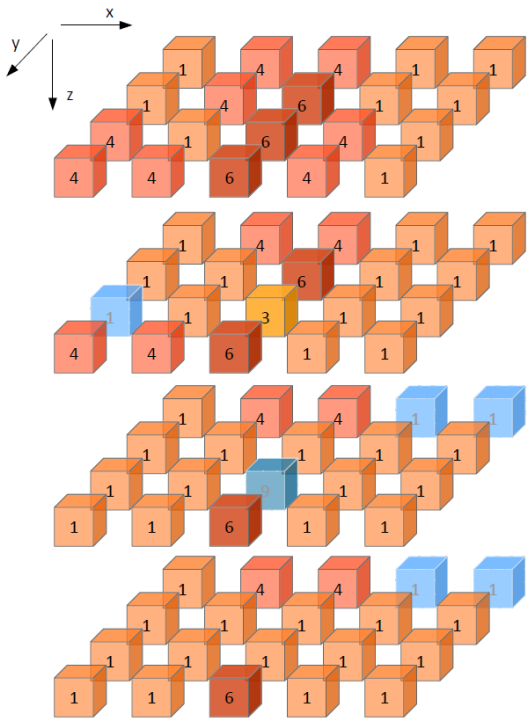



FIGURE 6.1 – Matrice fantôme utilisée pour la vérification des caractéristiques

La figure 6.2 donne un exemple de présentation des valeurs des biomarqueurs de la matrice GLCM :

| feature | dig. phantom | config. C | config. D | config. E |
|---|---|---|---|---|
| joint variance | 3.1 | $73.7 \pm 1.9$ | $17.6 \pm 0.3$ | 10.9 |
| joint entropy | 2.4 | $6.39 \pm 0.05$ | $4.95 \pm 0.02$ | $5.99 \pm 0.02$ |
| difference average | 1.43 | $2.17 \pm 0.04$ | $1.29 \pm 0.01$ | 1.83 |
| difference variance | 3.06 | $14.4 \pm 0.4$ | $5.37 \pm 0.11$ | $7.37 \pm 0.03$ |
| difference entropy | 1.56 | $2.64 \pm 0.02$ | 2.13 | 2.49 |
| sum average | 4.29 | $77.9 \pm 0.2$ | $37.7 \pm 0.8$ | $36.5 \pm 0.3$ |
| sum variance | 7.07 | $276 \pm 7$ | $63.4 \pm 1.3$ | $32.7 \pm 0.3$ |
| sum entropy | 1.92 | $4.56 \pm 0.03$ | $3.68 \pm 0.01$ | 4.03 |
| angular second moment | 0.303 | $(4.5 \pm 0.08) \cdot 10^{-2}$ | $0.11 \pm 0.002$ | $(3.43 \pm 0.09) \cdot 10^{-2}$ |
| contrast | 5.32 | $19.2 \pm 0.6$ | $7.07 \pm 0.14$ | 10.7 |
| dissimilarity | 1.43 | $2.17 \pm 0.04$ | $1.29 \pm 0.01$ | 1.83 |
| inverse difference | 0.677 | $0.583 \pm 0.003$ | $0.682 \pm 0.002$ | $0.548 \pm 0.002$ |
| inverse difference normalised | 0.851 | 0.966 | 0.965 | 0.951 |
| inverse difference moment | 0.618 | $0.548 \pm 0.003$ | $0.656 \pm 0.002$ | $0.505 \pm 0.003$ |
| inverse difference moment normalised | 0.898 | 0.994 | 0.994 | 0.991 |
| inverse variance | $6.04 \cdot 10^{-2}$ | $0.39 \pm 0.002$ | $0.341 \pm 0.004$ | 0.44 |
| correlation | 0.157 | 0.869 | $0.798 \pm 0.004$ | $0.503 \pm 0.003$ |
| autocorrelation | 5.06 | $1.58 \cdot 10^{3}$ | $370 \pm 16$ | $338 \pm 6$ |
| cluster tendency | 7.07 | $276 \pm 7$ | $63.4 \pm 1.3$ | $32.7 \pm 0.3$ |
| cluster shade | 16.6 | $(-1.06 \pm 0.02) \cdot 10^{4}$ | $(-1.27 \pm 0.04) \cdot 10^{3}$ | $-442 \pm 7$ |
| cluster prominence | 145 | $(5.69 \pm 0.1) \cdot 10^{5}$ | $(3.57 \pm 0.19) \cdot 10^{4}$ | $(1.15 \pm 0.01) \cdot 10^{4}$ |
| information correlation 1 | -0.157 | -0.236 | $-0.231 \pm 0.002$ | -0.115 |
| information correlation 2 | 0.52 | 0.9 | $0.845 \pm 0.002$ | 0.71 |

FIGURE 6.2 – Exemple de présentation des valeurs de biomarqueurs GLCM pour un calcul sur image 3D avec moyenne sur une matrice 3D. "Dig. phantom" correspond aux calculs réalisés sur le fantôme. Les configurations C, D et E, aux calculs faits sur des images cliniques avec 3 configurations différentes.

L'implémentation réalisée en python pour l'extraction des caractéristiques est donc vérifiée dans un premier temps en étant appliquée au fantôme de l'ISBI présent dans la figure 6.1 et comparée avec les résultats obtenus. Ensuite les résultats obtenus sur nos images TEP sont comparés avec les résultats de la vérité terrain (calculés en C par T. Carlier). Les calculs se font avec les paramètres suivants : une matrice 3D avec échantillonage absolu, et pas d'égalisation ou de normalisation de la taille des pixels (OMAR).

### 6.3.3 L'analyse de la survie

Le travail sur l'analyse de la survie comprend l'écriture de l'algorithme des RSF en python et donc sa validation en comparant à l'algorithme en R, l'optimisation des hyperparamètres, l'étude de la valeur prédictive de notre base de données au niveau de la prédiction de la rechute et de la détermination de groupes de bon et mauvais pronostic, et la détermination des caractéristiques ayant valeur prédictive dans la survie. Les étapes pour l'étude de la survie sont résumées dans la figure 6.3.

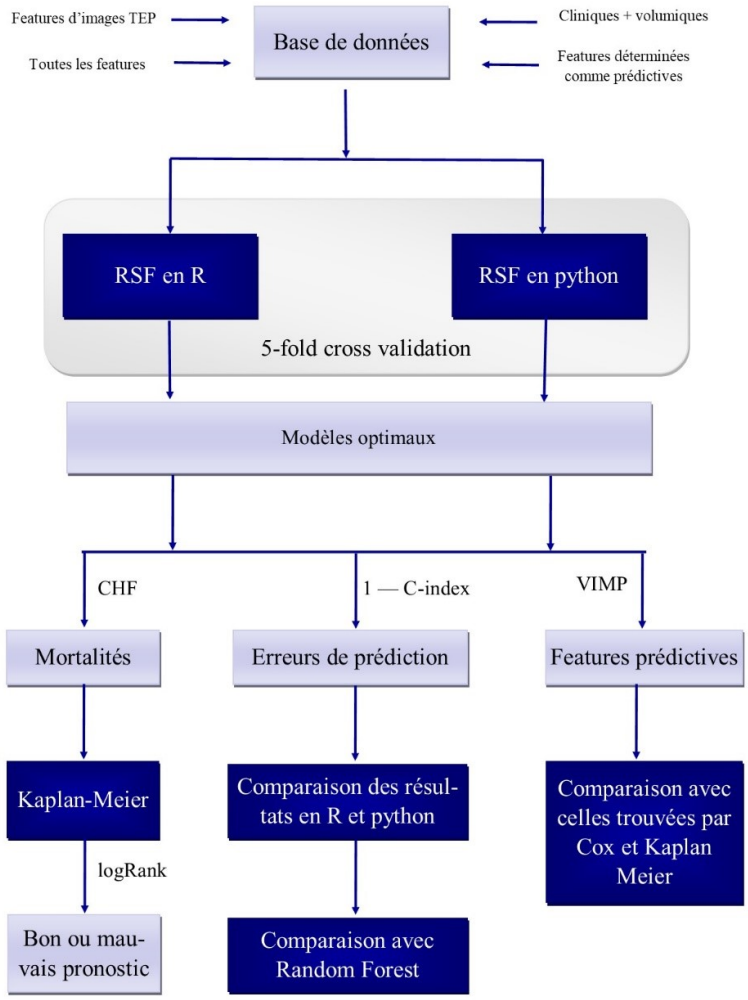



FIGURE 6.3 – Pipeline d'étude de la survie

#### 6.3.3.1 L'implémentation en python

Après écriture de l'algorithme des RSF, le modèle est ensuite optimisé pour chaque base de données utilisée afin d'en tirer les meilleures performances. Deux méthodes se sont naturellement présentées. L'utilisation de l'erreur de prédiction calculée sur l'Out-Of-Bag, comme présenté dans l'article [1], ou de la validation croisée (k-fold cross validation) avec k égal à 5. C'est la seconde méthode qui fut choisie car elle permet d'avoir les performances globales et non pas de chaque arbre. Pour ce faire la série de données est divisée en 5 sous-ensembles aléatoirement. Pour chaque sous-ensemble j, on réalise l'entraînement sur les 4 sous-ensembles qui diffèrent de j, et le test sur le sous-ensemble j. L'erreur de prédiction est calculée pour chaque test. Elle correspond à l'index de concordance (c-index) soustrait de un. Le c-index estime la probabilité que, dans une paire de cas sélectionnés aléatoirement, le cas qui échoue en premier ait la pire prédiction. [Calcul du C-index dans la section 5.3.2]

L'erreur de prédiction pour chaque sous-ensemble est ensuite moyennée pour donner une unique erreur de prédiction par validation croisée. On réalise un validation croisée avec différents paramètres, afin d'optimiser le modèle. Les paramètres utilisés sont :

— n_trees : le nombre d'arbres

— min_sample_split : le nombre minimal d'individus qu'il faut dans un noeud pour réaliser une séparation

— mode : séparation par LogRank ou LogRank random
— max_f : le nombre maximal de caractéristiques tirées par noeud

#### 6.3.3.2 Validation de l'implémentation par comparaison avec R

L'implémentation en python de la méthode RSF résulte du fait qu'elle n'est pour l'instant implémentée que en R. Or il est nécessaire de l'écrire en python pour obtenir un pipeline sur un seul langage. De plus, écrire l'algorithme en python permettra à l'avenir de le modifier plus facilement afin d'y apporter des améliorations. Cependant, une comparaison avec l'algorithme déjà présent en R est nécessaire pour vérifier sa véracité. La comparaison des algorithmes en R et en Python se fait sur notre base de données du myélome multiple ainsi que sur une base connue (veteran) présente dans R. Cela permet notamment de vérifier que la différence de performances ne soit pas dû à la base de données choisie.
Pour pouvoir faire une comparaison, il est nécessaire de traiter les deux algorithmes de la même façon. Ainsi, le calcul de l'erreur de prédiction se fait aussi en R par une validation croisée sur le set de test avec k égal à 5, et non sur OOB comme implémenté dans la fonction *rfsrc* du package *randomForestSRC* de R. Pour ce faire on utilise la fonction *estC* du package *CompareC* pour calculer le c-index.
La comparaison des deux algorithmes se fait sur la valeur de l'erreur prédiction du modèle optimisé.

#### 6.3.3.3 Optimisation du temps de calcul en python

Le temps de calcul en R avec la fonction rfsrc est beaucoup plus court que celui en python. Ainsi, pour pallier à ce problème plusieurs méthodes ont été testées.
Premièrement, le package joblib permet, par la fonction « parallel », d'utiliser au maximum les processeurs de la machine. La fonction a été testée afin de calculer en parallèle la meilleure caractéristique mais aussi de faire en parallèle la une validation croisée. Les deux ne pouvant pas être utilisés en même temps, c'est le calcul parallèle du « validation croisée » qui a été choisi car permettant un meilleur temps de calcul. Cython a aussi été testé, mais devant utiliser joblib avec la fonction "threading", cela ralentissait le calcul.

De plus, le calcul des arbres était de plus en plus long lorsque le nombre d'arbres à calculer était grand. La solution fut de pré-allouer la mémoire de la forêt. Enfin, afin d'éviter de tout recalculer, les calculs ont été faits avec un nombre d'arbres maximal (1000) et à paramètres équivalents (or nombre d'arbres) les calculs des forêts se faisait à partir d'arbres piochés dans les 1000 arbres calculés.

#### 6.3.3.4 Valeur prédictive de notre base de données et détermination de biomarqueurs

On voit dans le pipeline sur la figure 6.3 que quatre bases de données de myélome multiples sont utilisées :
— Avec toutes les caractéristiques
— Avec seulement les caractéristiques texturales
— Avec seulement les caractéristiques cliniques et volumiques
— Avec les caractéristiques trouvées comme prédictives dans la littérature

Cette dernière base de données a été constituée en se basant sur une recherche faite dans la littérature. Les caractéristiques considérées comme prédictives pour le myélome multiple ont été prises dans différents articles. En effet, des caractéristiques cliniques que nous possédons, l'article Amr Hanbali et al. [70] nous dit que le niveau d'hémoglobine, la calcémie et le nombre de lésions osseuses au scan ainsi que le R-ISS a un bon pouvoir de pronostic du stade de la maladie (L'hémoglobine et la calcémie sont utilisées pour les stades 1 et 3 et la créatinine pour les stades IIIA et B). Le calcium, l'hémoglobine et la créatinine sont retrouvés dans l'article de Bergsage et al. [71]. Le site www.cancer.ca [72] nous donne l'âge comme facteur pronostique en précisant que

les patients plus jeunes ont un meilleur pronostic.
L'article de C. Boder Milin [66] indique qu'un nombre de Lésions focales supérieur à 3 au diagnostic par FDG-TEP donne un pronostic plus faible pour la PFS et l'OS (confirmé par l'article de Bartel et al. [63]). Le nombre de lésions focales à l'IRM supérieur ou égal à 7 donne un moins bon pronostic en PSF mais n'est pas significatif en OS. De plus, avoir un SUV > 4,2, et la présence de lésions extra médullaires affecte aussi négativement la PFS et l'OS. Ainsi, les caractéristiques retenues comme pronostiques sont :

- SUV max des lésions focale osseuse en TEP
- Nombre de lésions focales
- Présence ou non de lésions extra ostéomédullaires
- Âge
- Nombre de lésions focales à l'IRM
- Hémoglobine
- Calcium
- Créatinine
- R-ISS

Ont été ajoutés :

- SUV osteo-medédullaire en TEP
- SUV lésions extra osteomédullaires en TEP
- Le traitement

Ces bases vont être utilisées dans un premier temps pour déterminer quelle est la base la plus prédictive, en prenant des bases avec le même nombre de patients (67 patients). C'est à dire que des validation croisées avec différents paramètres vont ête réalisés pour chaque base. Ensuite à partir de ces résultats, le modèle permettant d'obtenir la plus faible erreur de prédiction sera utilisé pour réaliser deux groupes pour chaque base : bon et mauvais pronostic.

La meilleure séparation entre ces deux groupes est déterminée à l'aide de la valeur du log rank test [voir calcul du logRank section 5.3.2] sur les mortalités prédites. Les courbes des deux groupes sont ensuite présentées à l'aide de kaplan-Meier avec un intervalle de confiance à 95%. Il est calculé en utilisant l'estimation de la variance par la formule de Greenwood.
En effet :

$$\sigma^2(S(t_j)) = S(t_j)^2 \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}$$

avec $n_i$ : le nombre d'observations restantes non censurées juste avant $t_i$ et $d_i$ : le nombre d'évènements observés à l'instant $t_i$
Grâce à cette variance, on obtient la formule de l'intervalle de confiance à 95% :

$$IC_{95\%}0 = S(t_j) \pm 1.96 * \sigma^2(S(t_j))$$

Dans un second temps, les bases seront utilisées avec tous les patients qu'elles peuvent posséder sans avoir de données manquantes afin de déterminer les caractéristiques les plus prédictives et voir l'influence du nombre de patients sur les résultats.
Le nombre de patients par base sont les suivants :

- Toutes les caractéristiques : 67 patients
- Caractéristiques prédictives : 89 patients
- Caractéristiques cliniques et volumiques : 89 patients
- Caractéristiques texturales : 70 patients

Pour chaque base, 50 ou 100 itérations (50 pour python et 100 pour R) sont réalisées et la valeur VIMP [voir partie 5.3.2] de chaque caractéristique ainsi que le TOP 10 (ou 5 pour la base des caractéristiques prédictives de la littérature car le nombre de caractéristiques n'est que de 14) sont récupérés. Les valeurs VIMP sont moyennées sur les 50 ou 100 itérations et la fréquence d'apparition dans le Top pour chaque caractéristique est calculée. Ceci permettra pour chaque base de données, de récupérer les 5 meilleures caractéristiques.

#### 6.3.3.5 Comparaison de la méthode RSF et autres méthodes de RF

Afin de savoir si la RSF obtient de bon résultats, il est nécessaire de la comparer avec les autres méthodes utilisées dans la littérature. Dans le contexte du stage, seuls les classifieurs RF ont été comparés. Par la suite la comparaison pourra être réalisée avec le modèle de Cox et les régresseurs Forêts aléatoires de Hothorn par exemple.

Le but du classifier RF est de classer les patients en deux groupes : mauvais et bon pronostic. Nous considérerons ici 4 cas :

— Un bon pronostic correspond à une survie supérieure à 2 ans.
— Un bon pronostic correspond à une survie supérieure à 3 ans.
— Un bon pronostic correspond à une survie supérieure à 4 ans.
— Un bon pronostic correspond à une survie supérieure à 5 ans.

pour déterminer le modèle optimisé, nous réaliserons une validation croisée. Ensuite une fois la classification réalisée, nous afficherons les courbes de Kaplan-Meier et appliquerons un test du logRank pour pouvoir comparer avec la séparation obtenue par RSF. La classification RF sera testée sur la base contenant toutes les caractéristiques et celles avec les caractéristiques cliniques et volumiques (les deux possédant un nombre de patients équivalent, 67). De plus, les patients ayant une censure avant la limite de séparation fixées (2, 3, 4 ou 5 ans), sont éliminés de la base car ne pouvant pas être considérés.

*Chapitre 7*

# Résultats

## 7.1 Segmentation

La segmentation par vote majoritaire, en utilisant les masques calculés par les différents logiciels, donnent des résultats équivalents à la vérité terrain. Cependant, au niveau des segmentations 2.5 et 40% il y a des variations comme le présente le tableau ... Des exemples de ces différences sur quelques patients sont présentés dans les tableaux 7.1 et 7.2. Étant donné le grand nombre de voxels, la précision (nombre d'éléments attribués à la classe A sur le nombre d'éléments appartement à la classe A) et le rappel (nombre d'éléments attribués à la classe A sur le nombre d'éléments attribués à la classe A) ne sont pas présenté car peu significatifs :

| Patient | Nombre pixels différents | nombre de pixels de la lésions dans la vérité terrain |
|---|---|---|
| A | 3 | $10.6 * 10^6$ |
| B | 14 | $5.1 * 10^5$ |
| C | 36 | $9.2 * 10^5$ |
| D | 0 | $4.8 * 10^5$ |
| E | 0 | $3.6 * 10^6$ |

TABLE 7.1 – Résultats de la segmentation 2.5 à l'aide des boîtes initiales.

| Patient | Nombre pixels différents | nombre de pixels de la lésions dans la vérité terrain |
|---|---|---|
| A | 3 | $10.6 * 10^6$ |
| B | 63 | $5.1 * 10^5$ |
| C | 82 | $9.2 * 10^5$ |
| D | 0 | $4.8 * 10^5$ |
| E | 0 | $3.6 * 10^6$ |

TABLE 7.2 – Résultats de la segmentation 40% à l'aide des boîtes initiales.

Pour certains patients les résultats sont équivalents à la vérité terrain, d'autres ont quelques pixels de différence (patients A). Ceci est dû à une légère différence dans les valeurs de SUV à la conversion. Ceci est acceptable étant donné que les pixels se trouvent généralement sur les bords et il est donc impossible de savoir si ils appartiennent à la lésion ou non. Cependant d'autres patients (B et C) ont des différences plus importantes. Après observation et comme on peut le voir sur la figure 7.1, la boîte initiale ne permet pas d'obtenir la vérité terrain. Les erreurs sont donc dues à la boîte initiale. Il sera donc nécessaire de vérifier la bonne récupération des boîtes initiales par la suite afin de corriger ces erreurs.

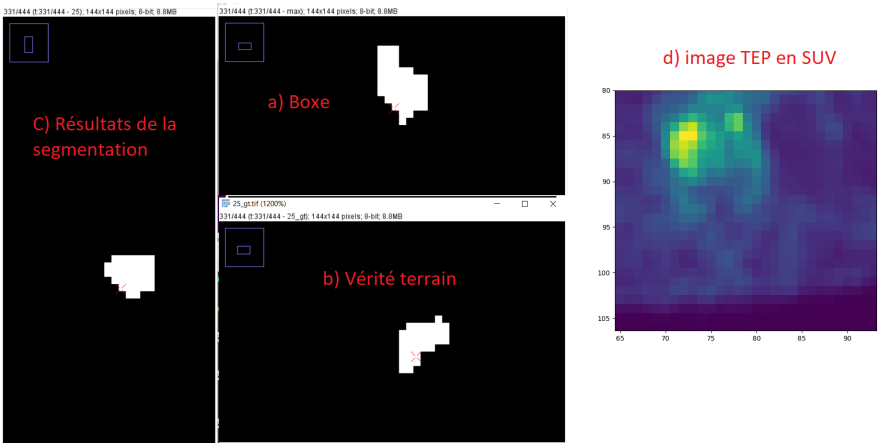



FIGURE 7.1 – Erreur dans la boîte initiale influençant les segmentations

La segmentation réalisée par F. Vogeli n'a pour l'instant été réalisée que par V-net. Cela correspond au deuxième V-net [voir section 5.1.2] qui permet la détection des lésions. Les deux modalités (PET et CT) ont été utilisées. Les résultats ne sont pour l'instant pas satisfaisants. Ceci donne cependant une base pour le futur.

## 7.2 Radiomics

Une fois les caractéristiques calculées à l'aide de pyradiomics, il est nécessaire de valider nos résultats. Deux méthodes sont utilisées. Dans une premier lieu, l'ISBI propose une matrice fantôme qui sera utilisée pour calculer les caractéristiques en utilisant pyradiomics. Ceci permet de valider les définitions utilisées pour les calculs, en comparant avec les résultats trouvés sur cette matrice par les membres de l'ISBI. Ensuite les résultats obtenus sur nos images TEP seront comparées avec la vérité terrain.

### 7.2.1 Validation par utilisation de l'ISBI

| | Pyradiomics | ISBI |
|---|---|---|
| HGLRE-GLRLM | 9,7 | 9,7 |
| LGLRE-GLRLM | 0,603 | 0,603 |
| LRE-GLRLM | 3,06 | 3,06 |
| SRE-GLRLM | 0,705 | 0,705 |
| SZHGE-GLSZM | 2,76 | 2,76 |
| ZLNU-GLSZM | 1 | 1 |
| ZP-GLSZM | 0,0676 | 0,0676 |
| contrast-GLCM | 5,32 | 5,32 |
| dissimilarity-GLCM | 1,43 | 1,43 |
| energy-GLCM | 0,303 | Non présente |
| entropy-GLCM | 2,4 | 2,4 |
| hgze-GLSZM | 15,6 | 15,6 |
| homogenity-GLCM | 0,677 | 0,678 |
| lgze-GLSZM | 0,253 | 0,253 |
| lzlge-GLSZM | 503 | 503 |
| maximum | 6 | 6 |
| sze-GLSZM | 0,255 | 0,255 |

TABLE 7.3 – Comparaison Pyradiomics et ISBI sur le fantôme

Les valeurs des caractéristiques calculées par le ISBI et avec pyradiomics sont équivalentes sur le fantôme (outre l'homogénéité avec un très faible écart-type de $7.7x10^{-4}$). Il n'y a donc pas d'erreur dans notre implémentation basée sur Pyradiomics. Les caractéristiques calculées sur le fantôme avec l'implémention utilisée pour le calcul de la vérité terrain sont elles aussi, équivalentes à l'ISBI. Cela implique donc que les caractéristiques ont la même définition dans Pyradiomics et avec l'implémention utilisée pour le calcul de la vérité terrain.

### 7.2.2 Comparaison des valeurs calculées par pyradiomics avec la vérité terrain

Les valeurs sur les images d'intérêt devraient donc être équivalents en Python et pour la vérité terrain. Lorsque l'on compare la vérité terrain, il réside cependant toujours des différences comme le montre le tableau 7.4. Les valeurs présentées correspondent aux écart-types et aux moyennes des différences observées entre les valeurs normalisées de pyradiomics et celles de la vérité terrain. La normalisation a été réalisée par l'écart-type de chaque variable, des valeurs de la vérité terrain. Les calculs ont été réalisés sur 5 patients.

| | **SUVmax** | **SRE** | **Homogeneity** | **Dissimilarity** | **LRE** | **Contrast** |
|---|---|---|---|---|---|---|
| **Moyenne** | 3,83E-08 | 1,01E-02 | 2,54E-02 | 3,58E-02 | 5,00E-02 | 5,38E-02 |
| **Ecart-type** | 5,52E-09 | 7,13E-03 | 1,64E-02 | 1,63E-02 | 4,18E-02 | 3,03E-02 |
| | **SZE** | **Entropy** | **ZP** | **ZLNU** | **Energy** | **SZHGE** |
| **Moyenne** | 7,18E-02 | 8,72E-02 | 1,21E-01 | 3,30E-01 | 3,91E-01 | 8,86E-01 |
| **Ecart-type** | 7,50E-02 | 7,32E-02 | 1,08E-01 | 6,91E-01 | 3,57E-01 | 1,48E+00 |
| | **HGZE** | **HGRE** | **LGRE** | **LGZE** | **LZLGE** | |
| **Moyenne** | 8,90E-01 | 9,13E-01 | 2,88E+01 | 3,20E+01 | 4,08E+01 | |
| **Ecart-type** | 1,37E+00 | 1,39E+00 | 1,37E+01 | 1,40E+01 | 7,15E+01 | |

TABLE 7.4 – Ecart-types et moyennes des différences observées entre les valeurs normalisées de pyradiomics et celles de la vérité terrain. (la normalisation a été réalisée par l'écart-type par variable des valeurs de la vérité terrain). Les calculs ont été réalisés sur 5 patients.

Les valeurs ont été triées de la plus faible différence (En haut à gauche) à la plus grande (en bas à droite). La caractéristique de premier ordre SuvMax est très proche de la vérité terrain. Il est donc peu probable que le problème se trouve au niveau de la segmentation ou de la conversion en SUV. La majorité des caractéristiques ont une différence qui n'est pas à négliger mais les valeurs restent dans le même ordre de grandeur que la vérité terrain. LGRE, LGZE et LZLGE ont un facteur de l'ordre de 10 entre les valeurs de la vérité terrain et celles calculées.

## 7.3 Validation de l'implémentation RSF en python

Les méthodes de RSF écrites en R et en python sont comparées afin de voir si les résultats obtenus sont équivalents ou si l'une surpasse l'autre.
Tout d'abord sur la base de données vétéran. Le tableau 7.5 donne des valeurs calculées sur les erreurs de prédictions trouvées lors de la détermination des paramètres optimaux. Il montre bien que la méthode en R donne de meilleurs résultats (la valeur de l'erreur de prédiction la plus proche de 0 étant la meilleure). La valeur de la meilleure erreur de prédiction en R est de 0,24. Pour comparaison celle obtenue par Iswaran [1] était de 0,29 environ mais trouvée à partir de l'OOB ce qui amène peut-être à expliquer qu'elle soit différente de celle du papier. La meilleure erreur de prédiction est de 0,39 en python, ce qui est plus élevée, mais reste tout de même raisonnable. La variance du set de test est aussi en moyenne plus élevé en python que en R.

| | Minimum | Maximum | Moyenne | Variance |
|---|---|---|---|---|
| En R | 0,2436 | 0,3479 | 0,3039 | 0,0003 |
| En Python | 0,3922 | 0,5944 | 0,5179 | 0,0012 |

TABLE 7.5 – Comparaison des résultats de l'erreur de prédiction moyenne sur le set de test sur la base vétéran en R et en python.Les valeurs ont été calculées sur les erreurs de prédictions trouvées lors de la détermination des paramètres optimaux.

Cependant, lorsque l'on regarde d'autres bases comme celle du myélome multiple, la variance, que se soit entre modèles de mêmes paramètres ou entre modèles à paramètres différents, la variance semble beaucoup plus faible en python comme le montre le tableau 7.6, qui donne les valeurs calculées sur les erreurs de prédictions trouvées lors de la détermination des paramètres optimaux.

| | Minimum | Maximum | Moyenne | Variance |
|---|---|---|---|---|
| En R | 0,4195 | 0,6896 | 0,5548 | 0,0023 |
| En Python | 0,4501 | 0,6380 | 0,5415 | 0,0005 |

TABLE 7.6 – Comparaison des résultats de l'erreur de prédiction moyenne sur le set de test sur la base vétéran en R et en python.Les valeurs ont été calculées sur les erreurs de prédictions trouvées lors de la détermination des paramètres optimaux

Il semble donc résider des différence entre les deux implémentations. Ces possibles différences seront discutées dans le chapitre 8.

De plus, au niveau du temps de calcul, la méthode en R est beaucoup plus rapide que celle en python. Si on compare le temps de calcul pour la réalisation d'une validation croisée avec nombre d'arbres = 20, max-caractéristiques = 136, min_sample_split = 8 et mode = logRank, sur la base contenant toutes les caractéristiques, on trouve que le temps de calcul en R est de 7 secondes contre 354.9 secondes en python. En prenant les même paramètres mais en remplaçant le logRank par le logRankRandom, on trouve 6 secondes pour R et 172,7 secondes pour python. Le fait que le temps de calcul en python est beaucoup plus rapide avec logRank random qu'avec logrank est dû au fait que le calcul des meilleures séparations au niveau des noeuds représente 98,8% du temps de calcul d'un arbre.

## 7.4 Résultats sur la base du myélome multiple

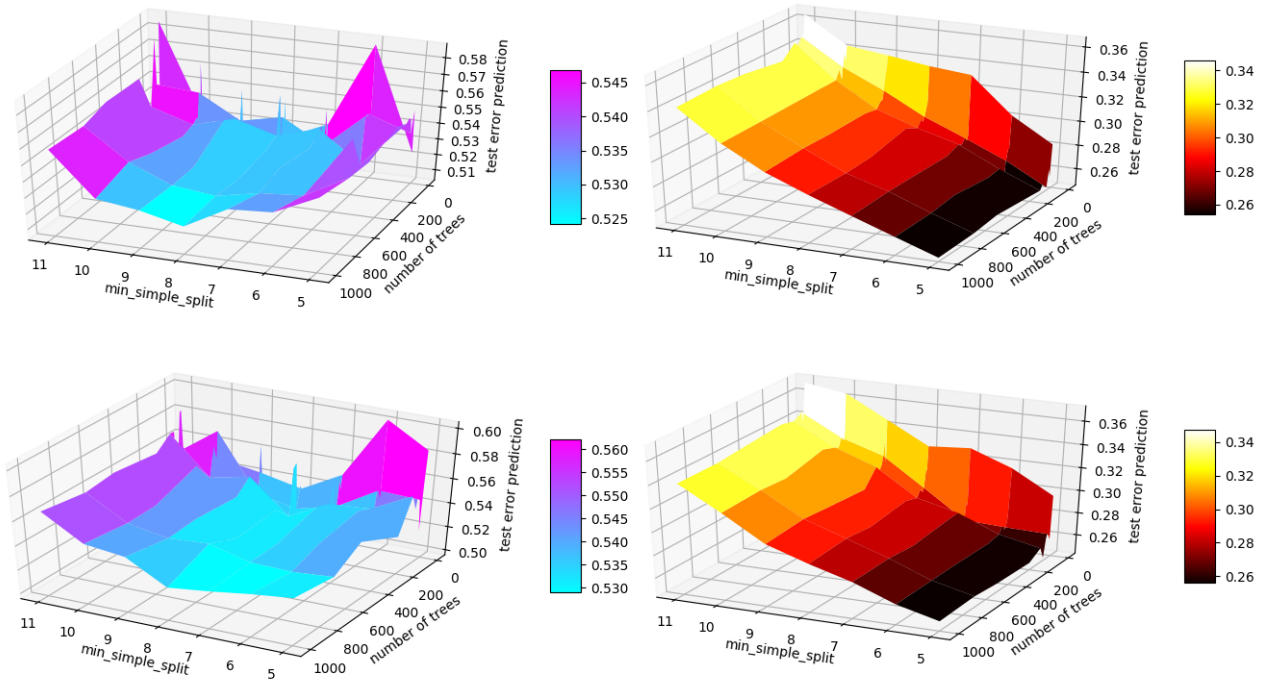
### 7.4.1 optimisation des hyperparamètres

Il est intéressant de regarder l'influence des différents paramètres sur les résultats afin de voir quels paramètres font varier l'erreur de prédiction et comment se comporte notre modèle. 4 paramètres sont testés :

- Le nombre d'arbres (variation entre 5 et 1000 abres)
- Le mode (logRankRandom ou LogRank)
- min-Sample-Split ou minimum d'individus nécéssaire pour réaliser une séparation au niveau d'un noeud (variation entre 5 et 12)
- max-features ou nombre maximal de caractéristiques tirées au niveau des noeuds pour réaliser une séparation

Cette influence des paramètres est observée sur la variation du minimum de l'erreur de prédiction, lorsque l'on applique la prédiction sur une base test 7.2a et lorsque l'on applique la prédiction sur la base d'entraînement

7.2b. Pour cela on utilise la base donnée du myélome multiple contenant toutes les caractéristiques et le modèle de RSF réalisé en python.




(a) set de test

(b) set d'entraînement

FIGURE 7.2 – Graphiques présentant l'erreur de prédiction minimale sur le set de test (a) et le set d'entraînement (b) avec python en fonction du mode (LogRank en haut : graphique du haut, LogRank Random : graphique du bas), du nombre d'arbres et de min-sample-split.

Les graphiques 7.2a et 7.2b présentent l'erreur de prédiction minimale sur le set de test en fonction du mode (LogRank ou LogRank Random), du nombre d'arbres et de min-sample-split. Le paramètre de caractéristiques maximales n'est pas présent dans ces graphiques. En effet, le calcul des points du graphique s'est fait de la manière suivante :
Pour chaque combinaison de paramètres mode/min-sample-split/arbres, une moyenne a été réalisée sur l'erreur de prédiction. Soit :

$$err_c(m, a, s) = \frac{\sum_{f \in F} err_o}{3},$$

avec :

- $err_o$ : l'erreur observée,
- m : le mode $\in$ {LogRank, LogRankRandom },
- s : $min_sample_split \in$ {i pour i dans 5 :11 },
- a : le nombre d'arbres $\in$ {5,15,25,40,50,70,100,200,500,700,1000 },
- $err_c$ : f le nombre maximal de caractéristiques $\in$ F avec F = {0.5*nombre de caractéristiques, 0.7*nombre de caractéristiques, nombre de caractéristiques }

,

On remarque que pour le set de test, les valeurs d'erreurs de prédiction en logrank Random sont légèrement supérieures. Cela ne semble cependant pas affecter les valeurs pour le train test.

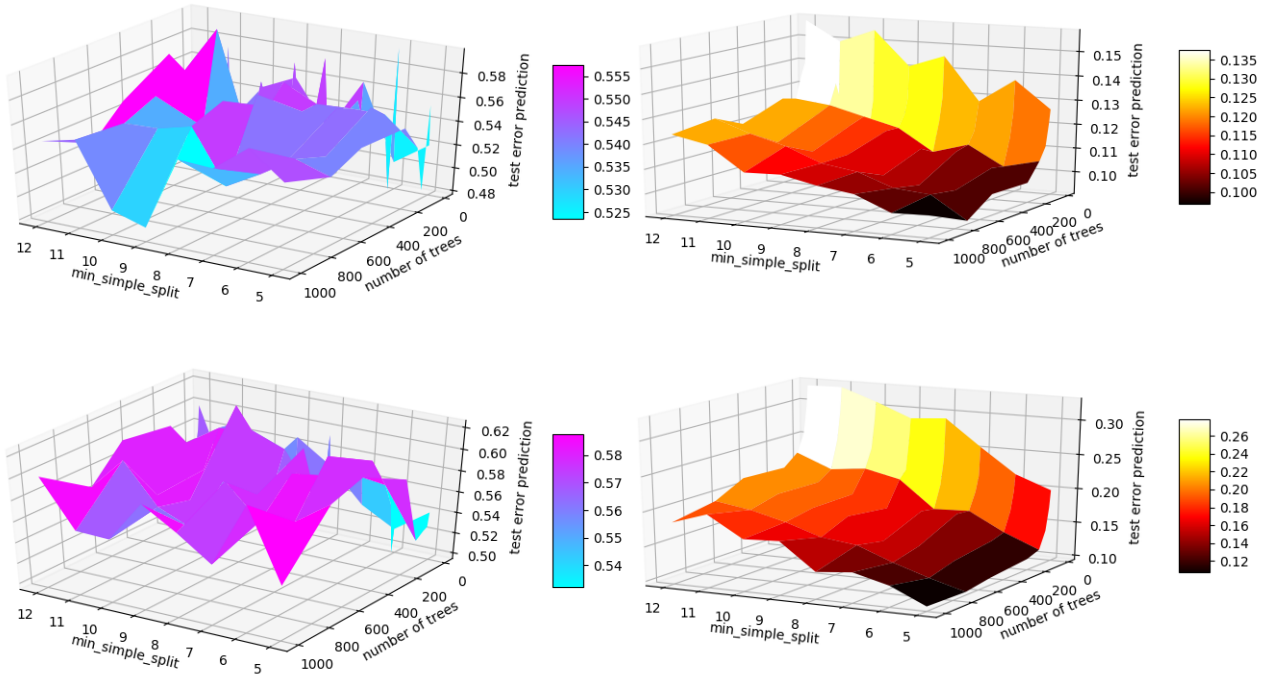
Que ce soit pour le set de test ou le set d'entraînement, le bruit semble plus grand quelque soit le mode lorsque le nombre d'arbres est inférieur à 200. Un nombre d'arbre supérieur à 200 semble induire une diminution du bruit mais l'augmentation d'un nombre d'arbres (lorsque supérieur à 200) ne semble pas avoir d'influence

avec LogRank. Une augmentation du nombre d'arbres semble cependant induire une diminution de l'erreur pour le logrankRandom. Ce même, la diminution de min-sample-split diminue l'erreur du test avec logRankRandom mais la valeur de min-sample-split optimal avec logRank se trouve entre 7 et 10. Pour le set d'entraînement, une diminution de min-sample-split diminue l'erreur quelque soit le mode. Ceci s'explique par le fait qu'une diminution du nombre minimal d'individus par noeud, induit une diminution du biais, mais augmente aussi de l'overfitting. Enfin, afin de voir l'influence du paramètre des caractéristiques maximum dans l'implémentation python,on réalise le tableau 7.7. Les valeurs correspondent à une moyenne sur le paramètre max-features pour toutes les valeurs du tableau.

| Max-features | Moyenne des errreurs sur le set de test | Moyenne des erreurs sur le set d'entraînement |
|---|---|---|
| 68 | 0.5435 | 0.305 |
| 95 | 0.5423 | 0.297 |
| 136 | 0.538 | 0.293 |

TABLE 7.7 – Erreurs de prédiction en fonction du paramètre max-features

Le fait d'augmenter le nombre de caractéristiques qui peuvent être sélectionnées par noeud, induit une diminution légère de l'erreur de prédiction mais est aussi risque d'"overfitting". On peut aussi comparer ces résultats avec les courbes trouvées en R. Ceci permet de comparer l'influence des paramètres dans nos deux modèles.




(a) set de test

(b) set d'entraînement

FIGURE 7.3 – Graphiques présentant l'erreur de prédiction minimale sur le set de test (a) et le set d'entraînement (b) avec R en fonction du mode (LogRank en haut : graphique du haut, LogRank Random : graphique du bas), du nombre d'arbres et de min-sample-split.

On remarque dans le graphique 7.3a que le bruit est plus fort avec les modèles obtenus en R qu'en python. De plus, le graphique 7.3b indique que le logRank a plus d'influence sur les valeurs d'erreurs de prédiction en R et en python. Ces différences de comportement des modèles confirment la différence d'implémentation.

Pour la suite des expériences, les modèles pour chaque base de données ont été réalisés avec les paramètres correspondant à ceux donnant la plus petite erreur de prédiction dans le tableau 7.8.

### 7.4.2 Valeur prédictive de notre base

L'erreur de prédiction trouvée pour la base des myélomes multiples avec toutes les caractéristiques (0,419 pour R et 0,45 pour python comme présenté dans le tableau 7.8) est relativement proche de 0,5 (une valeur à 0,5 induit une assignation aléatoire des résultats), mais tout de même inférieure ce qui induit une certaine prédiction. Dans son papier, Ishwaran [1] présente les résultats d'une base de patient à transplanter dont les meilleurs résultats se trouvent entre 0,45 et 0,5. Il considère que ce sont de bons résultats, car meilleurs que ceux en Cox ou RF. Il est donc nécessaire de regarder la comparaison avec Cox et RF pour notre base afin de voir si les résultats sont acceptables.

Le tableau 7.8 présente le meilleur résultat pour chaque base de données (les bases de données étant construites avec le même nombre de patients afin de les comparer).

| | toutes les caractéristiques | caractéristiques texturales | caractéristiques cliniques et volumiques | caractéristiques de la littérature |
|---|---|---|---|---|
| Error pred min. test | 0,4195 | 0,4287 | 0,3938 | 0,4075 |
| Var test | 0,2276 | 0,0099 | 0,0807 | 0,0594 |
| Mean train | 0,2324 | 0.3171 | 0,1331 | 0,1822 |
| Var train | 0,0363 | 0.0440 | 0,021 | 0,0297 |

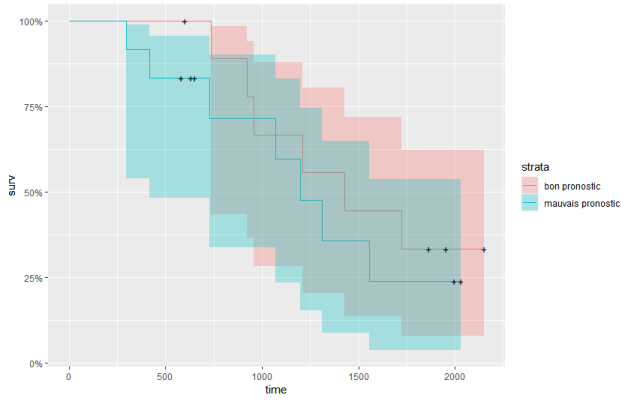
TABLE 7.8 – Erreurs de prédiction en fonction de la base de données

De plus, lorsque l'on change la base de données, par exemple en ne gardant que les caractéristiques cliniques et volumiques, les résultats peuvent s'améliorer. Ainsi, une base de données avec caractéristiques cliniques et volumiques seules permet d'avoir une erreur de prédiction minimale de 0,39. Le modèle a aussi été testé sur une base de données contenant tous les patients sans données manquantes avec les caractéristiques volumiques et cliniques (89 patients au lieu de 67) et on remarque que le résultat s'améliore (l'erreur de prédiction passe de 0.39 à 0.37), ce qui indique que le nombre de patients à un impact sur les résultats.

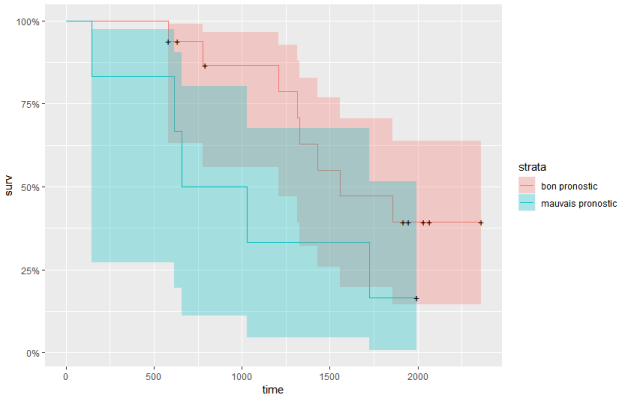
Un modèle est aussi construit avec une base de données comprenant seulement les caractéristiques cliniques trouvées comme prédictives dans la littérature [voir partie 6.3.3.4] et donne de meilleurs résultats qu'avec toutes les caractéristiques. Il est à noter que la variance reste peu élevée et notamment pour les images seules ou caractéristiques cliniques et volumiques seules.

Les résultats semblent meilleurs lorsque l'on utilise les caractéristiques cliniques et volumiques seules, et les résultats avec des caractéristiques texturales semblent meilleurs lorsque sont ajoutées les caractéristiques cliniques et volumiques.

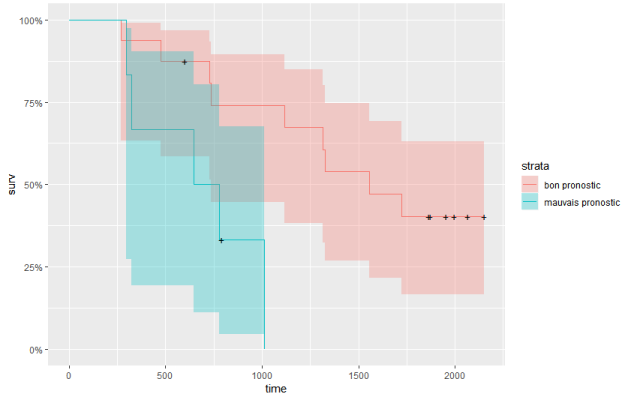
A la fin de notre modèle, nous souhaitons séparer en deux groupes les patients. Ceux avec un bon pronostic et ceux avec un mauvais pronostic. La meilleure séparation est déterminée à l'aide du log rank test sur les mortalités prédites. Sont présentées dans les figures 7.4a, 7.4b, 7.4c et 7.4d, les résultats obtenus pour chaque base de données. Les courbes présentées ci-dessous sont réalisées avec les résultats correspondant aux erreurs de prédiction présentées dans le tableau 7.8.



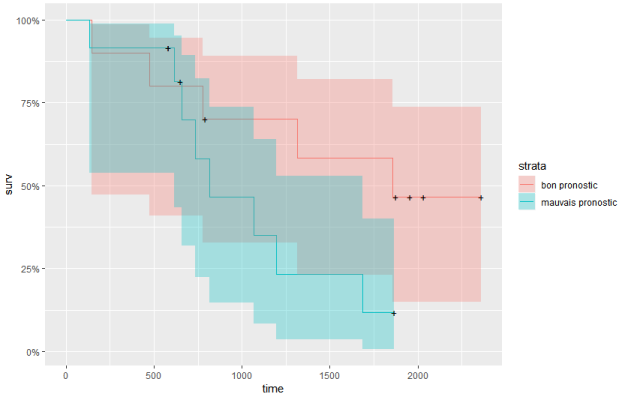

(a) Avec toutes les caractéristiques ; Err. = 0.42 ; logRank = 2.5 ; p-value = 0.1 ; nbre de patients par groupe = 14 et 8




(b) Avec les caractéristiques texturales ; Err. = 0.428 ; logRank = 2.4 ; p-value = 0.1 ; nbre de patients par groupe = 16 et 6




(c) Avec les caractéristiques cliniques et volumiques ; Err. = 0.39 ; logRank = 6.2 ; p-value = 0.01 ; nbre de patients par groupe = 13 et 9




(d) Avec les caractéristiques de la littérature ; Err. = 0.406 ; logRank = 2.4 ; p-value = 0.1 ; nbre de patients par groupe = 10 et 12

FIGURE 7.4 – Séparation en deux groupes (Bon et mauvais pronostic) selon les caractéristiques utilisées (67 patients)

La meilleure séparation correspond à celle obtenue avec les caractéristiques cliniques et volumiques. La p value est de 0.01 et la valeur de test de 6.3 ce qui indique donc un résultats interprétable. Cela permet bien de distinguer deux groupes.

Les autres séparations ont une p-value de 0.1 (supérieur à 0.05) qui pourrait indiquer que les courbes ne sont pas assez différentes.

### 7.4.3 Comparaison avec les Random Forest

Après détermination des paramètres optimaux, le random survival classifier est appliqué sur la base du myélome multiple avec toutes les caractéristiques. Nous observons la séparation en deux groupes. Pour la prédiction à 2, 4 et 5 ans, la répartition des patients dans les deux groupes au départ est telle que le classifieur prédit une seule classe pour tous les patients (ou seulement 2 ou 3 dans une classe). Les résultats ne sont donc pas interprétables. A 3 ans, on obtient une classification équilibrée (12 dans une classe et 11 dans l'autre). Les meilleurs résultats obtenus sont présentés dans le tableau 7.9.

Cela permet d'obtenir une courbe de Kaplan-Meier. La p-value est de 0.94 et la valeur du test du logRank de 0.05. On ne peut donc pas considérer cette séparation comme significative. Lorsque l'on utilise seulement les caractéristiques cliniques et volumiques, on obtient de meilleurs résultats qui sont présentés dans le tabelau 7.10. Les forêts aléatoires donnent d'assez bons résultats lorsque l'on utilise les caractéristiques cliniques et

| | Accuracy | Recall |
|---|---|---|
| set de test | 47.8% | 1.2 |
| set d'entraînement | 95.5% | 0.9 |

TABLE 7.9 – Résultats du RF classifier pour une séparation à 3 ans

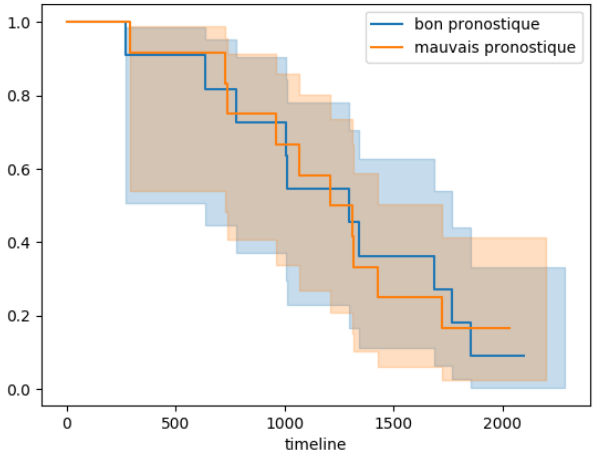



FIGURE 7.5 – Kaplan-Meier de la classification par RF à 3 ans sur la base du myélome multiple avec toutes les caractéristiques. Accuracy : 0.47, recall : 1.2, logRank = 0.05, p-value = .94, nombre de patients avec un bon et mauvais pronostique : 12 et 11

| | Accuracy | Recall |
|---|---|---|
| set de test | 60.8% | 0.9 |
| set d'entraînement | 91.9% | 0.95 |

TABLE 7.10 – Résultats du RF classifier pour une séparation à 3 ans, avec les caractéristiques cliniques et volumiques

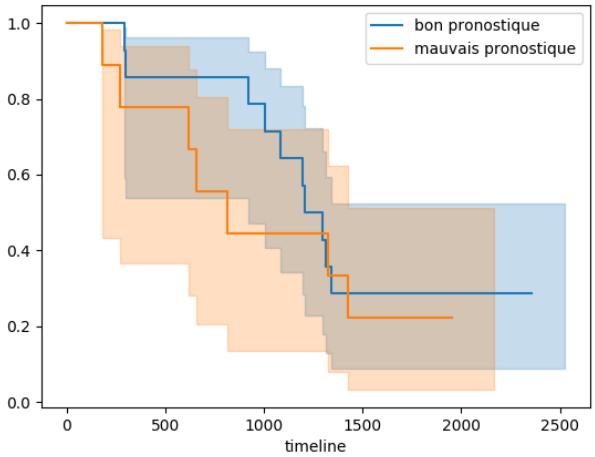



FIGURE 7.6 – Kaplan-Meier de la classification par RF à 3 ans sur la base du myélome multiple avec les caractéristiques cliniques et volumiques. Accuracy : 0.61, recall : 0.9, logRank = 0.56, p-value = 0.32, nombre de patients avec un bon et mauvais pronostique : 14 et 9

volumiques à 3 ans. Il ne permet cependant pas une séparation significative.

Il serait aussi intéressant de comparer nos résultats avec le modèle de Cox, qui est écrit pour l'étude de la survie, mais aussi au random forest regressor de Hothorn qui adapte les random forest regressor à la survie en utilisant des poids pour prendre en compte les données censurées (inverse probability of censoring (IPC) weights)

### 7.4.4 interprétation des biomarqueurs

Les 5 caractéristiques trouvées comme les plus prédictives pour chaque base de données sont présentées, en fonction de la moyenne de la valeur d'importance sur 100 tests/modèles (50 en python), mais aussi en fonction de leur fréquence s'apparition dans le top 5 pour les caractéristiques prédictives de la littérature (car seulement 14 caractéristiques en tout) et le top 10 pour les autres. La fréquence a un intérêt dans le cas où une caractéristique apparaît un nombre de fois très restreint mais avec une grande valeur VIMP, et reste avec une valeur faible le reste du temps. Dans ce cas là, on considère qu'elle est instable et donc non prédictive. L'utilisation de la valeur VIMP est utile quant à elle dans le cas où la caractéristique apparaîtrait souvent dans le Top 10 en raison du fait que peu de caractéristiques de la base ne soit prédictives et auraient une valeur négative ou nulle. On ne peut considérer une caractéristique prédictive si sa valeur est négative ou nulle. Ainsi, bien que dans le Top 10, elle ne pourrait pas être considérée comme prédictive. Il est à noter que les caractéristiques à valeur négative ne sont pas présentées dans le diagramme.

Les caractéristiques prédictives sont déterminées, dans un premier temps, en ne regroupant pas les caractéristiques texturales par technique de calcul (voir graphiques 7.7 et 7.8).

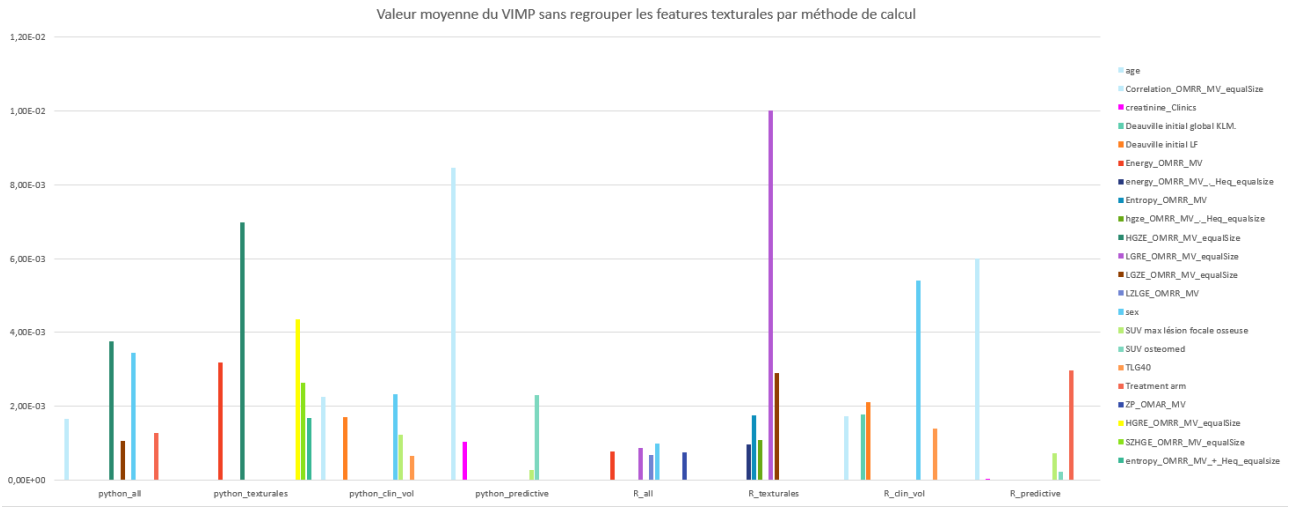



FIGURE 7.7 – Valeur moyenne du Vimp sans regroupement des caractéristiques texturales par méthode de calcul

Tout d'abord, si on se concentre sur les caractéristiques cliniques, on remarque que l'âge apparaît régulièrement, avec des valeurs de fréquence et VIMP relativement élevées. Il en est de même pour le sexe. Cependant, contrairement à l'âge qui apparaît comme prédictive dans la littérature, ce n'est pas le cas du sexe. On remarque aussi que les Deauvilles initiaux (surtout celui des lésions focales) apparaissent régulièrement et que le traitement apparaît avec avec des valeurs de VIMP correctes mais n'apparaît qu'une seule fois dans le diagramme de fréquence.

Au niveau des caractéristiques volumiques, le SUV max des lésions focales obtient de très bon résultats au niveau de la fréquences, mais de mauvais résultats au niveau des valeurs VIMP. La TLG40 et la TLGT40 apparaissent aussi mais seulement dans la base de caractéristiques cliniques et volumique, et non dans la base contenant toutes les caractéristiques.

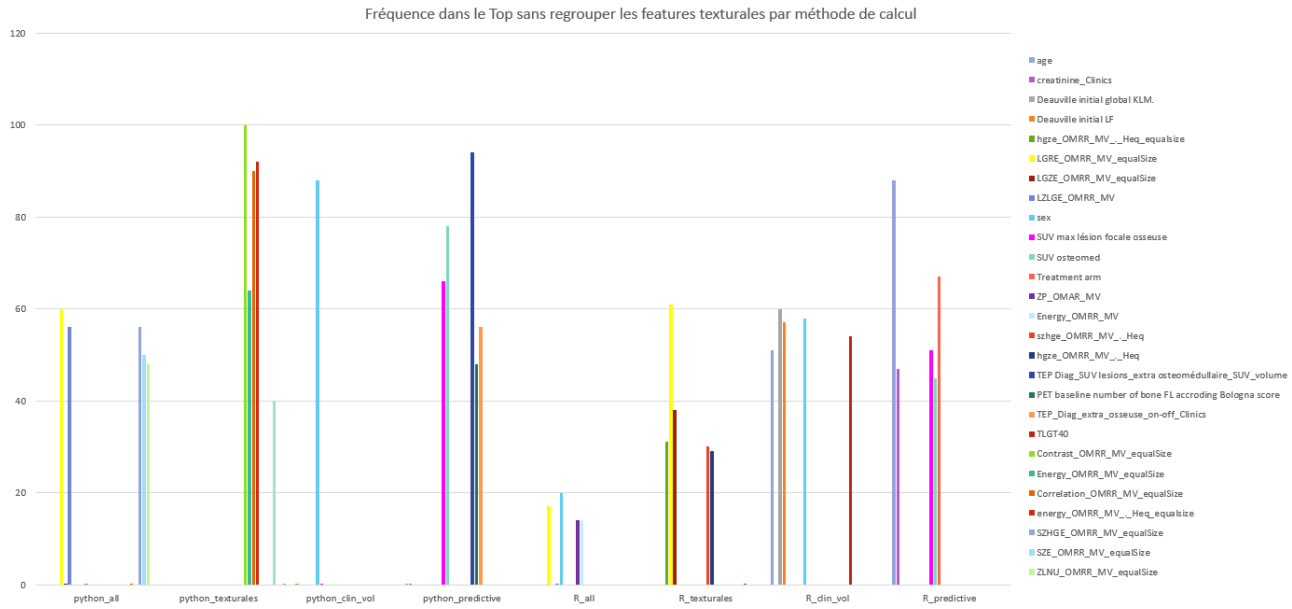



FIGURE 7.8 – Fréquence dans le Top sans regroupement des caractéristiques texturales par méthode de calcul

Enfin, au niveau des caractéristiques texturales prédictives hgze_OMRR_._Heq_equalsize, HGZE_OMRR_MV_equalSize, hgze_OMRR_MV_._Heq semblent apparaître régulièrement avec des valeurs relativement hautes. Il en est de même pour LGRE_OMRR_MV_equalSize et LGZE_OMRR_MV_equalSize, et dans une moindre mesure Energy_OMRR_MV_equalSize, Energy_OMRR_MV, energy_OMRR_MV_._Heq_equalsize, Entropy_OMRR_MV, entropy_OMRR_MV_._Heq_equalsize

LZLGE_OMRR_MV apparaît régulièrement dans le diagramme de fréquence avec des valeurs hautes, mais n'obtient que des valeurs VIMP faibles. Ceci peut peut-être, être expliquer par son caractère instable. Les autres caractéristiques texturales présentes dans le graphique n'apparaissent que dans une base avec un seul langage. On ne peut donc conclure sur leur caractère prédictif, en émettant l'hypothèse que le caractère prédicatif dépend de la base utilisée.

Enfin, un constat important pouvant être réalisé sur ces graphiques est que toutes les caractéristiques texturales trouvées comme prédictives ont été calculées sur une matrice avec un échantillonage relatif (0MRR, nombre de bins fixé à 64). L'égalisation de l'histogramme (Heq) ou l'égalisation de la taille des voxels (equalsize) ne semblent pas être un critère particulier de prédiction.

Dans un second temps nous regroupons les caractéristiques texturales pour éviter de prendre en compte la technique de calcul. Deux méthodes de regroupement pour les valeurs VIMP :

— En faisant la moyenne des valeurs de chaque méthode de calcul. (Figure 7.11)

— En faisant la somme des valeurs de chaque méthode de calcul (Figure 7.10)

Faire la moyenne induit que les caractéristiques cliniques et volumiques, qui n'apparaissent qu'une seule fois auront autant de chance d'apparaître dans le classement que les caractéristiques texturales. Cependant, si une caractéristique texturale n'a qu'une seule méthode de calcul permettant d'être prédictible, elle sera diminuée et n'apparaîtra pas dans le classement. On utilise donc une autre méthode de présentation des résultats qui est de faire la somme des valeurs de chaque méthode de calcul.

Les caractéristiques apparaîssant deux fois ou plus dans le diagramme de fréquence 7.9 sont HGZE, LGRE et l'énergie.

Au niveau des valeurs moyennes de VIMP en regroupant par somme HGZE apparaît 3 fois, l'entropie 2 fois et LGRE une fois mais avec une grand fréquence.

Bien que le sexe soit une caractéristique clinique et qu'il ne soit pas sommé, il apparaît comme étant prédictif,

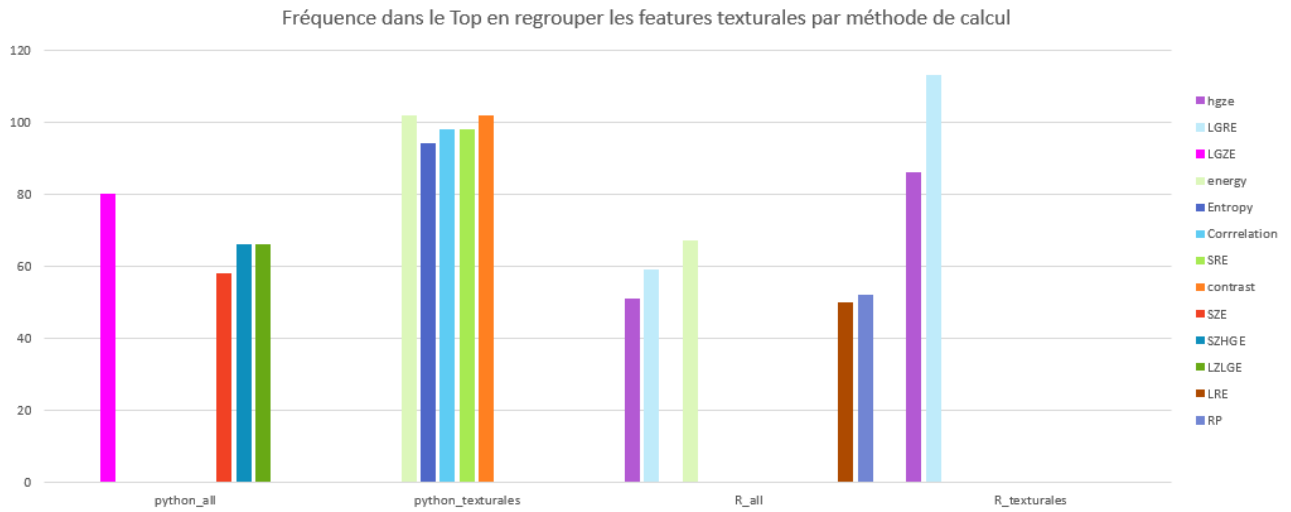



FIGURE 7.9 – Fréquence dans le Top avec regroupement des caractéristiques texturales par méthode de calcul

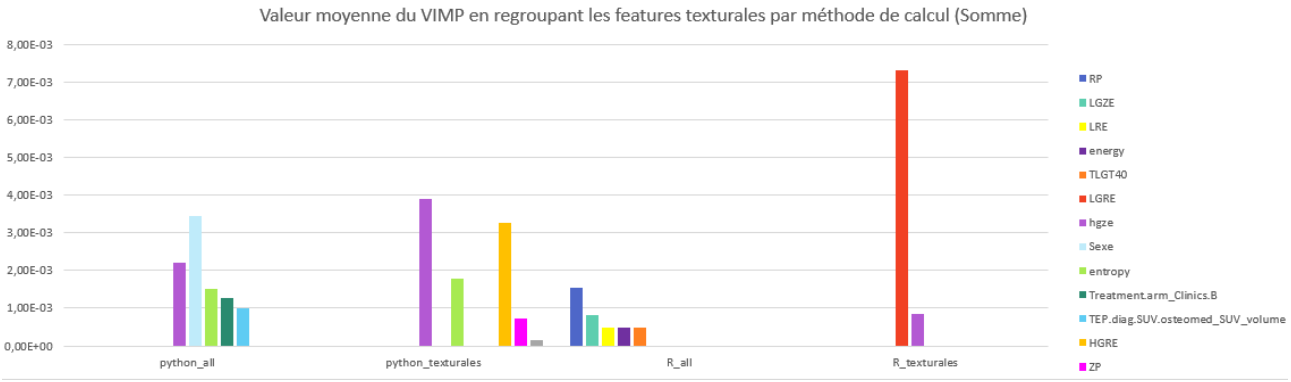



FIGURE 7.10 – Valeur moyenne du VIMP avec regroupement des caractéristiques texturales par méthode de calcul (Somme)

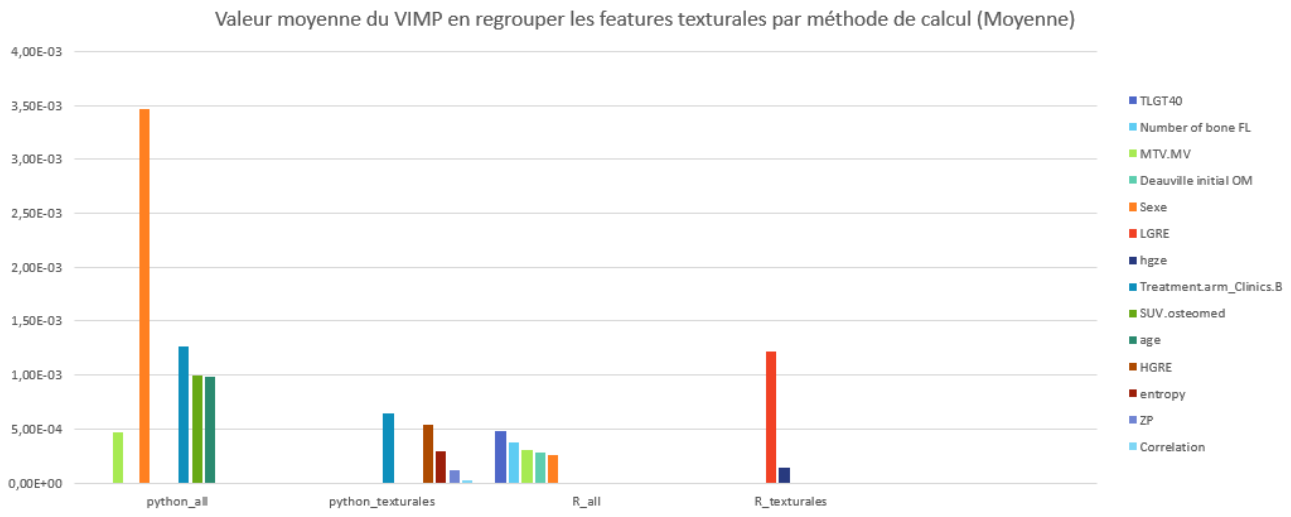
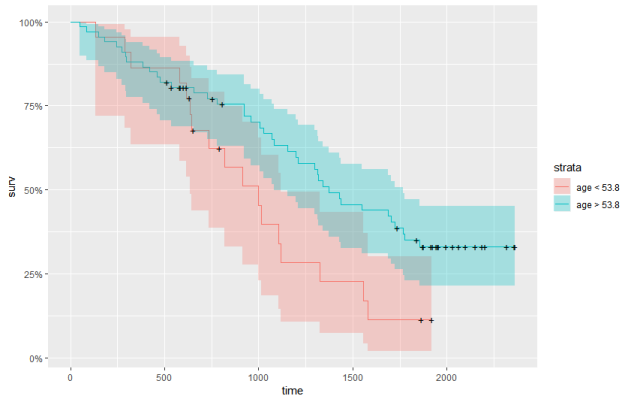



FIGURE 7.11 – Valeur moyenne du Vimp avec regroupement des caractéristiques texturales par méthode de calcul (Moyenne)

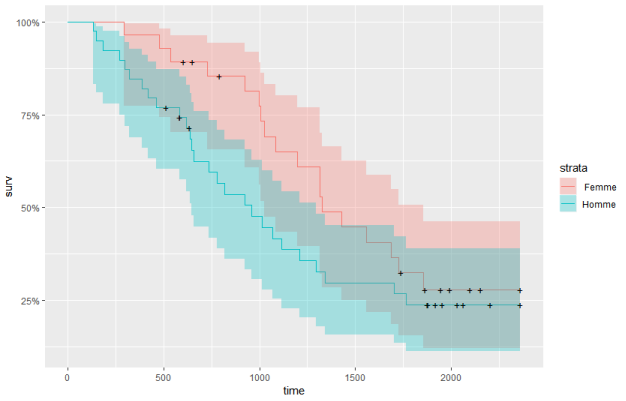
ce qui confirme sa valeur prédictive. Lorsque l'on moyenne, le sexe est donne ici la plus grande valeur. LGRE et le traitement paraissent avoir une valeur VIMP relativement grande. On retrouve dans une moindre mesure le suv Osteomedumaire et l'âge. HGZE, HGRE et l'entropie apparaissent aussi mais plus faiblement.

Ces résultats sont comparés avec des courbes de Kaplan-Meier. Ceux-ci sont réalisés pour chaque caractéristique en choisissant la valeur de séparation en deux groupes grâce au logRank lorsque la variable n'est pas binaire. Des modèles de Cox ont aussi été réalisés, mais les résultats n'étant pas interprétables car non significatifs (p-value trop élevée), ils ne sont pas présentés ici.

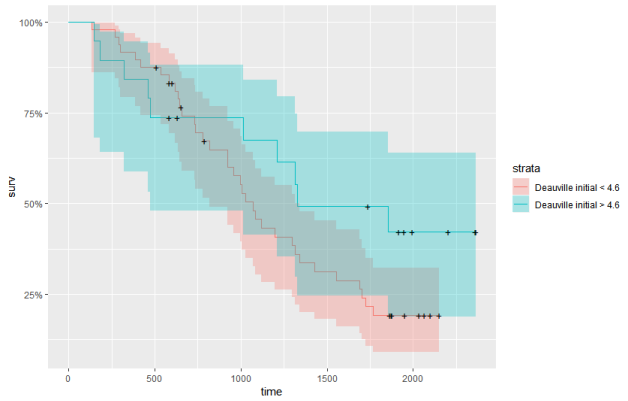
Les courbes sont réalisées sur la base contenant les 67 patients (toutes les caractéristiques) afin de pouvoir comparer nos courbes. On se base sur un modèle avec une erreur de prédiction de 0,39. Les courbes sont présentées avec un intervalle de confiance à 95% [voir méthode de calcul section 6.3.3.4]. Seules les courbes de caractéristiques présentées comme prédictives par notre base sont présentées ici.



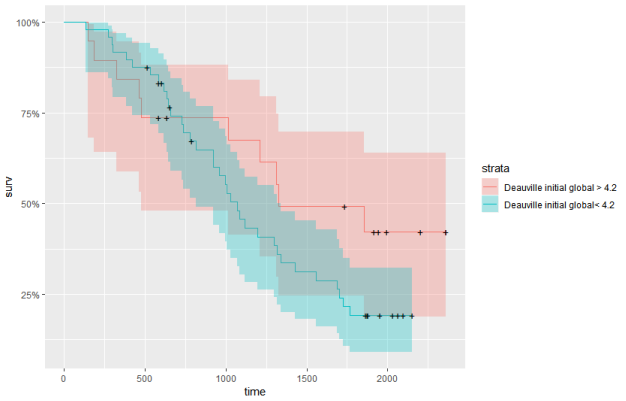

(a) Age : logRank = 2.9, p=0.09, nombre de patients par groupe : 47 et 20



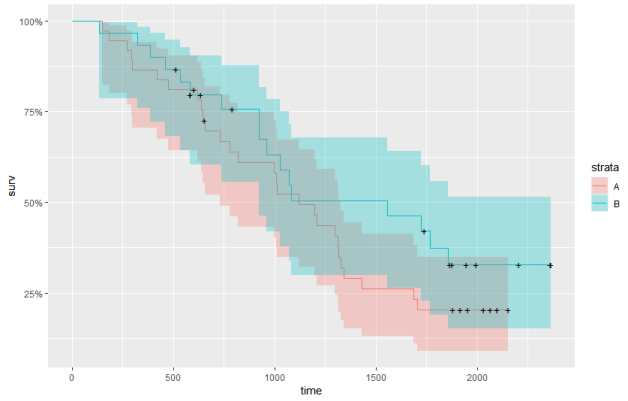

(b) Sexe : logRank = 1.9, p=1.45, nombre de patients par groupe : 28 et 39



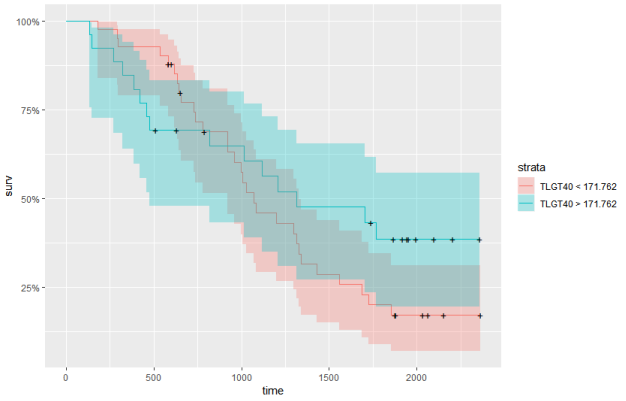

(c) Deauville initial des LF : logRank = 2.2, p=0.1, nombre de patients par groupe : 48 et 19




(d) Deauville initial global : logRank = 2.2, p=0.1, nombre de patients par groupe : 19 et 48

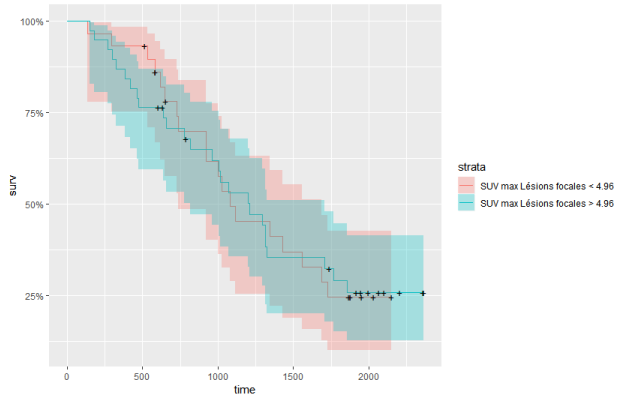



(e) Traitement : logRank = 1.4, p=0.1, nombre de patients par par groupe : 28 et 30

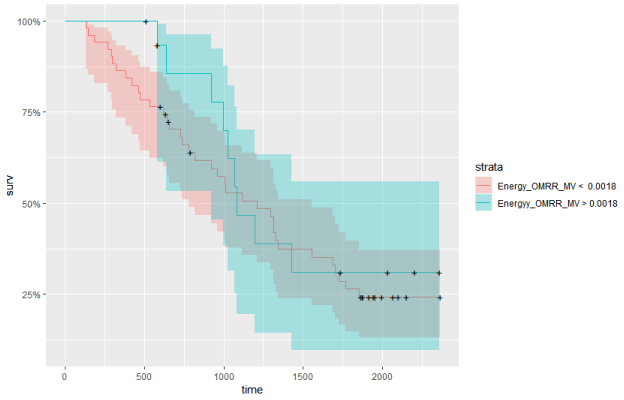



(f) TLGT40 : logRank = 1.3, p=0.2, nombre de patients par groupe : 41 et 26

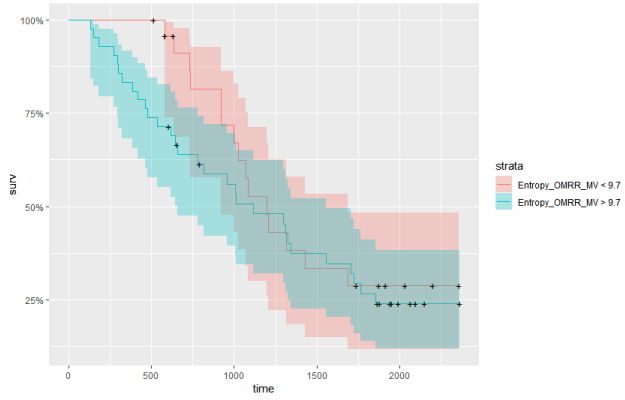
FIGURE 7.12 – Séparation par les caractéristiques cliniques (base de 67 patients)



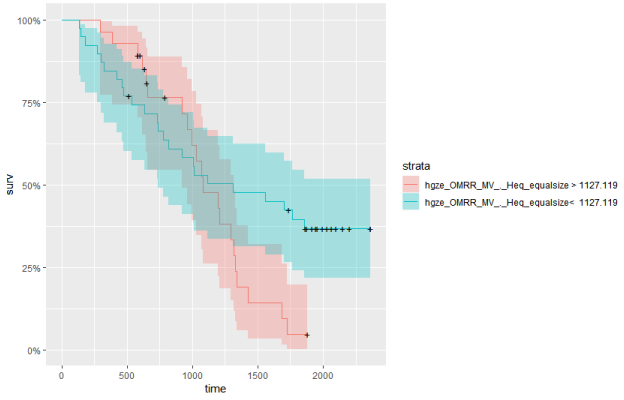

(a) SUV des Lésions focales : logRank = 0, p=1, nombre de patients par groupe : 29 et 38



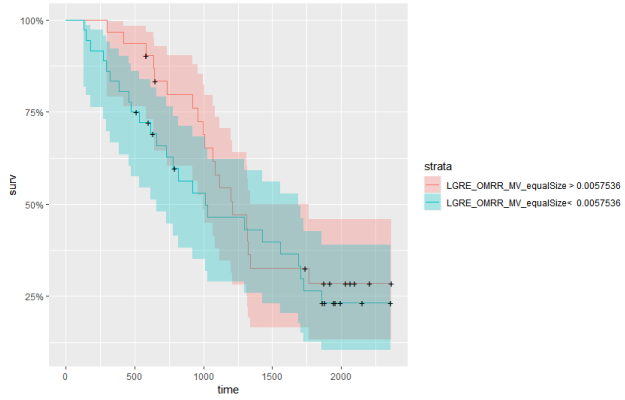

(b) $l'energy\_OMRR$ : logRank = 0.4, p=0.5, nombre de patients par groupe : 51 et 16



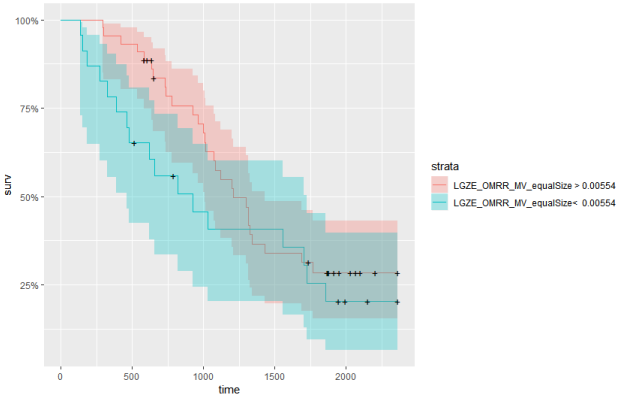

(c) $l'entropy\_OMRR$ : logRank = 0.6, p=0.4, nombre de patients par groupe : 25 et 42




(d) $hgze\_OMRR\_MV\_.\_Heq_equalsize$ : logRank = 2.9, p=0.09, nombre de patients par groupe : 28 et 24




(e) $LGRE\_OMRR\_MV\_equalSize$ : logRank = 0.7, p=0.4, nombre de patients par groupe : 31 et 36




(f) $LGZE\_OMRR\_MV\_equalSize$ : logRank = 1.4, p=0.2, nombre de patients par groupe : 44 et 23

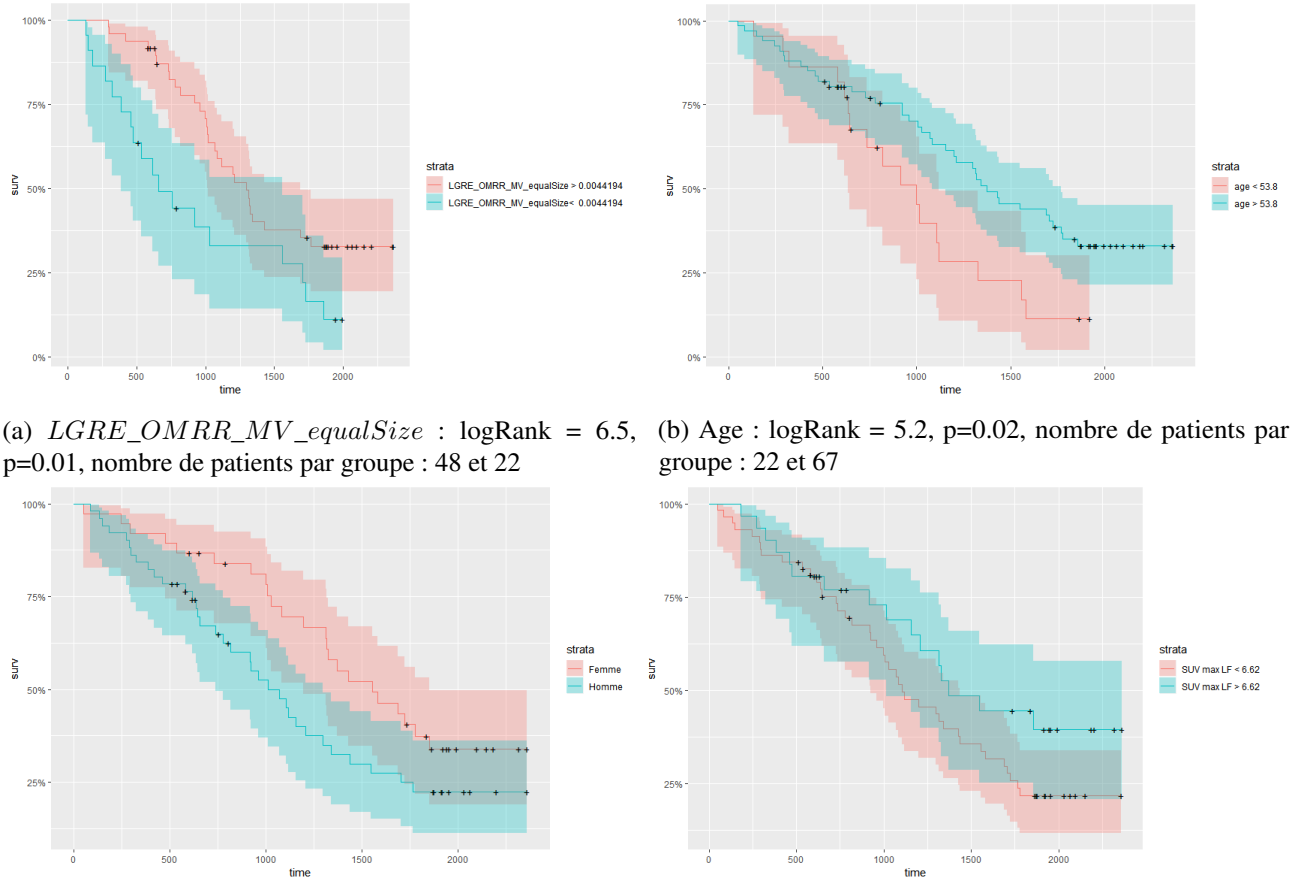
FIGURE 7.13 – Séparation par les caractéristiques image (base de 67 patients)

Les séparations selon le sexe, l'age, TLGT40 et les deauvilles sont correctes mais la p-value semble trop élevée pour être considérée comme significative. Ceci pourrait expliquer pourquoi le Deauville et la TLGT ne sont pas retrouvés dans la littérature comme significatifs.

Les femmes semblent avoir un meilleur pronostic que les hommes, tandis que des personnes plus jeunes ont un moins bon pronostic qu'une personne âgée de plus de 54 ans. Ceci est en contradiction avec ce qui est indiqué dans la littérature. Le traitement et la SUV ont des résultats qui ne permettent pas de considérer comme significative la séparation en deux groupes.

Les caractéristiques texturales semblent quant à elles, être significatives seulement au début de la maladie (avant 1000 jours soit 2,7 ans).

Cependant, si l'on applique ces modèles de Kaplan-Meier à des bases de données possédant plus de patients, les résultats sont significativement supérieurs pour certaines caractéristiques. C'est notamment le cas pour le sexe, l'age et la suv_max des lésions focales lorsque l'on utilise la base avec les caractéristiques cliniques et volumiques (89 patients, erreur de prédiction de 0,40) et pour LGRE_OMRR_MV_equalSize lorsque l'on utilise la base avec les caractéristiques texturales (70 patients, erreur de prédiction de 0,40))




(a) $LGRE\_OMRR\_MV\_equalSize$ : logRank = 6.5, p=0.01, nombre de patients par groupe : 48 et 22

(b) Age : logRank = 5.2, p=0.02, nombre de patients par groupe : 22 et 67

(c) Sexe : logRank = 3.8, p=0.02, nombre de patients par groupe : 23 et 51

(d) SUV des Lésions focales : logRank = 2.3, p=0.1, nombre de patients par groupe : 58 et 31

FIGURE 7.14 – Séparation par les caractéristiques induisant des différences dans les courbes en fonction du nombre de patients (70 patients pour la figure (a) et 89 patients pour les figures (b), (c), (d))

Ainsi, on remarque que les séparations pour ses caractéristiques donnent de bien meilleurs résultats lorsque le nombre de patients est supérieur. Notamment pour LGRE qui donne une valeur de test du logRank de 6.5 et une p-value de 0.01. Il en est de même pour l'age et le sexe. Le résultat du suv Max est plus mitigé. Cependant, sa prédictivité est aussi discutée dans la littérature, mais il semblerait que sa valeur prédictive est valable sur le

long terme et moins sur le court terme. Ceci est confirmé par le graphique.

Au vu de ces résultats, une nouvelle base est créée avec les caractéristiques que l'on considère comme prédictives (Apparaissant de façon significative au moins 2 fois dans les figures 7.7 à 7.11) qui sont les suivantes :

| | | |
|---|---|---|
| Entropy_OMRR_MV | Energy_OMRR_MV | HGZE_OMRR_MV_._Heq |
| Energy_OMRR_MV_equalSize | LGRE_OMRR_MV_equalSize | LGZE_OMRR_MV_equalSize |
| HGZE_OMRR_MV_equalSize | HGZE_OMRR_._Heq_equalsize | Energy_OMRR_MV_._Heq_equalsize |
| Entropy_OMRR_MV_._Heq_equalsize | Sexe | Âge |
| Deauville initial des lésions focales | Deauville initial global | Traitement |
| SUV max des lésions focales | SUV osteomedullaire | TLGT40 |
| TLG40 | | |

On réalise une validation croisée avec les mêmes paramètres que précédemment en R, afin de comparer les résultats de notre base avec les résultats obtenus avec les autres bases. Ainsi on trouve :

| | Nouvelle base de données |
|---|---|
| Error pred min. test | 0,2958 |
| Var test | 0,1312 |
| Mean train | 0,1240 |
| Var train | 0,0066 |

TABLE 7.11 – Erreurs de prédiction sur la nouvelle base de données

En comparant avec le tableau 7.11 on remarque que les résultats avec la nouvelle base sont bien meilleurs que ceux obtenus avec les bases précédentes. Lorsque l'on cherche à réaliser deux groupes avec cette base de données et un modèle avec la même erreur minimale de prédiction (0,29) on obtient la courbe de Kaplan Meier suivante :

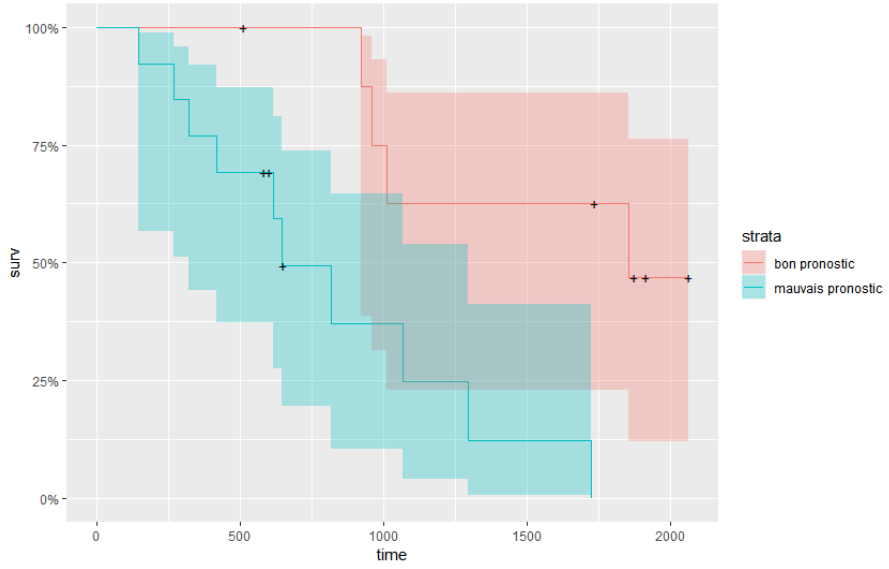



FIGURE 7.15 – Séparation sur notre nouvelle base de données ; erreur de prediction = 0.298 ; logRank = 7.9 ; p value = 0.005 ; nombre de patients dans le groupe de bon pronostic = 9 ; nombre de patients dans le groupe de bon pronostic = 13

Au regard de la p-value et de la valeur du test du logRank cette séparation est bien meilleure que celle obtenue avec les bases de données précédentes. Le graphique 7.16 donne l'importance des caractéristiques (valeur VIMP) et la fréquence dans le top 5 sur 100 itérations :

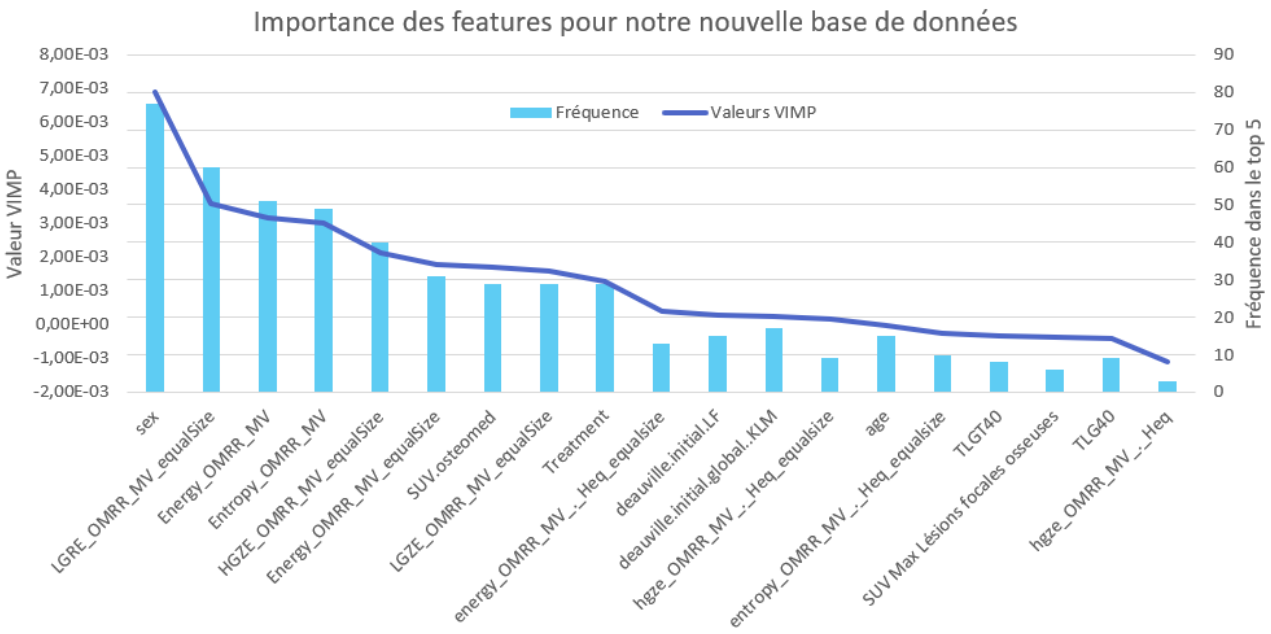



FIGURE 7.16 – Importance des caractéristiques obtenues avec notre nouvelle base de données

*Chapitre 8*

# Discussion

**La segmentation**

L'implémentation du vote majoritaire ainsi que des segmentations 2.5 et 40% sont correctes. Celle du k-means reste à implémenter. Il subsiste cependant des problèmes à régler au niveau des boîtes récupérées dans lesquelles sont réalisées les segmentations.

**L'extraction des caractéristiques**

La comparaison avec le fantôme de l'ISBI indique que les définitions utilisées sont bonnes. Cependant il reste des différences au niveau des images TEP par rapport à la vérité terrain. Il est possible que des problèmes persistent au niveau du paramétrage. Il serait alors intéressant d'utiliser les images cliniques de l'ISBI pour valider notre implémentation. Il reste néanmoins à prendre en compte que les personnes de l'ISBI ayant réalisé les calculs sur les images cliniques ont eu des difficultés (et n'ont toujours pas réussi) à converger vers des résultats exactement équivalents. Ceci indique que, même si l'on doit tout de même obtenir des valeurs dans le même ordre de grandeur, il est possible que les valeurs de caractéristiques calculées ne correspondent finalement jamais aux valeurs exactes de la vérité terrain.

**L'implémentation de RSF en Python**

Il réside des différences de résultats entre ceux trouvés en python et en R. L'algorithme des RSF possède beaucoup de paramètres et modifications possibles qui peuvent influencer les résultats, et qui ne sont pas forcément détaillés dans l'article de Ishwaran et al. [1]. Il est donc possible qu'il y ai encore des différences (la différence peut par exemple se trouver au nombre de valeurs testées pour déterminer la meilleure séparation). Il sera donc nécessaire de reprendre l'algorithme en s'aidant notamment du code de l'algorithme RSF disponible en C et en java sur internet. De plus, malgré les améliorations apportées, il reste encore une grande différence en ce qui concerne le temps de calcul. Il peut être ainsi intéressant de passer certaines fonctions en C pour accélérer le calcul des arbres.

**Comparaison avec les forêts aléatoires**

RF donne de bons résultats de classification à 3 ans lorsque l'on utilise les caractéristiques cliniques et volumiques, mais ne donne cependant pas une bonne séparation. Le nombre de patients est peut-être trop faible. En effet, il a été nécessaire d'éliminer les patients avec une censure avant 3 ans car ne pouvant être mis dans aucun des groupes.

RSF a pour avantage de prendre en compte la censure et de donner une meilleure séparation. Il donne aussi de meilleurs résultats lorsque l'on utilise toutes les caractéristiques.

**Valeur prédictive de notre base et recherche de biomarqueurs**

Lorsque l'on utilise toutes les caractéristiques qui sont à notre disposition nous obtenons une erreur de prédiction minimale de 0.41, et une p-value de 0.1, lors de la séparation en deux groupes (bon et mauvais

pronostic) de nos patients. On ne peut donc pas considérer cette séparation comme significative. Il advient donc de chercher à trouver de meilleurs résultats.
Pour cela, on teste la méthode des RSF sur différentes bases. Ainsi lorsque l'on ne garde que les caractéristiques cliniques et volumiques ont obtient de meilleurs résultats, avec une erreur de prédiction légèrement plus basse (0.39) et une p-value de 0.01 pour la séparation qui est donc significative. On en conclu aussi que l'utilisation des caractéristiques cliniques et volumiques reste nécessaire à l'obtention de bons résultats lorsque l'on utilise les caractéristiques texturales (erreur de prédiction de 0.42 pour les caractéristiques texturales seules contre 0.41 lorsque l'on ajoute les caractéristiques cliniques et volumiques). Ces résultats peuvent-être dû au fait qu'un grand nombre de caractéristiques texturales ont été utilisées et qu'elles sont fortement corrélées.

Afin de déterminer quelles sont les caractéristiques jouant un rôle dans ces résultats, l'importance des caractéristiques a été calculée pour chaque base à l'aide du VIMP. L'âge et le SUV max des lésions focales qui sont indiquées comme prédictives dans la littérature ont été retrouvés ici. Ce n'est cependant pas le cas de l'hémoglobine, la calcémie ou la créatinine par exemple. D'autres comme le sexe, le Deauville initial, le traitement ou la TLGT ont été trouvés comme prédictifs, bien que non indiqués dans la littérature.
Au niveau des caractéristiques texturales c'est l'HGZE, l'énergie, l'entropie, LGRE et LGZE qui ressortent.

A partir de ces caractéristiques nous avons créé une nouvelle base de données sur laquelle a été appliqué la RSF. Ce nouveau modèle donne de bien meilleurs résultats. En effet, on obtient une erreur de prédiction de 0.29 et une séparation qui est très significative ( p-value = 0.005 et valeur du test du log rank de 7.9). Ceci indique aussi qu'une sélection des caractéristiques avant d'appliquer la RSF pourrait être envisagée (avec par exemple Wilcoxon qui donne de bons résultats dans l'article de Parmar et al. [14]). Celle-ci permettrait d'améliorer les résultats et de gagner en temps. Enfin, il a été aussi observé que les résultats obtenus avec plus de patients donnent de meilleurs résultats. Il serait donc intéressant d'implémenter l'algorithme des données manquantes présenté par Ishwaran [1].

*Chapitre 9*

# Conclusion du stage et Perspectives

Différents problèmes ont été rencontrés lors du stage. Tout d'abord, il y a un certain nombre de données manquantes, ce qui induit une diminution du nombre de patients, et donc de devoir les traiter afin d'en utiliser le plus possible. La récupération des images TEP et CT d'environ 80 patients à l'hôpital a aussi été nécessaire. Cela inclut l'extraction de boîtes manuellement afin de réaliser les segmentations. L'extraction des boîtes a notamment due être refaite, à cause d'une erreur dans le logiciel. Un autre problème lié aux images fut la différence de traitement qu'elles ont subi au préalable. En effet, des problemes techniques ont induit des erreurs dans le calcul des segmentations de vérité terrain, ainsi que le calcul de leurs caractéristiques texturales.

Des problèmes de compatibilité de versions python et de systèmes d'exploitations ont aussi été rencontrés. De plus, le temps a manqué pour réaliser tout ce qui était souhaité. Cela est notamment dû au fait que le temps de calcul des arbres est grand et qu'une machine plus puissante n'a seulement été disponible que 1 mois avant la fin du stage.

Cependant, malgré les problèmes rencontrés et bien qu'il reste à améliorer, un pipeline entier a été écrit en python. Il inclut la segmentation des lésions, l'extraction des caractéristiques texturales, la prédiction de la rechute des patients par RSF et la classification dans des groupes de bon et mauvais pronostique. Ceci permettra d'automatiser, de diminuer les erreur possibles et de faire gagner du temps au médecin lors de l'étude de la survie des patients.

Au niveau de la segmentation , il reste à implémenter la méthode du k-mean et à régler les problèmes trouvés sur les boîtes. On peut aussi se lancer dans l'utilisation de l'apprentissage profond, en continuant la méthode du W-net. En ce qui concerne l'extraction caractéristiques texturales, les résultats obtenus sont plus proches de la vérité terrain qu'auparavant, il reste cependant des erreurs à corriger. Ceci peut être fait, tout d'abord en testant notre code avec les images proposées par l'ISBI [67] afin de vérifier que les erreurs ne sont pas dues à des des problèmes de paramétrage. Il peut être aussi intéressant de s'intéresser à l'extraction de d'autres paramètres, comme par exemple en utilisant l'analyse fractale sur des images TEP, comme Breki et al. [73] qui utilisent cette méthode sur des images TEP pour étudier l'impact d'un anticorps monoclonal sur des patients atteints de mélanome.

Au niveau de l'analyse de la survie, la prédiction de la survie chez des patients atteints de myélome multiple avec RSF est nouveau, et encore plus avec l'utilisation de caractéristiques texturales. Il n'y avait pas de package permettant de réaliser la RSF en python, d'où son implémentation en python lors du stage. Il permettra aussi plus tard de réaliser les modifications souhaitées. Il reste cependant à améliorer le RSF en python afin d'obtenir au minimum des résultats comme en R, puis si possible. Cela peut par exemple passer par l'utilisation de poids, pour contrebalancer la censure. On pourrait aussi se pencher sur l'inclusion des images CT.
Il a été vu que les résultats sont grandement influencés en fonction du nombre de patients et des caractéristiques

inclues dans le modèle. Il est donc prévu d'implémenter l'algorithme des données manquantes afin d'inclure les patients avec des données manquantes et de tester différentes méthodes de sélection des données. Enfin, des caractéristiques ont été déterminées comme prédictives grâce au VIMP. Certaines de ces caractéristiques sont déjà considérées dans la littérature mais d'autres sont nouvelles. Ces caractéristiques sont aussi bien cliniques que texturales ou volumiques.

Ces caractéristiques ont ainsi été utilisées pour composer une nouvelle base de données, à laquelle nous avons appliqué la RSF. C'est ainsi que l'on obtient de très bon résultats, meilleurs par exemple que la plupart des bases de données présentées par Ishwaran pour illustrer les Random Survival Forests. Enfin cela permet d'obtenir une séparation en deux groupes : bon et mauvais pronostique, qui est très significative. Pour conclure, outre l'amélioration des compétences en imagerie et en apprentissage automatique, ce stage a permis de préparer la thèse en s'appropriant le sujet et en commençant les différentes branches du projet.

*Annexe A*

# Annexes

## A.1 Annexe 1 : Les caractéristiques cliniques

| Caractéristiques cliniques | Définition |
|---|---|
| Nombre de lésions focales osseuses d'après le score de Bologne en TEP Baseline | 0 lésions : 0, 1 à 3 lésions : 1<br>4 à 10 lésions : 2, $\geq$ 10 lésions : 3 |
| Présence ou non de lésions extra-osseuses | |
| Deauville initial lésions focales | Comparaison avec intensité du foie. Equivalent : 3 |
| Deauville initial ostéomédullaire | Comparaison avec intensité du foie. Equivalent : 3 |
| Deauville initial global (KLM) | Comparaison avec intensité du foie. Equivalent : 3 |
| Nombre de lésions focales à l'IRM | <10 : 0, $\geq$ 10 : 1 |
| Âge | en années |
| Sexe | Femme : 0, Homme : 1 |
| Hémoglobine | en g/dl |
| Calcémie | en g/dl |
| Créatinémie | en $\mu$mol/l |
| R-ISS | Stade de la maladie |
| Bras de traitement | A ou B |

TABLE A.1 – Les caractéristiques cliniques

## A.2 Annexe 2 : Les caractéristiques volumiques

| |
|---|
| SUV Max lesions focales osseuses |
| SUV ostéomédulaire |
| SUV lésions extra-ostéomédullaire |
| Volume métabolique tumoral de la lésion la plus fixante, segmentation 40% |
| Volume métabolique tumoral de la lésion la plus fixante, segmentation au vote majoritaire |
| Volume métabolique tumoral total, segmentation 40% |
| Volume métabolique tumoral total, segmentation au vote majoritaire |
| Glycolyse totale de la lésion la plus fixante, segmentation 40% |
| Glycolyse totale de toutes les lésions, segmentation au vote majoritaire |
| Glycolyse totale de la lésion la plus fixante, segmentation 40% |
| Glycolyse totale de toutes les lésions segmentation au vote majoritaire |

TABLE A.2 – Les caractéristiques volumiques

## A.3 Annexe 3 : Les caractéristiques Texturales

| Premier ordre | GLCM | GLRLM | GLSZM |
|---|---|---|---|
| Maximum | Homogénéité | HGRE (High Gray Level Run Emphasis) | HGZE (High Gray Level Zone Emphasis) |
| | Entropie | LGRE (Low Gray Level Run Emphasis) | ZLNU (Size-Zone Non-Uniformity) |
| | Energie | SRE (Short Run Emphasis) | SZHGE (Small Area High Gray Level Emphasis) |
| | Corrélation | LRE (Long Run Emphasis) | LZLGE (Low Gray Level Zone Emphasis) |
| | Contraste | | SZE (Small Area Emphasis) |
| | Dissimilarité | | ZP (Zone pourcentage) |
| | | | RP (Run pourcentage) |

TABLE A.3 – Les caractéristiques texturales

# Table des figures


1.1 Organigramme du laboratoire Ls2n . . . . . 8

2.1 Le myélome multiple est un cancer de la moelle osseuse se caractérisant par la prolifération de plasmocytes mutés . . . . . 10
2.2 TEP-SCAN au CHU Nantes TEP/TDM (Tomographie par Emission de Positons/TomoDensitoMètre), Biogaph mCT (SIEMENS) . . . . . 11
2.3 Fonctionnement de la TEP. Les positons vont s'annihiler avec des électrons ce qui produira des photons $\gamma$ émis dos à dos. Ce sont ces photons $\gamma$ qui seront détectés et donneront l'orientation. . . 11
2.4 Exemples d'images CT (gauche), TEP (centre), et TEP-CT fusionnée (droite) . . . . . 12

3.1 Exemple de courbes de survie. Graphique représentant la probabilité de ne pas avoir eu d'évènements jusqu'au temps t, en fonction de t en jours. La courbe noir représente la courbe moyenne.[7] . . . 14
3.2 Représentation des types de censure. Le cas 1 représente la censure à gauche (exemple : ne pas connaître la date d'apparition de la maladie). Le cas 2 représente la censure à droite (exemple : perdre de vue un patient).Cas 3 : censure droite et gauche. Cas 4 : pas de censure (exemple : La date d'apparition de la maladie est connue et un évènement à été enregistré).[8] . . . . . 14
3.3 Exemple de courbes de Kaplan Meier présentant la survie de deux groupes. En rose, les femmes et en bleu les hommes. . . . . 16
3.4 Exemple de problème de séparation à deux classes. Il faut trouver une surface de séparation f qui permette de classer les éléments [17]. . . . . 18
3.5 schéma d'un réseau de neurones . . . . . 19

4.1 Comparaison de 12 méthodes de classification ( Nnet : Neural network, DT : Decision Tree, BST : Boosting, BY : Bayesian, BAG : Bagging, RF : Random Forest, MARS : Multi adaptive regression splines, SVM : Support vector machines, DA : Discriminant analysis, NN : Neirest neighbour, GLM : Generalized linear models, PLSR : Partial least squares and prinicipal componenet regression) et 14 méthodes de sélections de paramètres. La comparaison se fait sur la valeur de l'AUC (Area Under Curve) réalisée dans l'article de Parmar et al. [14] . . . . . 22

5.1 Diagramme UML du pipeline du projet . . . . . 25
5.2 Présentation d'un W-net simplifié. En gris le premier V-net pour la détection du squelette, et en rose le second V-net pour la détection des lésions [46]. . . . . 27
5.3 Construction de la matrice GLCM . . . . . 30
5.4 Construction de la matrice GLSZM . . . . . 30

5.5 Schéma du bagging avec pour règle de base un arbre CART [69] . . . 32
5.6 Schéma des forêts aléatoires RF-RI. Parmi les échantillons Ln, des sous échantillons sont sélectionnés aléatoirement pour construire des prédicteurs $\hat{h}(., \Theta_l, \Theta'_l)$. Les prédicteurs sont ensuite agrégés pour donner$\hat{h}_{RF-RI}(.)$ . . . 34
5.7 Exemple d'arbre, où le critère d'arrêt est d'avoir un minimum de 15 individus dans un nœud pour réaliser une séparation. Profondeur de l'arbre : 4. . . . 36
5.8 Exemple de courbes de mortalité prédites. Une courbe correspond à un patient. . . . 38

6.1 Matrice fantôme utilisée pour la vérification des caractéristiques . . . 41
6.2 Exemple de présentation des valeurs de biomarqueurs GLCM pour un calcul sur image 3D avec moyenne sur une matrice 3D. "Dig. phantom" correspond aux calculs réalisés sur le fantôme. Les configurations C, D et E, aux calculs faits sur des images cliniques avec 3 configurations différentes. 42
6.3 Pipeline d'étude de la survie . . . 43

7.1 Erreur dans la boîte initiale influençant les segmentations . . . 48
7.2 Graphiques présentant l'erreur de prédiction minimale sur le set de test (a) et le set d'entraînement (b) avec python en fonction du mode (LogRank en haut : graphique du haut, LogRank Random : graphique du bas), du nombre d'arbres et de min-sample-split. . . . 51
7.3 Graphiques présentant l'erreur de prédiction minimale sur le set de test (a) et le set d'entraînement (b) avec R en fonction du mode (LogRank en haut : graphique du haut, LogRank Random : graphique du bas), du nombre d'arbres et de min-sample-split. . . . 52
7.4 Séparation en deux groupes (Bon et mauvais pronostic) selon les caractéristiques utilisées (67 patients) 54
7.5 Kaplan-Meier de la classification par RF à 3 ans sur la base du myélome multiple avec toutes les caractéristiques. Accuracy : 0.47, recall : 1.2, logRank = 0.05, p-value = .94, nombre de patients avec un bon et mauvais pronostique : 12 et 11 . . . 55
7.6 Kaplan-Meier de la classification par RF à 3 ans sur la base du myélome multiple avec les caractéristiques cliniques et volumiques. Accuracy : 0.61, recall : 0.9, logRank = 0.56, p-value = 0.32, nombre de patients avec un bon et mauvais pronostique : 14 et 9 . . . 55
7.7 Valeur moyenne du Vimp sans regroupement des caractéristiques texturales par méthode de calcul 56
7.8 Fréquence dans le Top sans regroupement des caractéristiques texturales par méthode de calcul . . 57
7.9 Fréquence dans le Top avec regroupement des caractéristiques texturales par méthode de calcul . . 58
7.10 Valeur moyenne du VIMP avec regroupement des caractéristiques texturales par méthode de calcul (Somme) . . . 58
7.11 Valeur moyenne du Vimp avec regroupement des caractéristiques texturales par méthode de calcul (Moyenne) . . . 58
7.12 Séparation par les caractéristiques cliniques (base de 67 patients) . . . 59
7.13 Séparation par les caractéristiques image (base de 67 patients) . . . 60
7.14 Séparation par les caractéristiques induisant des différences dans les courbes en fonction du nombre de patients (70 patients pour la figure (a) et 89 patients pour les figures (b), (c), (d)) . . . 61
7.15 Séparation sur notre nouvelle base de données ; erreur de prediction = 0.298 ; logRank = 7.9 ; p value = 0.005 ; nombre de patients dans le groupe de bon pronostic = 9 ; nombre de patients dans le groupe de bon pronostic = 13 . . . 62
7.16 Importance des caractéristiques obtenues avec notre nouvelle base de données . . . 63

# Liste des tableaux


5.1 Exemples de caractéristiques sémantiques et agnostiques . . . . . 29

7.1 Résultats de la segmentation 2.5 à l'aide des boîtes initiales. . . . . 47
7.2 Résultats de la segmentation 40% à l'aide des boîtes initiales. . . . . 47
7.3 Comparaison Pyradiomics et ISBI sur le fantôme . . . . . 48
7.4 Ecart-types et moyennes des différences observées entre les valeurs normalisées de pyradiomics et celles de la vérité terrain. (la normalisation a été réalisée par l'écart-type par variable des valeurs de la vérité terrain). Les calculs ont été réalisés sur 5 patients. . . . . 49
7.5 Comparaison des résultats de l'erreur de prédiction moyenne sur le set de test sur la base vétéran en R et en python.Les valeurs ont été calculées sur les erreurs de prédictions trouvées lors de la détermination des paramètres optimaux. . . . . 50
7.6 Comparaison des résultats de l'erreur de prédiction moyenne sur le set de test sur la base vétéran en R et en python.Les valeurs ont été calculées sur les erreurs de prédictions trouvées lors de la détermination des paramètres optimaux . . . . . 50
7.7 Erreurs de prédiction en fonction du paramètre max-features . . . . . 52
7.8 Erreurs de prédiction en fonction de la base de données . . . . . 53
7.9 Résultats du RF classifier pour une séparation à 3 ans . . . . . 55
7.10 Résultats du RF classifier pour une séparation à 3 ans, avec les caractéristiques cliniques et volumiques 55
7.11 Erreurs de prédiction sur la nouvelle base de données . . . . . 62

A.1 Les caractéristiques cliniques . . . . . 68
A.2 Les caractéristiques volumiques . . . . . 68
A.3 Les caractéristiques texturales . . . . . 69

# Bibliographie


[1] E. Blackstone M. Lauer H. Ishwaran, U. Kogalur. Random surival forest. *The annals of applied statistics*, 2(3) :841–860.

[2] Le myélome multiple : Informations patients. *SHF*.

[3] Myélome multiple : du cœur des cellules à leur environnement.

[4] P. Gambert. On commence à parler de guérison du myélome. *Ouest-France Pays de la Loire*.

[5] Dr P.H. Granier. Le myélome multiple. *MN-net*.

[6] C. Bodet-Milin et al. P. Moreau, F. Caillon. 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.

[7] Joseph Rickert. Survival analysis with r.

[8] Iragaël Joly and Damien Rousselière. A propos de la capacité à survivre des coopératives agricoles : une étude de la relation entre âge et mortalité des coopératives agricoles françaises. 92 :259–289, 01 2011.

[9] Edward B. Garon et al. Pembrolizumab for the treatment of non-small-cell lung cancer. *The New England journal of medicine*, 372 21 :2018–28, 2015.

[10] Gilbert Colletaz. ModÈles de survie. *Notes de Cours MASTER 2 ESA voies professionnelle et recherche, université d'Orleans*.

[11] Philippe Saint-Pierre. Introduction à l'analyse des durées de survie. *Université Pierre et Marie Curie*.

[12] Andre Berchtold. Données longitudinales et modèles de survie, courbes de survie. *Département des sciences économiques, Université de Genève. Cours de Master*.

[13] t al. Primrose. Adjuvant capecitabine for biliary tract cancer : The bilcap randomized study. *Journal of Clinical Oncology*, 35(15_suppl) :4006–4006, 2017.

[14] Chintan A. Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo J. W. L. Aerts. Machine learning methods for quantitative radiomic biomarkers. In *Scientific reports*, 2015.

[15] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics : Images are more than pictures, they are data. In *Radiology*, volume 278, pages 563–577, 2016.

[16] Fayçal Ben Bouallègue et al. Association between textural and morphological tumor indices on baseline pet-ct and early metabolic response on interim pet-ct in bulky malignant lymphomas. *Medical Physics*, 44(9).

[17] Marin Ferecatu et Nicolas Thome Michel Crucianu. Cours - svm lineaires. *Cours Cnam RCP209*.

[18] Marc Parizeau. Réseaux de neurones. *Université de Laval*, 2006.

[19] L. Breiman. Bagging predictors. *Machine Learning*, 42 2 :123–140, 1996.

[20] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, volume 55 1, pages 119–139, 1997.

[21] L. Breiman. Random forests. *Machine Learning*, 45 1 :5–32, 2001.

[22] Martin Vallières et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports*, 7(10117) :2911–2918.

[23] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220, 1972.

[24] Yan Zhou and J J Mcardle. Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80 3 :811–33, 2015.

[25] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning : A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3) :651–674, 2006.

[26] Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23 1 :77–91, 2004.

[27] M. Ingrisch et al. Prediction of 90y radioembolization outcome from pretherapeutic factors with random survival forests. *J Nucl Med. Mai 2018*, 59(5) :769–773.

[28] Martin Vallières, C. R. Freeman, Sonia R Skamene, and Isaam El Naqa. A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in medicine and biology*, 60 14 :5471–96, 2015.

[29] Justine B. Nasejje and Henry Mwambi. Application of random survival forests in understanding the determinants of under-five child mortality in uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Research Notes*, 10(1) :459, Sep 2017.

[30] Claudia Bühnemann et al. Quantification of the heterogeneity of prognostic cellular biomarkers in ewing sarcoma using automated image and random survival forest analysis. *Plos one*.

[31] Fen Miao et al. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access*, 6 :7244–7253.

[32] Arabin Kumar Dey et al. Some variations on ensembled random survival forest with application to cancer research. *arXiv*.

[33] Edward Choi, Andy Schuetz, Walter Stewart, and J Sun. Using recurrent neural network models for early detection of heart failure onset. 24 :ocw112, 08 2016.

[34] Linxia Liao and Hyung-Il Ahn. Combining deep learning and survival analysis for asset health management. 2016.

[35] Jared L.Katzman et al. Deepsurv : Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med Res Methodol*, 35(25) :2911–2918.

[36] Mathieu Hatt, Florent Tixier, Dimitris Visvikis, and Catherine Cheze le Rest. Radiomics in pet/ct : More than meets the eye ? *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 58 3 :365–366, 2017.

[37] Ruben T H M Larue, Gilles Defraene, Dirk K. M. De Ruysscher, Philippe Lambin, and Wouter J. C. van Elmpt. Quantitative radiomics studies for tissue characterization : a review of technology and methodological procedures. In *The British journal of radiology*, 2017.

[38] C. Bourgier et al. Définition et applications cliniques des radiomics. radiomics : Definition and clinical development. *Cancer / Radiothérapie*, 19(6-7) :532–537.

[39] Hugo J. W. L. Aerts et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. In *Nature communications*, volume 5, 2014.

[40] Devinder Kumar et al. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction.

[41] C. Lartizien, M. Rogez, E. Niaf, and F. Ricard. 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, May 2014.

[42] Lei Bi, Jinman Kim, Ashnil Kumar, Lingfeng Wen, David Dagan Feng, and Michael J. Fulham. Automatic detection and classification of regions of fdg uptake in whole-body pet-ct lymphoma studies. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 60 :3–10, 2017.

[43] Florent Tixier, Catherine Cheze Le Rest, and Mathieu Hatt et al. Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 52 3 :369–78, 2011.

[44] Seraina Steiger, Michael Arvanitakis, Beate Sick, Walter Weder, Sven Hillinger, and Irene A. Burger. Analysis of prognostic values of various pet metrics in preoperative 18f-fdg pet for early-stage bronchial carcinoma for progression-free and overall survival : Significantly increased glycolysis is a predictive factor. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 58 12 :1925–1930, 2017.

[45] Constantin Lapa et al. 18fdg-pet/ct for prognostic stratification of patients with multiple myeloma relapse after stem cell transplantation. In *Oncotarget*, volume 5, page 7381–7391, 2014.

[46] Lina Xu et al. 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).

[47] Olivier Decaux et al. Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients : A study of the intergroupe francophone du myélome. *Journal of Clinical Oncology*, 26(29).

[48] B. Amin Samir kumar and Wai-Ki Yip et al. Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. In *Leukemia*, volume 28, pages 2229–2234, Nov. 2014.

[49] M. Hauser et S. Minvielle H. Pang. Pathway-based identification of snps predictive of survival. *European Journal of Human Genetics*, 19(25) :704–709.

[50] U. Bross-Bach et al. M. Horger, C. D. Claussen. Whole-body low-dose multidetector row-ct in the diagnosis of multiple myeloma : an alternative to conventional radiography. *European Journal of Radiology*, 54(2) :289–297.

[51] J. C. Dutoit and K. L. Verstraete. Mri in multiple myeloma : a pictorial review of diagnostic and post-treatment findings. *Insights into Imaging*, 7(4) :553–569.

[52] S. Usmani et al. M. A. Dimopoulos, J. Hillengass. Role of magnetic resonance imaging in the management of patients with multiple myeloma : a consensus statement. *Journal of Clinical Oncology*, 333(6) :657–664.

[53] I. I. Riphagen S. Zweegman O. S. Hoekstra D. Van Lammeren-Venema, J. C. Regelink and J. M. Zijlstra. 18f-fluoro-deoxyglucose positron emission tomography in assessment of myeloma-related bone disease : a systematic review. *Cancer*, 118(8) :1971–1981.

[54] Y. Nakamoto. Clinical contribution of pet/ct in myeloma : from the perspective of a radiologist. *Clinical Lymphoma, Myeloma & Leukemia*, 14(1) :10–11.

[55] S. J. Eustace J. Madewell P. J. O'Gorman C. F. Healy, J. G. Murray and P. O'Sullivan. Multiple myeloma : a review of imaging features and radiological techniques. *Bone Marrow Research*, 2011 :1–9.

[56] C. Nanni et al. M. Cavo, E. Terpos. Role of 18 f-fdg pet/ct in the diagnosis and management of multiple myeloma and other plasma cell disorders : a consensus statement by the international myeloma working group. *The Lancet Oncology*, 18(4) :206–217.

[57] A. Schirbel et al. C. Lapa, M. Schreder. [68ga]pentixafor-pet/ct for imaging of chemokine receptor cxcr4 expression in multiple myeloma—comparison to [18f]fdg and laboratory values. *Theranostics*, 7(1) :205–212.

[58] C. Lartizien et al. 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.

[59] Marie-Charlotte Desseroit et al. Reliability of pet/ct shape and heterogeneity features in functional and morphological components of non-small cell lung cancer tumors : a repeatability analysis in a prospective multi-center cohort. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 58 3 :406–411, 2017.

[60] Mathieu Hatt et al. 18f-fdg pet uptake characterization through texture analysis : investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 56 1 :38–44, 2015.

[61] Clément Bailly et al. Revisiting the robustness of pet-based textural features in the context of multi-centric trials. In *PloS one*, 2016.

[62] T. Carlier et al. Valeur pronostique des paramètres de texture tep au diagnostic dans le myélome multiple symptomatique (mm). *Médecine Nucléaire*, 41 3 :143, 2017.

[63] Twyla B. Bartel et al. F18-fluorodeoxyglucose positron emission tomography in the context of other imaging techniques and prognostic factors in multiple myeloma. *Blood*, 114 10 :2068–76, 2009.

[64] Elena Zamagni et al. 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.

[65] James E. McDonald et al. Assessment of total lesion glycolysis by 18f fdg pet/ct significantly improves prognostic value of gep and iss in myeloma. *Clinical Cancer Research*, 2016.

[66] Caroline Bodet-Milin, Thomas Eugène, Clément Bailly, Marie Lacombe, Eric Frampas, Benoît Dupas, Philippe Moreau, and Françoise Kraeber-Bodéré. Fdg-pet in the evaluation of myeloma in 2012. *Diagnostic and interventional imaging*, 94 2 :184–9, 2013.

[67] M. Vallières S. Löck A. Zwanenburg, S. Leger. Image biomarker standardisation initiative. *Journal of Clinical Oncology*, 35(25) :2911–2918.

[68] pyradiomics documentation. 2016.

[69] Jean-Michel Poggi Robin Genuer. Arbres cart et forêts aléatoires, importance et sélection de variables. *HAL*.

[70] Amr Hanbali, Mona Hassanein, Walid Rasheed, Mahmoud Deeb Aljurf, and Fahad Alsharif. The evolution of prognostic factors in multiple myeloma. In *Advances in hematology*, 2017.

[71] P. Leif Bergsagel. Prognostic factors in multiple myeloma. *Clinical Cancer Research*, 9(2) :533–534, 2003.

[72] Prognosis and survival for multiple myeloma. *www.cancer.ca*.

[73] Christina-Marina Breki et al. Fractal and multifractal analysis of pet/ct images of metastatic melanoma before and after treatment with ipilimumab. In *EJNMMI research*, volume 6, 2016.